A Novel Model for Semantic Learning and Retrieval of Images

Zhixin Li, Zhiping Shi, Zhengjun Tang, Weizhong Zhao

HAL Id: hal-01524960
https://inria.hal.science/hal-01524960

Submitted on 19 May 2017

A Novel Model for Semantic Learning and Retrieval of Images

Zhixin Li[1], ZhiPing Shi[2], ZhengJun Tang[1], Weizhong Zhao[3]

[1] College of Computer Science and Information Technology,
Guangxi Normal University, Guilin 541004, China
[2] College of Information Engineering, Capital Normal University, Beijing 100048, China
[3] College of Information Engineering, Xiangtan University, Xiangtan 411105, China
`lizx@gxnu.edu.cn, shizhiping@gmail.com, zjtang@gxnu.edu.cn,`
`zhaoweizhong@gmail.com`

Abstract. In this paper, we firstly propose an extended probabilistic latent semantic analysis (PLSA) to model continuous quantity. In addition, corresponding EM algorithm is derived to determine the parameters. Then, we apply this model in automatic image annotation. In order to deal with the data of different modalities according to their characteristics, we present a semantic annotation model which employs continuous PLSA and traditional PLSA to model visual features and textual words respectively. These two models are linked with the same distribution over all aspects. Furthermore, an asymmetric learning approach is adopted to estimate the model parameters. This model can predict semantic annotation well for an unseen image because it associates visual and textual modalities more precisely and effectively. We evaluate our approach on the Corel5k and Corel30k dataset. The experiment results show that our approach outperforms several state-of-the-art approaches.

Keywords: semantic learning; automatic image annotation; continuous PLSA; aspect model; image retrieval

1 Introduction

As an important research issue, Content-based image retrieval (CBIR) searches relative images of given example in visual level. Under this paradigm, various low-level visual features are extracted from each image in the database and image retrieval is formulated as searching for the best database match to the feature vector extracted from the query image. Although this process is accomplished quickly and automatically, the results are seldom semantically relative to the query example due to the notorious semantic gap [5]. As a result, automatic image annotation has emerged as a crucial problem for semantic image retrieval.

The state-of-the-art techniques of automatic image annotation can be categorized into two different schools of thought. The first one defines auto-annotation as a traditional supervised classification problem [4, 12], which treats each word (or semantic category) as an independent class and creates different class models for every word.

This approach computes similarity at the visual level and annotates a new image by propagating the corresponding class words. The second perspective takes a different stand and treats image and text as equivalent data. It attempts to discover the correlation between visual features and textual words on an unsupervised basis [1, 2, 7, 8 10, 11, 13, 15], by estimating the joint distribution of features and words. Thus, it poses annotation as statistical inference in a graphical model.

As latent aspect models, both PLSA [9] and latent Dirichlet allocation (LDA) [3] have been successfully applied to annotate and retrieve images. PLSA-WORDS [15] is a representative approach, which acquires good annotation performance by constraining the latent space. However, this approach quantizes continuous feature vectors into discrete visual words for PLSA modeling. As a result, its annotation performance is sensitive to the clustering granularity. In our previous work [13], we propose PLSA-FUSION which employs two PLSA models to capture semantic information from visual and textual modalities respectively. Furthermore, an adaptive asymmetric learning approach is presented to fuse the aspects of these two models. Consequently, PLSA-FUSION can acquire higher accuracy than PLSA-WORDS. Nevertheless, PLSA-FUSION also needs to quantize visual features into discrete visual words for PLSA modeling. In the area of automatic image annotation, it is generally believed that using continuous feature vectors will give rise to better performance [2, 11]. However, since traditional PLSA is originally applied in text classification, it can only handle discrete quantity (such as textual words). In order to model image data precisely, it is required to deal with continuous quantity using PLSA.

This paper proposes continuous PLSA, which assumes that each feature vector in an image is governed by a Gaussian distribution under a given latent aspect other than a multinomial one. In addition, corresponding EM algorithm is derived to estimate the parameters. Then, as general treatment, each image can be viewed as a mixture of Gaussians under this model. Furthermore, based on the continuous PLSA and the traditional PLSA, we present a semantic annotation model to learn the correlation between the visual features and textual words. An asymmetric learning approach is adopted to estimate the model parameters. We evaluate our approach on standard Corel datasets and the experiment results show that our approach outperforms several state-of-the-art approaches.

The rest of the paper is organized as follows. Section 2 presents the continuous PLSA model and derives corresponding EM algorithm. Section 3 proposes a semantic annotation model and describes the asymmetric learning approach. Experiment results are reported and analyzed in section 4. Finally, the overall conclusions of this work are presented in section 5.

2 Continuous PLSA

Just like traditional PLSA, continuous PLSA is also a statistical latent class model which introduces a hidden variable (latent aspect) z_k ($k \in 1, \ldots, K$) in the generative process of each element x_j ($j \in 1, \ldots, M$) in a document d_i ($i \in 1, \ldots, N$). However, given this unobservable variable z_k, continuous PLSA assumes that elements x_j are sampled

from a multivariate Gaussian distribution, instead of a multinomial one in traditional PLSA. Using these definitions, continuous PLSA [14] assumes the following generative process:

1. Select a document d_i with probability $P(d_i)$;
2. Sample a latent aspect z_k with probability $P(z_k|d_i)$ from a multinomial distribution conditioned on d_i;
3. Sample $x_j \sim P(x_j|z_k)$ from a multivariate Gaussian distribution $N(\mathbf{x}|\boldsymbol{\mu_k},\boldsymbol{\Sigma_k})$ conditioned on z_k.

Continuous PLSA has two underlying assumptions. First, the observation pairs (d_i, x_j) are generated independently. Second, the pairs of random variables (d_i, x_j) are conditionally independent given the latent aspect z_k. Thus, the joint probability of the observed variables is obtained by marginalizing over the latent aspect z_k,

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

A representation of the model in terms of a graphical model is depicted in Figure 1.
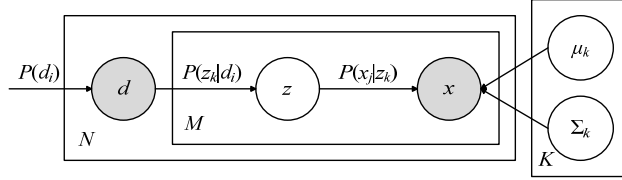
Fig. 1. Graphical model representation of continuous PLSA.

The mixture of Gaussian is assumed for the conditional probability P(.|z). In other words, the elements are generated from K Gaussian distributions, each one corresponding a z_k. For a specific latent aspect z_k, the condition probability distribution function of elements x_j is

$$P(x_j \mid z_k) = \frac{1}{(2\pi)^{D/2} \mid \Sigma_k \mid^{1/2}} \exp(-\frac{1}{2}(x_j - \mu_k)^T \Sigma_k^{-1}(x_j - \mu_k)), \tag{2}$$

where D is the dimension, μ_k is a D-dimensional mean vector and Σ_k is a $D{\times}D$ covariance matrix.

Following the maximum likelihood principle, $P(z_k|d_i)$ and $P(x_j|z_k)$ can be determined by maximization of the log-likelihood function

$$\mathcal{L} = \sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i,\boldsymbol{x}_j)\log P(d_i,\boldsymbol{x}_j),$$
$$= \sum_{i=1}^{N} n(d_i)\log P(d_i) + \sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i,\boldsymbol{x}_j)\log \sum_{k=1}^{K} P(z_k \mid d_i)P(\boldsymbol{x}_j \mid z_k). \tag{3}$$

where $n(d_i, x_j)$ denotes the number of element x_j in d_i.

The standard procedure for maximum likelihood estimation in latent variable models is the EM algorithm [6]. In E-step, applying Bayes' theorem to (1), one can obtain

$$P(z_k \mid d_i, x_j) = \frac{P(z_k \mid d_i)P(x_j \mid z_k)}{\sum_{l=1}^{K} P(z_l \mid d_i)P(x_j \mid z_l)}. \tag{4}$$

In M-step, for any d_i, z_k and x_j, the parameters are determined as

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

$$\Sigma_k = \frac{\sum_{i=1}^{N} \sum_{j=1}^{M} n(d_i, x_j)P(z_k \mid d_i, x_j)(x_j - \mu_k)(x_j - \mu_k)^T}{\sum_{i=1}^{N} \sum_{j=1}^{M} n(d_i, x_j)P(z_k \mid d_i, x_j)}, \tag{6}$$

$$P(z_k \mid d_i) = \frac{\sum_{j=1}^{M} n(d_i, x_j)P(z_k \mid d_i, x_j)}{\sum_{j=1}^{M} n(d_i, x_j)}. \tag{7}$$

Alternating (4) with (5)–(7) defines a convergent procedure. The EM algorithm terminates by either a convergence condition or *early stopping* technique.

3   Semantic Learning Model

3.1   Gaussian-Multinomial PLSA

In order to deal with the data of different modalities in terms of their characteristics, we employ continuous PLSA and traditional PLSA to model visual features and textual words respectively. These two models are linked by sharing the same distribution over latent aspects $P(z|d)$. We refer to this semantic annotation model as *Gaussian-multinomial PLSA* (GM-PLSA), which is represented in Figure 2.
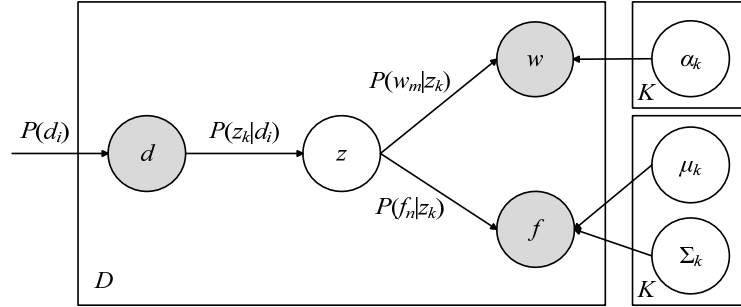
Fig. 2. Representation of the semantic annotation model GM-PLSA.

GM-PLSA assumes the following generative process:
1. Select a document d_i with probability $P(d_i)$;
2. Sample a latent aspect z_k with probability $P(z_k|d_i)$ from a multinomial distribution conditioned on d_i;

3. For each of the words, Sample $w_m \sim P(w_m|z_k)$ from a multinomial distribution Mult($\mathbf{x}|\boldsymbol{\alpha_k}$) conditioned on z_k.

4. For each of the feature vectors, Sample $f_n \sim P(f_n|z_k)$ from a multivariate Gaussian distribution $N(\mathbf{x}|\boldsymbol{\mu_k},\boldsymbol{\Sigma_k})$ conditioned on the latent aspect z_k.

Under this modeling approach, each image can be viewed as either a mixture of continuous Gaussian in visual modality or a mixture of discrete words in textual modality. Therefore, it can learn the correlation between features and words effectively and predict semantic annotation precisely for an unseen image.

3.2 Algorithm Description

We adopt asymmetric learning approach to estimate the model parameters because an asymmetric learning gives a better control of the respective influence of each modality in the latent space definition [15]. In this learning approach, textual modality is firstly chosen to estimate the mixture of aspects in a given document, which constrains the definition of latent space to ensure its consistency in textual words, while retaining the ability to jointly model visual features. The flow of learning and annotating is described in Figure 3.
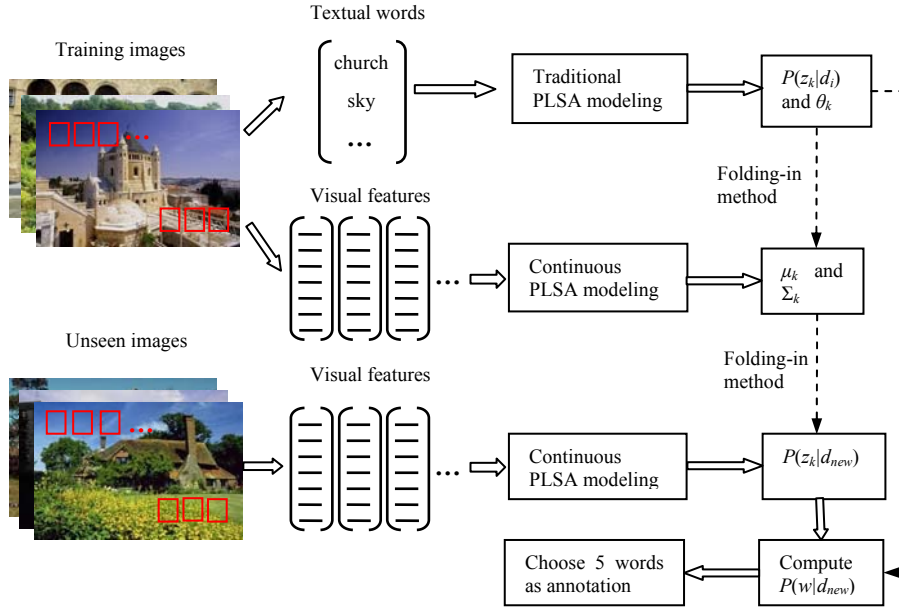

Fig. 3. The flow of learning and annotating algorithms of GM-PLSA.

In training stage, each training image is processed and represented as a bag of visual features and textual words. The aspect distributions $P(z_k|d_i)$ are firstly learned for all training documents from textual modality only. At the same time, the parameter $P(w_m|z_k)$ (i.e. $\boldsymbol{\alpha_k}$) is determined too. Then we use folding-in method to infer the parameters $\boldsymbol{\mu_k}$ and $\boldsymbol{\Sigma_k}$ for the visual modality with the aspect distributions $P(z_k|d_i)$ kept

fixed. Consequently, we can get the model parameters $\boldsymbol{\alpha}_k$, $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$, which remain valid in images out of the training set. The learning procedure is described in detail in Algorithm 1.

Algorithm 1. Estimation of the parameters: $\boldsymbol{\alpha}_k$、 $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$

Input: Visual features f_n and textual words w_m of training images.

Output: Model parameters $\boldsymbol{\alpha}_k$、 $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$.

Process:
1. random initialize the $P(z_k|d_i)$ and $P(w_m|z_k)$ probability tables;
2. while increase in the likelihood of validation data $\Delta L_t > T_t$ **do**
    (a) {E step}
        for $k \in 1, 2, \ldots, K$ and all (d_i, w_j) pairs in training documents **do**
        compute $P(z_k|d_i, w_j)$ with EM algorithm of traditional PLSA;
        end for
    (b) {M step}
        for $k \in 1, 2, \ldots, K$ and $j \in 1, 2, \ldots, M$ **do**
        compute $P(w_j|z_k)$ with EM algorithm of traditional PLSA;
        end for
        for $k \in 1, 2, \ldots, K$ and $i \in 1, 2, \ldots, N$ **do**
        compute $P(z_k|d_i)$ with EM algorithm of traditional PLSA;
        end for
    (c) compute the likelihood of validation data L_t;
    end while
3. save $\boldsymbol{\alpha}_k = \{P(w_1|z_k), P(w_2|z_k), \ldots, P(w_{Mw}|z_k)\}$;
4. initialize $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$;
5. **while** increase in the likelihood of validation data $\Delta L_c > T_c$ **do**
    (a) {E step}
        for $k \in 1, 2, \ldots, K$ and all $(d_i, \boldsymbol{f}_j)$ pairs in training documents **do**
        compute $P(z_k|d_i, \boldsymbol{f}_j)$ with eq.(4);
        end for
    (b) {M step}
        for $k \in 1, 2, \ldots, K$, $i \in 1, 2, \ldots, N$ and $j \in 1, 2, \ldots, M$ **do**
        compute $\boldsymbol{\mu}_k$ with eq.(7);
        compute $\boldsymbol{\Sigma}_k$ with eq.(8);
        end for
    (c) compute the likelihood of validation data L_c with eq.(3);
    end while
6. save $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$.

In annotation stage, given visual features of each test image and the previously estimated parameters $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$, the aspect distribution $P(z_k|d_{new})$ can be inferred using the folding-in method. The posterior probability of each word in the vocabulary is then computed by

$$P(w\,|\,d_{new}) = \sum_{k=1}^{K} P(z_k\,|\,d_{new}) P(w\,|\,z_k). \tag{8}$$

As usual, we choose five words with the largest posterior probabilities as annotations of an unseen image.Having estimated the parameters of GM-PLSA, the stage of semantic retrieval can be put into practice directly. The retrieval algorithm takes as inputs a semantic word w_q and a database of test images. After annotating each image in the test database, the retrieval algorithm ranks the images labeled with the query word by decreasing posterior probability $P(w_q|d_{new})$.

4   Experimental Results

In order to test the effectiveness and accuracy of the proposed approach, we conduct our experiments on two standard datasets, i.e. Corel5k[7] and Corel30k[4].

The focus of this paper is not on image feature selection and our approach is independent of visual features. We simply decompose images into a set of 32×32 blocks, then compute a 36 dimensional feature vector for each block, consisting of 24 color features and 12 texture features.

4.1   Results on Corel5k dataset

In this section, the performance is evaluated by comparing the captions automatically generated with the original manual annotations. We compute the recall and precision of every word in the test set and use the mean of these values to summarize the system performance.

We report the results 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. The systematic evaluation results are shown in Table 1. From the table, we can see that our model performs much better than the first three approaches. Besides, the model performs slightly better than MBRM. We believe that the application of the continuous PLSA is the reason for this result.

Table 1. Performance comparison on the task of automatic image annotation.

Models	Translation	CMRM	CRM	MBRM	PLSA-WORDS	GM-PLSA
#words with recall > 0	49	66	107	122	105	125
Results on 49 best words, as in [7,8,10,11]						
Mean Recall	0.34	0.48	0.70	0.78	0.71	0.79
Mean Precision	0.20	0.40	0.59	0.74	0.56	0.76
Results on all 260 words						
Mean Recall	0.04	0.09	0.19	0.25	0.20	0.25
Mean Precision	0.06	0.10	0.16	0.24	0.14	0.26

Several examples of annotation obtained by our prototype system are shown in Table 2. Here top five words are taken as annotation of the image. We can see that even the system annotates an image with a word not contained in the ground truth, this annotation is frequently plausible.

Table 2. Comparison of annotations made by PLSA-WORDS and GM-PLSA.

Image				
Ground Truth	mule, deer, buck, doe	sphinx, stone, statue, sculpture	branches, grass, frost, ice	people，sand，water，sky
PLSA-WORDS Annotation	deer,mule,sculpture, stone, rabbit	sculpture, stone, sky, hut, statue	clouds, fox, frost, snow, ice	sky, beach, coast, sand, oahu
GM-PLSA Annotation	deer, mule, grass, sculpture, sand	stone, sculpture, statue, sky, sand	ice，frost，snow，clouds，branches	beach, water, sky, sand, coast

Mean average precision (mAP) is employed as a metric to evaluate the performance of single word retrieval. We only compare our model with CMRM, CRM, MBRM and PLSA-WORDS, because mAPs of other models cannot be accessed directly from the literatures.

The annotation results ignore rank order. However, rank order is very important for image retrieval. Our system will return all the images which are automatically annotated with the query word and rank the images according to the posterior probabilities of that word. Table 3 shows that GM-PLSA is slightly better than other models.

Table 3. Comparison of ranked retrieval results

Mean Average Precision on Corel5k Dataset		
Models	All 260 words	Words with recall ≥ 0
CMRM	0.17	0.20
CRM	0.24	0.27
MBRM	0.30	0.35
PLSA-WORDS	0.22	0.26
GM-PLSA	0.32	0.37

In summary, the experiment results show that GM-PLSA outperforms several previous models in many respects, which proves that the continuous PLSA is effective in modeling visual features.
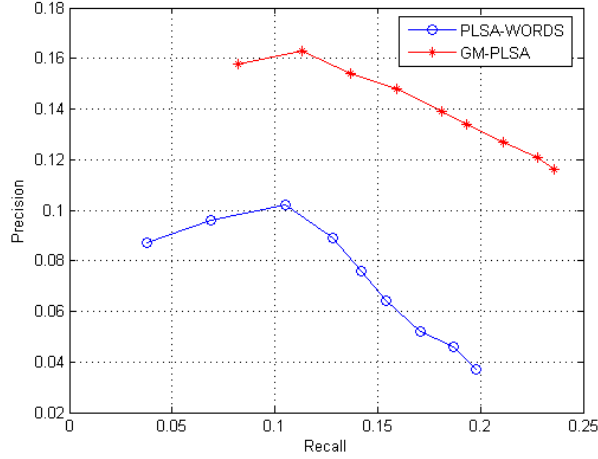
4.2   Results on Corel30k dataset

The Corel30k dataset provides a much large database size and vocabulary size compared with Corel5k. Since Corel30k is a new database, we only compare our model with PLSA-WORDS.

Figure 4 presents the precision-recall curves of PLSA-WORDS and GM-PLSA on the Corel30k dataset, with the number of annotations from 2 to 10. The precision and

recall values are the mean values calculated over all words. From the figure we can see that GM-PLSA consistently outperforms PLSA-WORDS.

The superior performance of GM-PLSA on precision and recall directly results in its great semantic retrieval performance. From Table 4 we can also see the great improvements of GM-PLSA over PLSA-WORDS.

Fig. 4. Precision-recall curves of PLSA-WORDS and GM-PLSA.

Table 4. Comparison of ranked retrieval results

Mean Average Precision on Corel30k Dataset		
Models	All 950 words	Words with recall $\geq$ 0
PLSA-WORDS	0.14	0.17
GM-PLSA	0.23	0.28

Overall, the experiments on Corel30k indicate that GM-PLSA is fairly stable with respect to its parameters setting. Moreover, since this annotation model integrates traditional PLSA and continuous PLSA, it has better robustness and scalability.

5   Conclusion

In this paper, we have proposed continuous PLSA to model continuous quantity and develop an EM-based iterative procedure to learn the conditional probabilities of the continuous quantity given a latent aspect. Furthermore, we present a semantic annotation model, which employ continuous PLSA and traditional PLSA to deal with the visual and textual data respectively. An adaptive asymmetric learning approach is adopted to estimate the model parameters. Experiments on the Corel dataset prove that our approach is promising for automatic image annotation. In comparison to pre-

vious proposed annotation methods, higher accuracy and effectiveness of our approach are reported.

6 Acknowledgments

7 References

1. K. Barnard, P. Duygulu, D. Forsyth, et al. Matching words and pictures. *Journal of Machine Learning Research*, 3: 1107–1135, 2003.
2. D.M. Blei, M.I. Jordan. Modeling annotated data. In: *Proc. 26th Intl. ACM SIGIR Conf.*, pp. 127–134, 2003.
3. D.M. Blei, A.Y. Ng, M.I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3: 993–1022, 2003.
4. G. Carneiro, A.B. Chan, P.J. Moreno, N. Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *IEEE Trans. PAMI*, 29(3): 394–410, 2007.
5. R. Datta, D. Joshi, J. Li, J.Z. Wang. Image retrieval: ideas, influences, and trends of the new age. *ACM Computing Surveys*, 40(2): article 5, 1–60, 2008.
6. A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1): 1–38, 1977.
7. P. Duygulu, K. Barnard, N. de Freitas, D. Forsyth. Object recognition as machine translation: learning a lexicon for a fixed image vocabulary. In: *Proc. 7th ECCV*, pp. 97–112, 2002.
8. S.L. Feng, R. Manmatha, V. Lavrenko. Multiple Bernoulli relevance models for image and video annotation. In: *Proc. CVPR*, pp. 1002–1009, 2004.
9. T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1–2): 177–196, 2001.
10. J. Jeon, V. Lavrenko, R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In: *Proc. 26th Int'l ACM SIGIR Conf.*, pp. 119–126, 2003.
11. V. Lavrenko, R. Manmatha, J. Jeon. A model for learning the semantics of pictures. In: *Proc. NIPS*, pp. 553–560, 2003.
12. J. Li, J.Z. Wang. Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Trans. PAMI*, 25(9): 1075–1088, 2003.
13. Zhixin Li, Zhiping Shi, Xi Liu, Zhiqing Li, Zhongzhi Shi. Fusing semantic aspects for image annotation and retrieval. *Journal of Visual Communication and Image Representation*, 2010, 21(8): 798–805.
14. Zhixin Li, Zhiping Shi, Xi Liu, Zhongzhi Shi. Automatic image annotation with continuous PLSA. In: *Proc. 35th ICASSP*, pp. 806–809, 2010.
15. F. Monay, D. Gatica-Perez. Modeling semantic aspects for cross-media image indexing. *IEEE Trans. PAMI*, 2007, 29(10): 1802–1817.