Journal of Information Hiding and Multimedia Signal Processing ©2018 ISSN 2073-4212
Ubiquitous International Volume **9**, Number **5**, September **2018**

Research on PLSA Model based Semantic Image Analysis: A Systematic Review

Dongping Tian[1,2]

[1]Institute of Computational Information Science
Baoji University of Arts and Sciences
No.1 Hi-Tech Avenue, Hi-Tech District, Baoji, Shaanxi 721013, P.R. China

[2]Institute of Computer Software
Baoji University of Arts and Sciences
No.44 Baoguang Road, Weibin District, Baoji, Shaanxi 721016, P.R. China
tiandp@ics.ict.ac.cn, tdp211@163.com

Received November, 2017; revised May, 2018

Abstract. *Semantic image analysis is an active topic of research in computer vision and pattern recognition. In the last decade, a huge number of related models have been proposed, among which probabilistic latent semantic analysis (PLSA) is one of the most popular topic models due to its potential ability to capture the aspects that contain general information about visual co-occurrence. However, compared with various PLSA models and their corresponding applications in semantic image analysis, there is almost no review research and analysis about PLSA related studies. So the current paper, to begin with, elaborates the basic principles of the PLSA model, followed by summarizes PLSA with applications to image annotation, image retrieval, image classification and several other applications. Subsequently, a detailed survey on PLSA model itself improvement, mainly including its initialization methods, visual words construction, adding hidden layers and integration with other models, is comprehensively and systematically reviewed. Finally, this paper is ended with a summary of some important conclusions and several potential research directions for semantic image analysis in the near future.*
Keywords: PLSA, AIA, Image retrieval, Image classification, Visual words

1. **Introduction.** Probabilistic models with hidden topic variables, originally developed for statistical text modeling of large document collections such as latent semantic analysis, probabilistic latent semantic analysis [1], latent Dirichlet allocation [2], and correlated topic model [3], have recently become an active research topic for multimedia representation and annotation in both 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 that is 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 multimedia processing, such as object recognition, scene classification, automatic segmentation

and image annotation. Except for the pros and cons of PLSA model aforementioned, it also should be noted that most of the existing works related to PLSA have focused on both its improvements and applications whereas there is almost no review research. So in this paper, we present a comprehensive survey on PLSA for semantic image analysis. Specifically, we describe it gradually along two aspects. To start with, the basic principle of PLSA as well as its parameter estimation is introduced, followed by the PLSA with applications to image annotation, image retrieval, image classification and several other applications are summarized. On the other hand, we have provided a comprehensive review on PLSA itself improvement, mainly including its initialization, visual words, hidden layers and integration with other models.

The rest of this paper is organized as follows. Section 2 introduces the basic PLSA model and its parameter estimation approach. In section 3, we elaborate some representative applications of PLSA from the perspective of image annotation, image retrieval, image classification and several other applications, respectively. Section 4 presents a detailed description of the improvements for PLSA model itself, including its initialization methods, visual words construction, adding hidden layers and integration with other models, respectively. In section 5, we draw some important conclusions and highlight the potential research directions of PLSA in semantic image analysis for the future.

2. PLSA Model.

Note that in this section, the basic principle of PLSA model and its parameter estimation will be succinctly introduced. More details can be gleaned from the following subsections.

2.1. Introduction of PLSA.

Probabilistic latent semantic analysis (PLSA) [1] is a statistical latent aspect model for co-occurrence data that associates an unobserved class variable $z_k \in \{z_1, z_2, \cdots, z_K\}$ with each observation, an observation being the occurrence of a word in a particular document. Let's introduce the following probabilities: $P(d_i)$ is used to denote the probability that a word occurrence will be observed in a particular document d_i, $P(w_j|z_k)$ denotes the class-conditional probability of a specific word conditioned on the unobserved class variable z_k, and $P(z_k|d_i)$ denotes a document-specific probability distribution over the latent variable space. Based on these definitions, one can define a generative model for word/document co-occurrences by the following scheme:

- select a document d_i with probability $P(d_i)$.
- pick a latent class z_k with probability $P(z_k|d_i)$.
- generate a word w_j with probability $P(w_j|z_k)$.

As a result, one can obtain an observation pair (d_i, w_j) while the latent class variable z_k is discarded. Note that to translate the data generation process into a joint probability model results in the following expression.

$$P(d_i, w_j) = P(d_i)P(w_j|d_i) \tag{1}$$

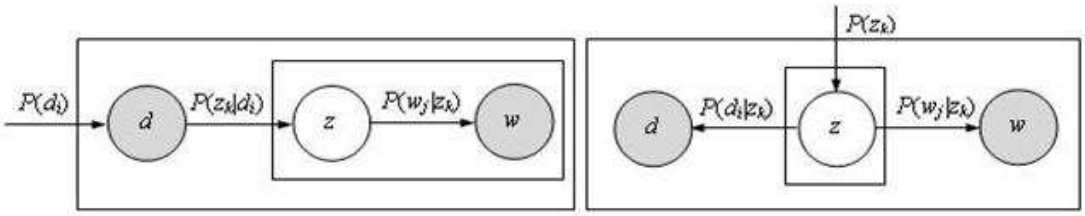
$$P(w_j|d_i) = \sum_{k=1}^{K} P(w_j|z_k)P(z_k|d_i) \tag{2}$$

$$P(d_i, w_j) = P(d_i)\sum_{k=1}^{K} P(w_j|z_k)P(z_k|d_i) \tag{3}$$

Applying Bayes rule, the above joint probability (Eq.(3)) can be further expressed as below.

$$P(d_i, w_j) = \sum_{k=1}^{K} P(z_k)P(d_i|z_k)P(w_j|z_k) \tag{4}$$

In addition, a representation of the aspect model in terms of a graphical model is depicted in Fig. 1(a). Since the cardinality of the latent aspects is typically smaller than the number of documents (and terms) in the collection, $K \leftarrow \min\{N, M\}$, it acts as a bottleneck variable in predicting words. Correspondingly, Fig. 1(b) illustrates the symmetric graphical model of the PLSA denoted by Eq.(4).



(a) Asymmetric representation (b) Symmetric representation

FIGURE 1. Graphical model representation of PLSA

2.2. **Model fitting with EM.** Note that the PLSA model (Eq.(3)) expresses each document as a convex combination of K aspect vectors. This amounts to the matrix decomposition. Essentially, each document is modeled as a mixture of aspects — the histogram for a particular document being composed of a mixture of the histograms corresponding to each aspect. The model parameters of PLSA are the two conditional distributions: $P(w_j|z_k)$ and $P(z_k|d_i)$, in which $P(w_j|z_k)$ characterizes each aspect and remains valid for documents out of the training set, on the other hand, $P(z_k|d_i)$ is only relative to the document-specific and cannot carry any prior information to an unseen document. Due to the existence of the sums inside the logarithm, direct maximization of the log-likelihood by partial derivatives is very difficult. Thus the expectation maximization (EM) is commonly employed to estimate the parameters through maximizing the log-likelihood of the observed data.

$$L = \sum_{i=1}^{N} \sum_{j=1}^{M} n(d_i, w_j) \log P(d_i, w_j) \tag{5}$$

where $n(d_i, w_j)$ denotes the number of times the term w_j occurred 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:

$$P(z_k|d_i, w_j) = \frac{P(z_k|d_i)P(w_j|z_k)}{\sum_{l=1}^{K} P(z_l|d_i)P(w_j|z_l)} \tag{6}$$

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

$$P(w_j|z_k) = \frac{\sum_{i=1}^{N} n(d_i, w_j)P(z_k|d_i, w_j)}{\sum_{m=1}^{M} \sum_{i=1}^{N} n(d_i, w_m)P(z_k|d_i, w_m)} \tag{7}$$

$$P(z_k|d_i) = \frac{\sum_{j=1}^{M} n(d_i, w_j)P(z_k|d_i, w_j)}{\sum_{j=1}^{M} n(d_i, w_j)} \tag{8}$$

Note that the E-step and M-step are alternated until a termination condition is met. In particular, one can make use of a technique known as early stopping, in which one does

not necessarily optimize until convergence but instead stops updating the parameters once the performance on hold-out data is not improving. This is a standard procedure that can be used to avoid over-fitting in the context of iterative fitting methods. In addition, as for the two parameters $P(w_j|z_k)$ and $P(z_k|d_i)$, if one of them is known, the other one can be inferred by leveraging fold-in method — a partial version of the EM algorithm, which updates the unknown parameters with the known parameters kept fixed so that it can maximize the likelihood with respect to the previously trained parameters.

3. **Applications of PLSA.**

Note that in this section, some representative applications of PLSA model will be reviewed from the aspects of image annotation, image retrieval, image classification and several other applications, respectively.

3.1. **Image annotation.**

Image annotation has been an active topic of research in computer vision for decades due to its potentially large impact on both image understanding and web image search. To be specific, automatic image annotation (AIA) refers to a process to automatically generate textual words to describe the content of a given image, which plays a crucial role in semantic based image retrieval. In this subsection, we will review some pioneer works for automatic image annotation by using PLSA related models. In general, the state-of-the-art research on AIA has proceeded along two categories, i.e., generative model and discriminative model. The basic idea of generative model is to construct a model from joint probability of image features and words, and then use the Bayes rule and marginalization of probability to estimate the conditional probability of words given image features. But in the discriminative case, the model is directly being constructed the conditional probability of words given image features. In the context of PLSA model for image annotation, it obviously belongs to the former. As the representative work of this perspective, Monay et al. present a series of PLSA models for AIA [4-6], among which PLSA-MIXED [4] learns a standard PLSA model on a concatenated representation of the textual and visual features while PLSA-WORDS or PLSA-FEATURES [5,6] allows modeling of an image as a mixture of latent aspects that is defined either by its text captions or by its visual features for which the conditional distributions over aspects are estimated from one of the two modalities only. In order to extract effective features to reflect the intrinsic content of images as complete as possible, Zhang et al.[7] put forward a multi-feature PLSA (MF-PLSA) to tackle the problem of combining low-level visual features for image region annotation in that it handles data from two different visual feature domains by attaching one more leaf node to the graphical structure of the original PLSA, where the two types of visual descriptors extracted from the same key point location are jointly modeled yet with their conditional distributions constrained via a single latent variable. In recent years, Peng et al.[8] present a new approach and algorithm for the semantic concept annotation based on audio PLSA model, which includes two main aspects: audio vocabulary construction and audio PLSA. To be exact, an audio-clip is firstly partitioned into a few homogeneous audio-segments according to its content change. An audio vocabulary is then constructed by the rival penalized competitive learning clustering of audio-segments so as to form the bag-of-word representation for each audio-clip. Afterwards the PLSA is employed to discover the latent topics existing in audio-clips and the concept classification is carried out by a SVM further. In addition, a supervised PLSA (S-PLSA) model [9] is formulated 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. Particularly, more recent work [10] has begun to introduce some semi-supervised learning

techniques to enhance the quality of the training image data so as to improve the performance of PLSA model for automatic image annotation. In summary, all of the models mentioned above are able to obtain promising annotation results. However, note that the number of latent aspects, in most cases, is set empirically rather than based on some theorem and deduction. As a result, how to appropriately determine the aspect number during PLSA training for AIA is well worth exploring. On the other hand, how to relax the independence assumption of PLSA as well as to introduce the contexts of images and keywords is also worth pursuing.

3.2. **Image retrieval.** From the perspective of probability theory, image retrieval can be seen as a procedure of ranking images in the database according to their posterior probabilities of being relevant to the query concept. As a latent topic model, PLSA has been widely used in the area of image retrieval [11-13]. In [11], Shah-hosseini and Knapp construct a novel PLSA based image retrieval system by utilizing the search history to find hidden image semantics of the database [10]. In addition, image features are integrated into the model as well to improve the retrieval performance. Subsequently, the PLSA model is employed to infer which visual patterns describe each object for image retrieval [12]. In literature [13], PLSA is exploited to build a high-level representation appropriate for retrieval by considering images as mixtures of topics. To start with, two PLSA models based on visual features and tags are constructed respectively, followed by these two models are integrated by learning a third PLSA model on the already derived topic mixtures and thus a multimodal high-level image representation is captured for retrieval. Fig. 2 illustrates the framework of the multimodal retrieval system. Thereafter, Sayad et al.[14] introduce a new method using multilayer PLSA for image retrieval, which can effectively eliminate the noisiest words generated by the vocabulary building process. In the meanwhile, a spatial weighting scheme is adopted to reflect the information about the spatial structure of the images. Finally, they construct visual phrases from groups of visual words that are involved in strong association rules. In more recent work [15], the standard PLSA model is extended to higher order for image indexing by treating images, visual features and tags as the three observable variables of an aspect model so as to learn a space of latent topics that incorporates the semantics of both visual and tag information. Alternatively, the relevance feedback technique can be introduced as an effective solution to improve the performance of content-based image retrieval systems based on the PLSA model. To our knowledge, there is no relevant research in this point.

3.3. **Image classification.** Image classification is a challenging problem in computer vision, whose aim is to decide whether an image belongs to a certain category or not. As the classic work, Quelhas et al.[16] apply PLSA along with the support vector machine (SVM) to the task of scene classification. Through evaluation on four datasets, they show that as an unsupervised probabilistic model for collections of discrete data, PLSA has the dual ability to generate a robust, low-dimensional scene representation, and to automatically capture the meaningful scene aspects. Similar to [16], a combination of PLSA and multi-class discriminative classifier (k-NN or SVM) is proposed in [17] for scene classification. The principal difference between them is the former uses sparse features and is applied to classify images into only three scene types. Followed by Bosch et al.[18] exploit PLSA model for scene classification under changes in the visual vocabulary and number of latent topics learnt. In recent years, Zhuang et al.[19] propose a semi-supervised PLSA (SS-PLSA) for image classification that can greatly prevent the inter-impact between different categories. Compared with the classic non-supervised PLSA, this method is able to overcome the shortcoming of poor classification performance when the features of two categories are quite similar. Especially the iteration process can be

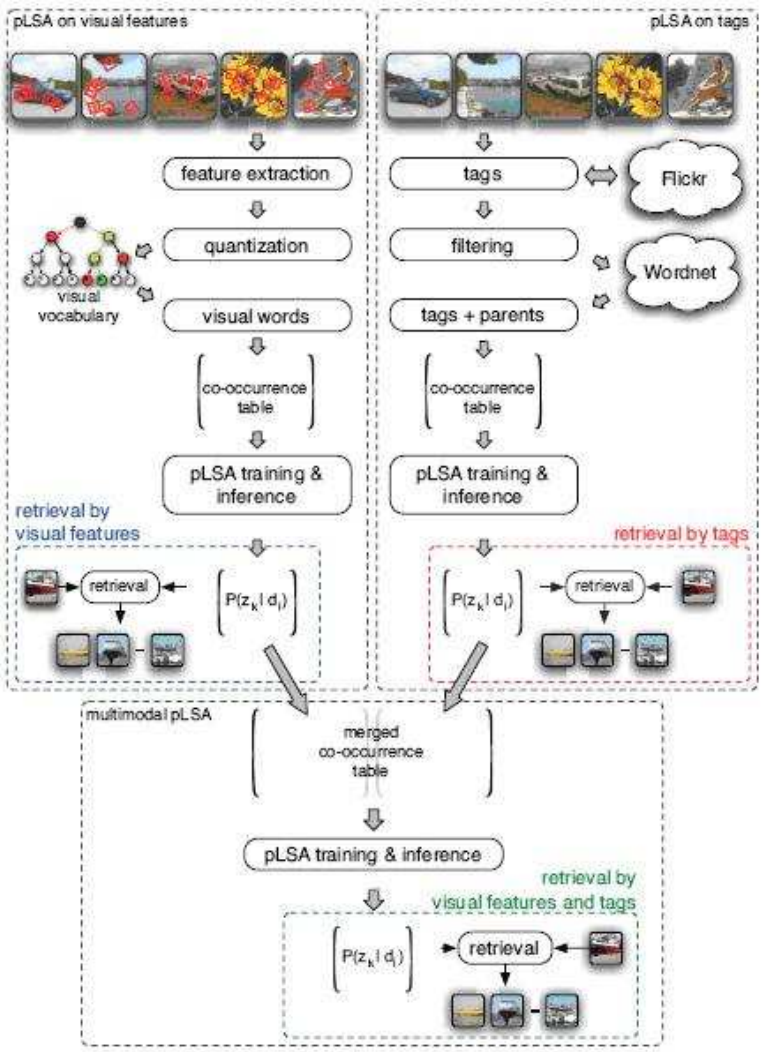


FIGURE 2. The framework of the multimodal retrieval system

directed carefully to the desired result via introducing category label information into the EM algorithm. The work by Lu et al.[20] puts forward a rival penalized competitive learning based method to provide a good initial estimate for the PLSA model used in image categorization through directly maximizing the likelihood function based on the observed data. More recently, Jin et al.[21] leverage a histogram to represent the spatial relationships between objects. To be specific, they first use fuzzy k-nearest neighbors to classify the spatial relationships (left, right, above, below, near, far, inside, outside) with soft labels, and then extend PLSA by taking into account the spatial relationships between topics (SR-PLSA), which is subsequently applied to model the image as the input for support vector machine to classify the scene. Fig. 3 illustrates its graphical model representation.

Correspondingly, the joint probability is defined as follows:

$$P(d_i, z_k, z_l, R, w_j) = P(R|h) = \sum_{n=1}^{8} w_n(\mu_n - |\mu_n - P(r_n|h)|) \tag{9}$$

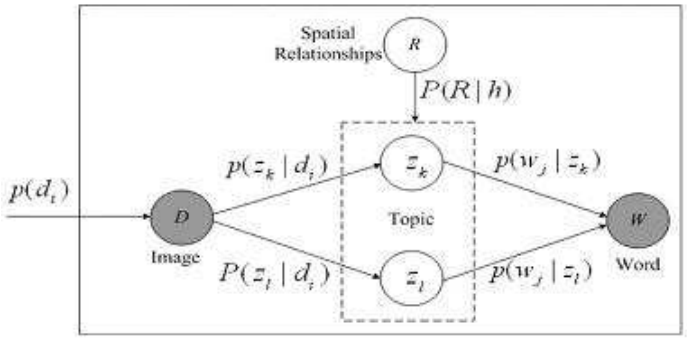


FIGURE 3. Graphical model representation of the SR-PLSA

where

$$P(r_n|h) = \frac{\sum_{i=1}^{k} P(r_n|z_i)(d(h,z_i))^{-2/(n-1)}}{\sum_{i=1}^{k} (d(h,z_i))^{-2/(n-1)}} \tag{10}$$

3.4. **Other applications.** Except for automatic image annotation, image retrieval and image classification, PLSA has also been extensively applied for other tasks [22-24]. As the representative work, Jiang et al.[22] propose a co-regularized probabilistic latent semantic analysis (Co-PLSA) model for multi-view clustering. It attempts to perform PLSA in different views collaboratively through a constraint of pairwise co-regularization. The central idea behind co-regularization is that the sample similarities in the topic space from one view should agree with those from another view. Formally, Co-PLSA can be modeled as maximizing the following objective function:

$$O(\varphi^v, \varphi^w) = \tau^v L(\varphi^v) + \tau^w L(\varphi^w) - \lambda R \tag{11}$$

where $L(\varphi^v)$ and $L(\varphi^w)$ are the log-likelihood functions of PLSA on view V and view W respectively. The parameter $\tau^v(\tau^w)$ is the weight of view $V(W)$ which satisfies the constraint $\tau^v + \tau^w = 1$. λ is utilized to trade off the two PLSA log-likelihood objectives and the pairwise co-regularization R is to bridge these two individual views together. Fig. 4 illustrates the structure of Co-PLSA model in two-view case. Note that the small circles with the same color means Co-PLSA encourages the pairwise similarities based on topic distribution as consistent as possible across different views. In the meanwhile, Zhou and Luo [23] bring forward a geo-topic extraction framework for geolocation inference, including location name entity recognition, location related image association and a multimodal location dependent PLSA geo-topic model. Besides, Kim et al.[24] propose a novel scene classification method that represents a scene as a latent aspect distribution using the probabilistic latent semantic analysis with visterm spatial location and determines the class label of input image's latent aspect distributions based on the support vector machine.

4. **Improvements of PLSA.** This section will summarize the improvements of PLSA model itself, including its initialization methods, visual words construction, adding hidden layers and integration with other models, respectively.

4.1. **To improve the initialization.** An important consideration in PLSA modeling is the performance of the model which is strongly affected by the initialization of the model prior to training. Thus a method for identifying a good initialization or alternatively a good trained model is in urgent need. Since the expectation maximization (EM) algorithm

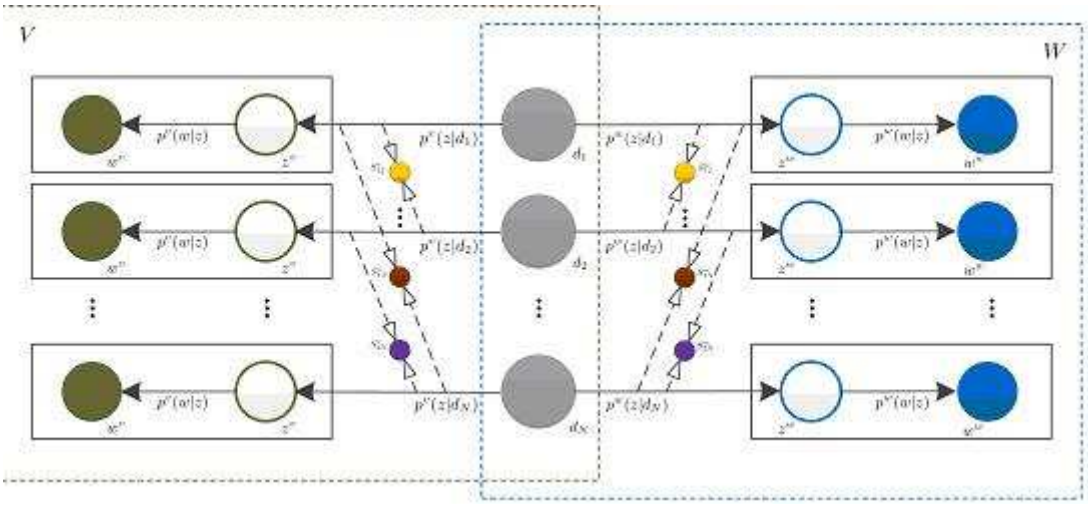


FIGURE 4. The structure of Co-PLSA model

used to train the PLSA model is sensitive to its initialization. Hence, the early notable research mainly focuses on finding a good way to initialize the PLSA model. As the representative work, Farahat et al.[25] present a framework for using latent semantic analysis (LSA) to better initialize the parameters of a corresponding PLSA model, and the EM algorithm is then used to further refine the initial estimate. Subsequently, Rodner and Denzler [26] propose to use an ensemble of PLSA models that are trained using random fractions of the training data. They analyze empirically the influence of the degree of randomization and the size of the ensemble on the overall classification performance of a scene recognition task. Recently, Lu et al.[20] use rival penalized competetive learning method to initialize the PLSA model. What's more, they have made a direct comparison with the results obtained by the PLSA model with LSA based initialization, random initialization and RPCL based initialization, respectively.

4.2. **To improve the visual words.** It is well known that the performance of image annotation heavily depends on the image feature representation. In recent years, the bag-of-visual-words model [27] has been successfully applied in multimedia community and shown the promising performance. However, since the standard PLSA can only handle discrete quantity (such as textual words), this approach quantizes feature vectors into discrete visual words for PLSA modeling. Hence its annotation performance is sensitive to the clustering granularity. In addition, due to the positional relationship between image features is usually ignored, which will undoubtedly affect the performance of the PLSA model. Thus Sivic et al.[28] develop the doublet visual words to encode spatially local co-occurring image regions for image segmentation. For more details of this model, please refer to the corresponding sections described in [27]. Horster et al.[29] present three versions of PLSA with different continuous visual word models such as Gaussian mixture models, shared Gaussian words and fixed shared Gaussian words (abbreviated as GM-PLSA, SGW-PLSA and FSGW-PLSA respectively) instead of the discrete quantized high-dimensional descriptors. Note that the joint probabilities of the generative process for GM-PLSA and SGM-PLSA models, by introducing a multivariate Gaussian mixture

over the feature space for each latent topic, can be defined as follows:

$$P(d_i, w_j) = P(d_i) \sum_{h=1}^{H} \left(P(z_h|d_i) \sum_{k=1}^{K} \pi_{kh} \times N(w_j|\mu_{kh}, \Sigma_{kh}) \right) \tag{12}$$

$$P(d_i, w_j) = P(d_i) \sum_{h=1}^{H} \sum_{k=1}^{K} P(w_j|g_k) P(g_k|z_h) P(z_h|d_i) \tag{13}$$

where $P(w_j|g_k) = N(w_j|\mu_k, \Sigma_k)$, H and K denote the total number of the topics and the Gaussian words in the model, respectively.

Motivated by the ideas of literature [29], Wu et al.[30] employ PLSA model to region-based image classification and propose two soft vector quantization methods to tackle the small sample problem in visual vocabulary construction. Afterwards, Wang et al.[31] propose a method to build an effective visual vocabulary by using hierarchical Gaussian mixture model instead of traditional clustering methods. In addition, PLSA is employed to explore semantic aspects of visual concepts and to discover topic clusters among documents and visual words so that each image can be projected on to a lower dimensional topic space for more efficient and effective annotation. In the area of AIA, it is generally believed that using continuous image feature vectors will result in better performance. To this end, Li et al.[32] propose a continuous PLSA that assumes the feature vectors in an image are governed by a Gaussian distribution under a given latent aspect rather than a multinomial one, and it can be viewed as an effective extension of work [29]. Different from Eq.(7), it should be noted that the conditional probability density function of element w_j in [32] is defined as follows:

$$P(w_j|z_k) = \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} \exp\{-\frac{1}{2}(w_j - \mu_k)^T \Sigma_k^{-1}(w_j - \mu_k)\} \tag{14}$$

where d is the dimension, μ_k is a d-dimensional mean vector and Σ_k is a $d \times d$ covariance matrix. The representation of this model in terms of a graphical model can be depicted in Fig. 5.

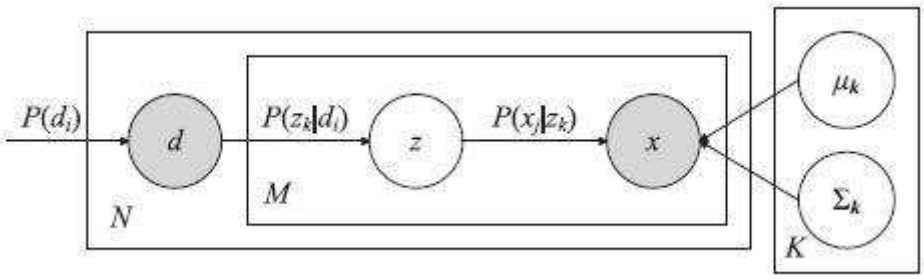


FIGURE 5. The structure of continuous PLSA model

4.3. **To add the hidden layers.** Inspired by the ideas of neuroscience, Lienhart et al.[33,34] extend the standard single-layer probabilistic latent semantic analysis to multiple multimodal layers, denoted by MM-PLSA, which consists of two leaf-PLSAs (here from two different data modalities: image tags and visual image features) and a single top-level PLSA node merging the two leaf-PLSAs. Especially the proposed fast initialization technique, stepwise forward procedure, makes the MM-PLSA very practical and computable. It's worth noting that the smallest possible multi-layer multimodal PLSA

model considering two modes with their respective observable word occurrences and hidden topics as well as a single top-level of hidden aspects is graphically depicted in Fig. 6 (left). In [35], the standard PLSA model is extended by integrating an additional variable associated with the time stamp to better model the temporal topic trends among images from an unified perspective, in which the time stamp is the added hidden layer to PLSA. The joint probability can be formally expressed as below:

$$P(d_i, w_j) = \sum_{z_k} \sum_{t_s} P(d_i)P(z_k|d_i)P(t_s)P(w_j|z_k, t_s) \tag{15}$$

In addition, Li et al.[36] propose a correlated probabilistic latent semantic analysis (C-PLSA) model by introducing a correlation layer between images and latent topics to incorporate the image correlations, in which the correlations are parameterized by the image correlation matrix C as is shown in the dashed box in Fig. 6 (right).

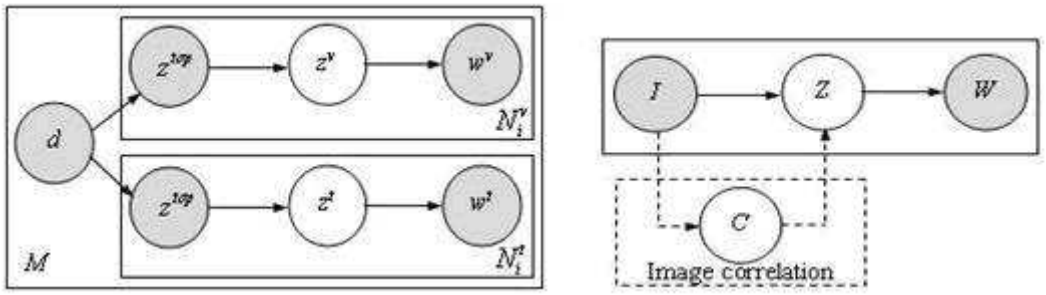


FIGURE 6. The structures of MM-PLSA (left) and C-PLSA (right) models

Different from the methods with explicit hidden layers added by the PLSA model, a novel two-probabilistic latent semantic analysis (two-PLSA) model has been proposed in the work of [37] based on two hidden random variables, in which the first latent aspect is used for representing the images on a corpus relate their words, and the second latent aspect is used for representing visual features of each image associated with their words. The two-PLSA can support multi-functionality, including not only image retrieval function but also automatic image annotation for image retrieval systems. Its graphical model is shown in Fig. 7.

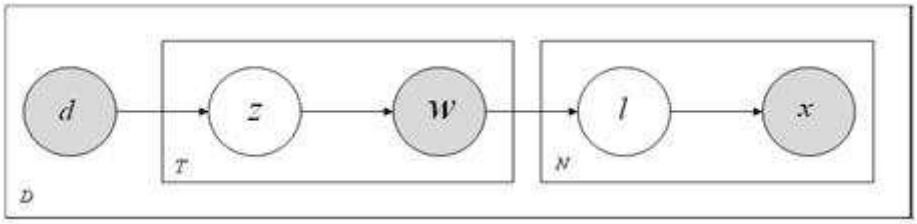


FIGURE 7. The structure of two-PLSA model

4.4. **To integrate with other models.** In general, commonly used methods in semantic image analysis community can be roughly divided into two categories: generative methods and discriminative methods. Concretely, the discriminative model treats semantic image analysis as a classification problem by considering each semantic concept or keyword as

a class and building concept classifiers to predict the relevance between images and given concepts, while the generative method addresses semantic image analysis by constructing the probabilistic relevance between visual features and textual description. Both of them have their own advantages and disadvantages. It is generally believed that the more effective approaches can be constructed by integrating these two kinds of methods simultaneously, and most of the existing studies related to the PLSA model demonstrate their complementary nature [16,18,20,24,38,39,40,41,45,46]. Based on this fact, Lu et al.[20] integrate PLSA model with the rival penalized competitive learning initialization as well as ensemble-based SVM for image categorization, which is robust to the changes of the visual vocabulary and the number of latent topics. Quelhas et al.[16] combine PLSA with SVM for scene classification, especially they have used the ability of PLSA to generate a compact scene representation and to automatically extract visually meaningful aspects for aspect-based image ranking and context-sensitive image segmentation, respectively. Besides, the task of scene classification has been done by integrating PLSA with k-nearest neighbor classifier, support vector machine, and multi-instance multi-label learning [18,24,39]. Note that in [38], PLSA is combined with visual attention model to create AM-PLSA, in which the attention model is used to detect salient regions and non-salient regions in an image to alleviate the influence of background clutter to object. However, this kind of algorithm just adds a preprocessing to probabilistic latent semantic analysis and does not change the essence of the PLSA model itself. Followed by Ergul and Arica [43] fuse spatial pyramid matching and probabilistic latent semantic analysis for scene classification (called as cascaded PLSA), which performs PLSA in a hierarchical sense after the soft-weighted bag-of-words histograms representation based on the dense local features is extracted. Fig. 8 illustrates the hierarchy structure of the cascaded PLSA model.

In recent years, the combination of PLSA with canonical correlation analysis is used to develop image annotation systems [44]. Meanwhile, PLSA based topo-Markov random field is proposed for synthetic aperture radar image classification [45]. In the most recent years, a unified two-stage image annotation framework is constructed in [46], which is implemented by integrating a probabilistic latent semantic analysis with random walk. Specifically, a PLSA with asymmetric modalities is first exploited to accomplish the initial image annotation. The random walk process over the constructed label similarity graph is then implemented to further mine the correlation among the candidate annotations generated by the PLSA. More details can be gleaned from the corresponding literature. In a word, the integration of PLSA with other methods is easy to understand and many of them are able to obtain superior performance. However, an important consideration for this integration is that the trade-off between computational complexity and model reconstruction error. To a large extent, it heavily depends on the real-world applications.

5. **Conclusions and Future Work.** In this paper, we have made a comprehensive review on the PLSA model for semantic image analysis in literature. To be specific, we have focused on two major aspects to describe it. On one side, the PLSA with applications to automatic image annotation, image retrieval, image classification and several other applications are systematically summarized. On the other side, we have provided a comprehensive review on PLSA itself improvement, including its initialization methods, visual words construction, adding hidden layers and integration with other methods. All of them are able to obtain better performance for different tasks from different point of view. The ultimate goal of this paper is to attempt to look into them through a unified view, which may help grasp the essentials of these probabilistic latent semantic analysis models for other researchers. Meanwhile, this paper also illustrates the pros and cons of

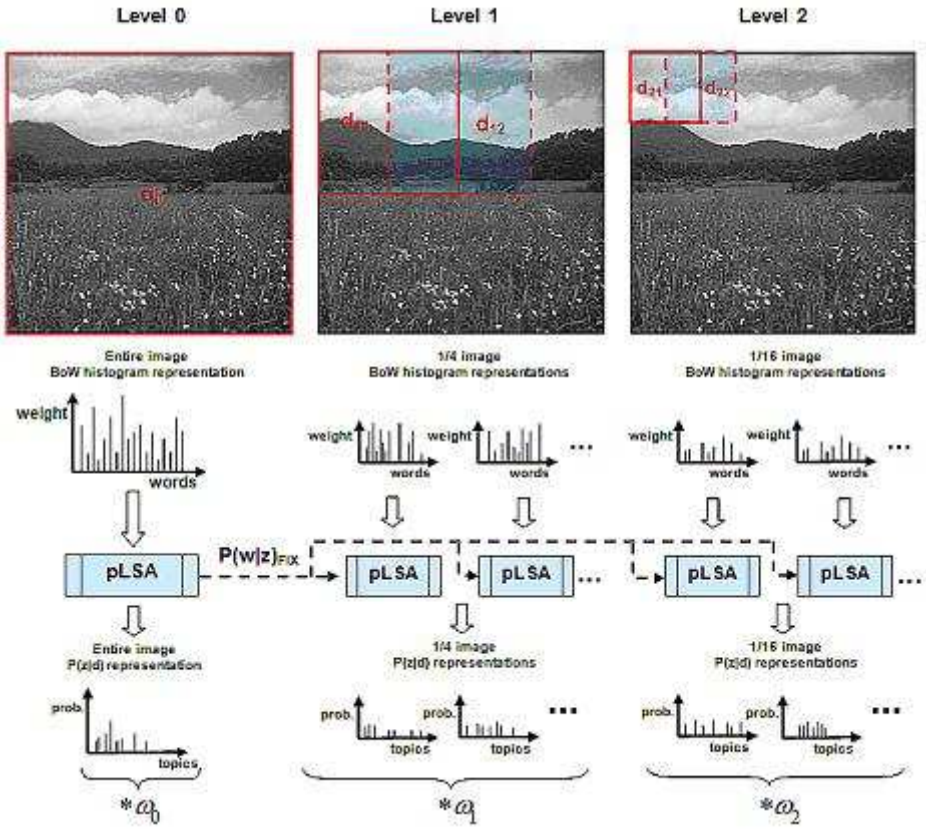


FIGURE 8. The structure of the cascaded PLSA model

PLSA combined with a great deal of existing research as well as to point out the promising research directions for semantic image analysis in the future.

However, there are still several major issues to be further explored for PLSA model. First, it is well known that due to the fact that the EM algorithm is sensitive to the initialization, an important consideration for the PLSA model trained via EM is that the performance of the model is strongly affected by the initialization of the model. Hence, how to formulate a method to identify a good initialization for PLSA model is in urgent need. Second, since 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 may make this region-based feature representation imprecise and therefore undermine the performance of the PLSA-based approaches. So to explore more efficient image segmentation methods is helpful to boost the performance. What is more, image segmentation itself is a worthy of further research direction. Third, since the standard PLSA can only handle discrete quantities (such as textual words), that is to say, PLSA usually quantizes feature vectors into discrete visual words for modeling, thus its performance is sensitive to the clustering granularity. So how to construct an effective representation of visual words is a promising research direction. Fourth, due to the lack of commonly acceptable image databases for PLSA related models evaluation, which results in the phenomenon that different PLSA related approaches use different image datasets for their performance evaluation and thus it is difficult to make a fair comparison with each other. Therefore, some standard image datasets are expected to

be created for researches in the future. Last but not the least, due to the complementary performance of hybridizing two or more machine learning techniques together, which can usually make them benefit from each other. Based on this recognition, how to efficiently integrate PLSA with other methods based on the trade-off between computational complexity and model reconstruction error is a valuable research direction in the future. In addition, as for future work, PLSA model should be applied in a wider selection of practical machine learning domain to deal with more multimedia related tasks, such as speech recognition, action recognition, music information retrieval and other multimedia event detection tasks, etc. At the same time, it is worth noting that the parallelization of PLSA model to very large-scale image datasets is also an important issue to be further studied, especially in the current circumstances of cloud computing, cloud services, hadoop, smartwatch, fingerprint password, web of things, 3D printing and deep learning techniques, etc.

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 for stimulating discussions and helpful hints. This work is partially supported by the National Program on Key Basic Research Project (No.2013CB329502), the National Natural Science Foundation of China (No.61202212) the Key R&D Program of the Shaanxi Province of China (No.2018GY-037), and the Special Research Project of the Educational Department of Shaanxi Province of China (No.18JK1038).

REFERENCES

[1] T. Hofmann, Unsupervised learning by probabilistic latent semantic analysis, *Machine Learning*, vol. 42, no. 1, pp. 177–196, 2001.

[2] D. Blei, A. Ng and M. Jordan, Latent Dirichlet allocation, *Journal of Machine Learning Research*, vol. 3, no. 4-5, pp. 993–1022, 2003.

[3] D. Blei and J. Lafferty, Correlated topic models, *Annals of Applied Statistics*, vol. 1, no. 1, pp. 17–35, 2007.

[4] F. Monay and D. Gatica-Perez, On image auto-annotation with latent space models, *Proc. of the 11th Int'l Conf. on Multimedia (MM'03)*, pp. 275–278, 2003.

[5] F. Monay and D. Gatica-Perez, PLSA-based image auto-annotation: constraining the latent space, *Proc. of the 12th Int'l Conf. on Multimedia (MM'04)*, pp. 348–351, 2004.

[6] 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.

[7] 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.

[8] Y. Peng, Z. Lu and J. Xiao, Semantic concept annotation based on audio PLSA model, *Proc. of the 17th Int'l Conf. on Multimedia (MM'09)*, pp. 841–844, 2009.

[9] 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.

[10] D. Tian, Semantic image annotation based on robust probabilistic latent semantic analysis, *Journal of Information Hiding and Multimedia Signal Processing*, vol. 8, no. 1, pp. 228–237, 2017.

[11] A. Shah-hosseini and G. Knapp, Semantic image retrieval based on probabilistic latent semantic analysis, *Proc. of the 14th Int'l Conf. on Multimedia (MM'06)*, pp. 703–706, 2006.

[12] R. Lienhart and M. Slaney, PLSA on large scale image databases, *Proc. of the 32nd Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP'07)*, pp. 1217–1220, 2007.

[13] S. Romberg, E. Horster and R. Lienhart, Multimodal PLSA on visual features and tags, *Proc. of the Int'l Conf. on Multimedia and Expo (ICME'09)*, pp. 414–417, 2009.

[14] 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.

[15] 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.

[16] P. Quelhas, F. Monay, J. Odobez, et al., Modeling scenes with local descriptors and latent aspects, *Proc. of the 10th Int'l Conf. on Computer Vision (ICCV'05)*, pp. 883–890, 2005.
[17] A. Bosch, A. Zisserman and X. Munoz, Scene classification using a hybrid generative/discriminative approach, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 4, pp. 712–727, 2008.
[18] A. Bosch, A. Zisserman and X. Munoz, Scene classification via PLSA, *Proc. of the 9th European Conf. on Computer Vision (ECCV'06)*, pp. 517–530, 2006.
[19] L. Zhuang, L. She, Y. Jiang, et al., Image classification via semi-supervised PLSA, *Proc. of the 5th Int'l Conf. on Image and Graphics (ICIG'09)*, pp. 205–208, 2009.
[20] Z. Lu, Y. Peng and H. Ip, Image categorization via robust PLSA, *Pattern Recognition Letters*, vol. 31, no. 1, pp. 36–43, 2010.
[21] 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.
[22] 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.
[23] Y. Zhou and J. Luo, Geo-location inference on news articles via multimodal PLSA, *Proc. of the 20th Int'l Conf. on Multimedia (MM'12)*, pp. 741–744, 2012.
[24] S. Kim and D. Kim, Scene classification using PLSA with visterm spatial location, *Proc. of the 1st Int'l Workshop on Interactive Multimedia for Consumer Electronics (IMCE'09)*, pp. 57–66, 2009.
[25] A. Farahat and F. Chen, Improving probabilistic latent semantic analysis with principal component analysis, *Proc. of the 11th Conf. of the European Chapter of the Association for Computational Linguistics (EACL'06)*, pp. 105–112, 2006.
[26] 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.
[27] A. Bosch, X. Munoz and R. Marti, Which is the best way to organize/classify images by content? *Image and Vision Computing*, vol. 25, no. 6, pp. 778–791, 2007.
[28] J. Sivic, B. Russell, A. Efros, et al., Discovering objects and their localization in images, *Proc. of the 10th Int'l Conf. on Computer Vision (ICCV'05)*, pp. 370–377, 2005.
[29] E. Horster, R. Lienhart and M. Slaney, Continuous visual vocabulary models for PLSA-based scene recognition, *Proc. of the 7th Int'l Conf. on Image and Video Retrieval (CIVR'08)*, pp. 319–328, 2008.
[30] 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.
[31] 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.
[32] Z. Li, Z. Shi, X. Liu, et al., Modeling continuous visual features for semantic image annotation and retrieval, *Pattern Recognition Letters*, vol. 32, no. 3, pp. 516–523, 2011.
[33] R. Lienhart, S. Romberg and E. Horster, Multilayer PLSA for multimodal image retrieval, *Proc. of the 8th Int'l Conf. on Image and Video Retrieval (CIVR'09)*, pp. 1–8, 2009.
[34] 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.
[35] X. Liao, Y. Wang and L. Ding, Discovering temporal patterns from images using extended PLSA, *Proc. of the 1st Int'l Conf. on Multimedia Technology (ICMT'10)*, pp. 1–7, 2010.
[36] 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.
[37] N. Watcharapinchai, S. Aramvith and S. Siddhichai, Two-probabilistic latent semantic model for image annotation and retrieval, *Proc. of the 10th Asian Conf. on Computer Vision Workshop (ACCVW'10)*, pp. 359–369, 2010.
[38] 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.
[39] 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.
[40] D. Tian, X. Zhao and Z. Shi, Fusing PLSA model and Markov random fields for automatic image annotation, *High Technology Letters*, vol. 20, no. 4, pp. 409–414, 2014.

[41] D. Tian, Exploiting semantic context relationships for automatic image annotation, *Journal of Information Hiding and Multimedia Signal Processing*, vol. 8, no. 6, pp. 1203–1217, 2017.

[42] 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.

[43] 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.

[44] Y. Zheng, T. Takiguchi and Y. Ariki, 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.

[45] M. Liu, C. He, X. Su, et al., A PLSA based on topo-MRF model method for SAR images classification, *Geomatics and Information Science of Wuhan University*, vol. 36, no. 1, pp. 122–125, 2011.

[46] 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.