Automatic Image Annotation and Retrieval Using Hybrid Approach

Zhixin Li[1], Weizhong Zhao[2], Zhiqing Li[2], Zhiping Shi[3]

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

Abstract. We firstly propose continuous probabilistic latent semantic analysis (PLSA) to model continuous quantity. In addition, corresponding Expectation-Maximization (EM) algorithm is derived to determine the model parameters. Furthermore, we present a hybrid framework which employs continuous PLSA to model visual features of images in generative learning stage and uses ensembles of classifier chains to classify the multi-label data in discriminative learning stage. Since the framework combines the advantages of generative and discriminative learning, it can predict semantic annotation precisely for unseen images. Finally, we conduct a series of experiments on a standard Corel dataset. The experiment results show that our approach outperforms many state-of-the-art approaches.

Keywords: automatic image annotation; continuous PLSA; semantic learning; hybrid approach; image retrieval

1 Introduction

Content-based image retrieval (CBIR) has been studied and explored for decades. Its performance, however, is not ideal enough due to the notorious *semantic gap* [18]. CBIR retrieves images in terms of their visual features, while users often prefer intuitive text-based image searching. Since manual image annotation is expensive and difficult to be extended to large image databases, automatic image annotation has emerged as a striking and crucial problem [5].

The state-of-the-art techniques of automatic image annotation can be categorized into two different schools of thought. The first one is based on discriminative model. It defines auto-annotation as a traditional supervised classification problem [3,4,12,17], which treats each semantic concept as an independent class and creates different classifiers for different concepts. This approach computes similarity at the visual level and annotates a new image by propagating the corresponding words. The second perspective takes a different stand. It is based on generative model and treats image and text as equivalent data. It attempts to discover the correlation between

visual features and textual words on an unsupervised basis by estimating the joint distribution of features and words. Thus, it poses annotation as statistical inference in a graphical model. Under this perspective, images are treated as bags of words and features, each of which are assumed generated by a hidden variable. Various approaches differ in the definition of the states of the hidden variable: some associate them with images in the database [8,10,11], while others associate them with image clusters [1,7] or latent aspects (topics) [2,14,15]. Both these two kind of approaches have their own advantages and disadvantages. This paper will show that it is feasible to combine the advantages of these two formulations.

As a latent aspect model, PLSA [9] has been successfully applied to annotate and retrieve images. PLSA-WORDS [15] is a representative approach, which achieves the annotation task by constraining the latent space to ensure its consistency in words. However, since traditional PLSA can only handle discrete quantity (such as textual words), this approach quantizes feature vectors into discrete visual words for PLSA modeling. Therefore, its annotation performance is sensitive to the clustering granularity. In the area of automatic image annotation, it is generally believed that using continuous feature vectors will give rise to better performance [2,3,11]. In order to model image data precisely, it is required to deal with continuous quantity using PLSA.

This paper presents continuous PLSA, which assumes that feature vectors of images are 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, each image can be treated as a mixture of Gaussians under this model. Furthermore, we propose a hybrid framework to learn semantic classes of images. The framework employs continuous PLSA to model visual features of images in generative learning stage, and uses ensembles of classifier chains to classify the multi-label data in discriminative learning stage. We compare our approach with some state-of-the-art approaches on a standard Corel dataset and the experiment results show that our approach performs more effectively and precisely.

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 hybrid framework and describes the training and annotation procedure. 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 [13] 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.

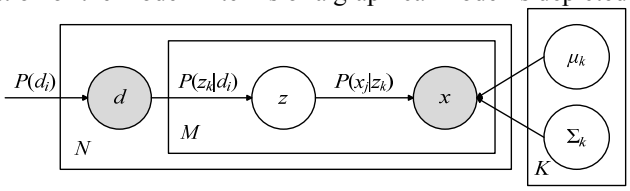


Fig. 1. Graphical model representation of continuous PLSA.

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

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

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

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

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

The standard procedure for maximum likelihood estimation 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.

As for the parameters, if parameter $P(x_j|z_k)$ is known, we could quickly infer the other parameters μ_k and Σ_k using folding-in method, and vice versa. Folding-in method is a partial version of the EM algorithm. It 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 Hybrid Generative/Discriminative Approach

3.1 Hybrid Framework

On the basis of continuous PLSA, we propose a hybrid framework which combines generative and discriminative learning. The framework firstly employs continuous PLSA to model visual features of images. As a result, each image can be represented as an aspect distribution. Then, this intermediate representation can be used to build ensembles of classifier chains, which can learn semantic classes of images and consider the correlation between the labels at the same time. The framework is shown in figure 2.

In training procedure, we firstly get the parameters μ_k and Σ_k given aspect z_k by modeling visual features of training images with continuous PLSA. At the same time, the aspect distribution $P(z_k|d_i)$ of each image is determined. This is the generative learning stage. The parameters μ_k and Σ_k are parameters of continuous PLSA. According to the independence assumption, these parameters remain valid for documents out of the training set. On the other hand, the aspect distribution $P(z_k|d_i)$ is only relative to the specific documents and cannot carry any prior information to an unseen image. This distribution, however, can represent each training image as a K-dimension vector. In addition, all the vectors can construct a simplex. Then, by making use of the aspect distribution and original annotation labels of each training image, we build a series of classifiers in which every word in the vocabulary is treated as an independent class. This is the discriminative learning stage. At this time, every image is represented as an aspect distribution, but has several semantic labels. This circumstance is

in conformity with multi-label learning, which can construct multiclass classifiers and integrate correlative information of textual words at the same time.

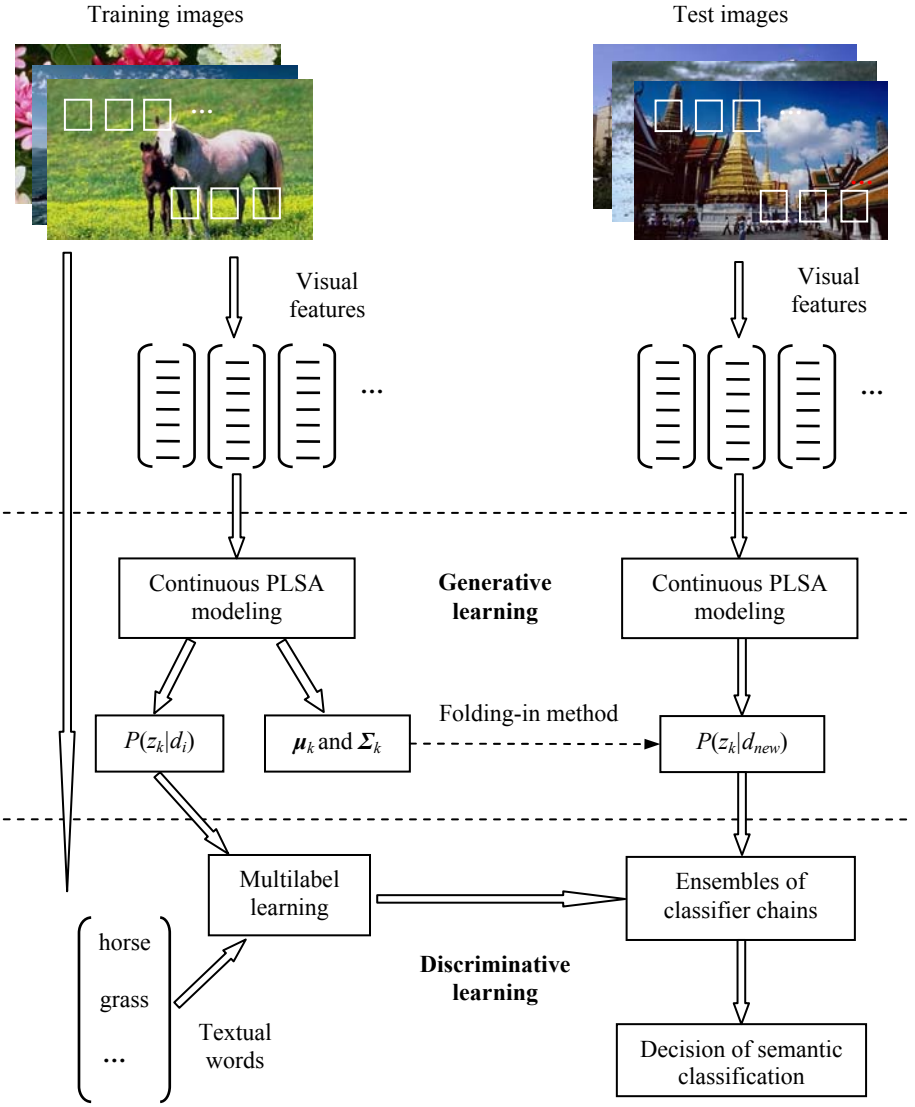


Fig. 2. Learning procedure of hybrid framework.

Correspondingly, there are two steps in annotation procedure. Firstly, since model parameters μ_k and Σ_k are determined in training procedure, we can compute the aspect distribution of each test image using folding-in method. Secondly, we classify the aspect distribution of each test image with the trained ensembles of classifier chains. Furthermore, we choose 5 words with highest confidence as annotations of the test

image. After each image in the database is annotated, the retrieval algorithm ranks the images labeled with the query word by decreasing confidence.

3.2 Ensembles of Classifier Chains

In discriminative learning stage, we employ ensembles of classifier chains [16] to accomplish the task of multi-label classification. Each binary classifier is implemented with SVM in classifier chains. Having taken the correlation between semantic labels into consideration, this approach can classify images into several semantic classes and it has higher confidence with acceptable computation complexity.

The classifier chain model involves $|L|$ binary classifiers, where L denotes the label set. Classifiers are linked along a chain where each classifier deals with the binary relevance problem associated with label $l_j \in L$. The feature space of each link in the chain is extended with the 0/1 label associations of all previous links. The training procedure is described in Algorithm 1. Note the notation for a training example $(\mathbf{x}, S)$, where $S \subseteq L$ and $\mathbf{x}$ is an instance feature vector.

Algorithm 1. Training procedure of classifier chain

Input: Example set $D = \{(\mathbf{x_1}, S_1), (\mathbf{x_2}, S_2), \ldots, (\mathbf{x_n}, S_n)\}$.

Output: Classifier chains $\{C_1, C_2, \ldots, C_{|L|}\}$.

Process:

1. **for** j∈1, 2, …, $|L|$
2. **do** single-label transformation and training
3. $D' \leftarrow \{\}$
4. **for** $(\mathbf{x}, S) \in D$
5. **do** $D' \leftarrow D' \cup ((\boldsymbol{x}, l_1, l_2, \ldots, l_{j-1}), l_j)$
6. Train C_j to predict binary relevance of l_j
7. C_j: $D' \rightarrow l_j \in \{0, 1\}$

Hence a chain $C_1, C_2, \ldots, C_j$ of binary classifiers is formed. Each classifier C_j in the chain is responsible for learning and predicting the binary association of label l_j given the feature space, augmented by all prior binary relevance predictions in the chain $l_1, l_2, \ldots, l_{j-1}$. The classification process begins at C_1 and propagates along the chain: C_1 determines $Pr(l_1|\mathbf{x})$ and every following classifier $C_2, \ldots, C_j$ predicts $Pr(l_j|\boldsymbol{x}_i, l_1, l_2, \ldots, l_{j-1})$. This classification procedure is described in Algorithm 2.

Algorithm 2. Classifying procedure of classifier chain

Input: Test example $\mathbf{x}$.

Output: Results of all classifiers in the chain $Y = \{ l_1, l_2, \ldots, l_{|L|}\}$.

Process:

1. $Y \leftarrow \{\}$
2. **for** j∈1, 2, …, $|L|$
3. **do** $Y \leftarrow Y \cup (l_j \leftarrow C_j: (\boldsymbol{x}, l_1, l_2, \ldots, l_{j-1}))$
4. **return** $(\boldsymbol{x}, Y)$

This training method passes label information between classifiers, allowing classifier chain take into account label correlations and thus overcoming the label independence problem of binary relevance method. However, classifier chain still remains advantages of binary relevance method including low memory and runtime complexity.

The order of the chain itself clearly has an effect on accuracy. This problem can be solved by using an ensemble framework with a different random train ordering for each iteration. Ensemble of classifier chains trains m classifiers C_1, C_2, …, C_m. Each C_k is trained with a random chain ordering of L and a random subset of D. Hence each C_k model is likely to be unique and able to give different multi-label predictions. These predictions are summed by label so that each label receives a number of votes. A threshold is used to select the most popular labels which form the final predicted multi-label set.

4 Experimental Results

In order to test the effectiveness and accuracy of the proposed approach, we conduct our experiments on an annotated image data set which was originally used in [7]. The dataset consists of 5000 images from 50 Corel Stock Photo cds. Each cd includes 100 images on the same topic. We divided this dataset into 3 parts: a training set of 4000 images, a validation set of 500 images and a test set of 500 images. The validation set is used to determine system parameters. After fixing the parameters, we merged the 4000 training set and 500 validation set to form a new training set. This corresponding to the training set of 4500 images and the test set of 500 images used by [7].

4.1 Parameters Setting

An important parameter of the experiment is the number of latent aspects for the PLSA-based models. Since the number of latent aspects defines the capacity of the model — the number of model parameters, it can determine the training time and system efficiency to a large extent. We choose three aspect numbers (90, 120 and 150) to do experiments. Through a series of experiments, we found that the system performs better when aspect number is 150. Therefore, we use 150 as aspect number, without ruling out the possibility that another aspect number would make the system performs much better. Furthermore, our approach constructs an ensemble including 90 classifier chains. Each classifier chain randomly chooses a subset of 500 images for training.

The focus of this paper is not on image feature selection and our approach is independent of visual features. So our prototype system uses similar features to [8] for easy comparison. We simply decompose images into a set of blocks (the size of each block is empirically determined as 16×16 through a series of experiments on the validation set), then compute a 36 dimensional feature vector for each block, consisting of 24 color features (auto correlogram computed over 8 quantized colors and 3 Manhattan Distances) and 12 texture features (Gabor energy computed over 3 scales and 4

orientations). As a result, each block is represented as a 36 dimension feature vector. Then each image is represented as a bag of features, that is, a set of 36 dimension vectors. All the feature vectors of training images compose the inputs of continuous PLSA. Therefore, this preprocessing procedure provides a uniform interface for continuous PLSA modeling.

4.2 Results of Automatic Image Annotation

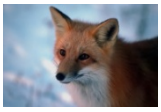
In this section, the performance of our approach is compared with some state-of-the-art approaches — the Translation Model [7], CMRM [10], CRM [11], MBRM [8], PLSA-WORDS [15] and SML [3]. We evaluate the performance of image annotation by comparing the captions automatically generated with the original manual annotations. Similarly to [11], 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 approach performs significantly better than all other approaches. We believe that using hybrid framework to learn semantic classes is the reason for this result.

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

Models	**Translation**	**CMRM**	**CRM**	**MBRM**	**PLSA-WORDS**	**SML**	**Hybrid Approach**
#words with recall > 0	49	66	107	122	105	137	137
Results on 49 best words, as in [7,8,10,11]							
Mean Recall	0.34	0.48	0.70	0.78	0.71	—	0.83
Mean Precision	0.20	0.40	0.59	0.74	0.56	—	0.78
Results on all 260 words							
Mean Recall	0.04	0.09	0.19	0.25	0.20	0.29	0.32
Mean Precision	0.06	0.10	0.16	0.24	0.14	0.23	0.28

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

Image				
Ground truth	grizzly, bear, meadow, grass	head, fox, snow, closeup	trees, sky, frost, ice	sand, water, people, sky
Annotations of PLSA-WORDS	bear, grizzly, horse, meadow, sand	clouds, sky, stone, sculpture, rabbit	grass, desert, ice, path, sculpture	beach, iceburg, snow, ice, water
Annotations of hybrid approach	bear, grizzly, grass, meadow, trees	fox, snow, sky, head, clouds	sky, trees, branch, ice, grass	water, sky, beach, people, snow

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.

4.3 Results of Ranked Image Retrieval

In this section, 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, PLSA-WORDS and SML, because mAP of the Translation model cannot be accessed directly from the literatures.

The annotation results ignore rank order. However, users always like to rank retrieval images and hope that the top ranked ones are relative images. In fact, most users do not want to see more than even 10 or 20 images in a query. Therefore, rank order is very important for image retrieval. Given a query word, 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 our hybrid approach performs better than other models.

Table 3. Comparison of mAPs in ranked image retrieval.

Models	CMRM	CRM	MBRM	PLSA-WORDS	SML	Hybrid Approach
All 260 words	0.17	0.24	0.30	0.22	0.31	0.35
Words with recall ≥ 0	0.20	0.27	0.35	0.26	—	0.41
Words with recall > 0	—	—	—	0.55	0.49	0.67

In summary, the experiment results show that our approach outperforms some state-of-the-art approaches in many respects, which proves that the continuous PLSA and our hybrid approach is effective in modeling visual features and learning semantic classes of images.

5 Conclusion

In this paper, we have proposed continuous PLSA to model continuous quantity and develop an EM-based iterative procedure to estimate the parameters. Furthermore, we present a hybrid generative/discriminative approach, which employs continuous PLSA to deal with the visual features and uses ensembles of classifier chains to learn semantic classes of images. Experiments on the Corel dataset prove that our approach is promising for automatic image annotation and retrieval. In comparison to some state-of-the-art approaches, higher accuracy and superior effectiveness of our approach are reported.

6 Acknowledgments

This work is supported by the National Natural Science Foundation of China (Nos. 61165009, 60903141, 61105052), the Guangxi Natural Science Foundation (2012GXNSFAA053219) and the "Bagui Scholar" Project Special Funds.

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. 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.
4. E. Chang, K.Goh, G. Sychay, and G. Wu. CBSA: content-based soft annotation for multimodal image retrieval using bayes point machines. *IEEE Trans. CSVT*, 13(1): 26–38, 2003.
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, Zhongzhi Shi. Automatic image annotation with continuous PLSA [C]. In: *Proc. 35th ICASSP*, pp. 806–809, 2010.
14. J. Liu, M. Li, Q. Liu, et al. Image annotation via graph learning. Pattern Recognition, 42(2): 218–228, 2009.
15. F. Monay, D. Gatica-Perez. Modeling semantic aspects for cross-media image indexing. *IEEE Trans. PAMI*, 29(10): 1802–1817, 2007.
16. J. Read, B. Pfahringer, G. Holmes, and E. Frank. Classifier chains for multi-label classification. In: *Proc. ECML*, pp. 254–269, 2009.
17. C. Wang, S. Yan, L. Zhang, et al. Multi-label sparse coding for automatic image annotation. In: *Proc. CVPR*, pp. 1643–1650, 2009.
18. A.W.M. Smeulders, M. Worring, S. Santini, *et al.*, "Content-based image retrieval at the end of the early years," *IEEE Trans. PAMI*, 22(12): 1349-1380, 2000.