Learning Similarity with Probabilistic Latent Semantic Analysis for Image Retrieval

Xiong Li[1], Qi Lv[2], Wenting Huang[1]*
[1] National Computer Network Emergency Response Technical Team
Beijing, 100029, China
[2]School of Management and Economics, North China University of Water Resources and Electric Power
Zhengzhou, 450000, China
[e-mail: hwt_chn@163.com]
*Corresponding author: Wenting Huang

*Received December 14, 2014; revised January 31, 2014; accepted February 27, 2015;
published April 30, 2015*

Abstract

It is a challenging problem to search the intended images from a large number of candidates. Content based image retrieval (CBIR) is the most promising way to tackle this problem, where the most important topic is to measure the similarity of images so as to cover the variance of shape, color, pose, illumination etc. While previous works made significant progresses, their adaption ability to dataset is not fully explored. In this paper, we propose a similarity learning method on the basis of probabilistic generative model, i.e., probabilistic latent semantic analysis (PLSA). It first derives Fisher kernel, a function over the parameters and variables, based on PLSA. Then, the parameters are determined through simultaneously maximizing the log likelihood function of PLSA and the retrieval performance over the training dataset. The main advantages of this work are twofold: (1) deriving similarity measure based on PLSA which fully exploits the data distribution and Bayes inference; (2) learning model parameters by maximizing the fitting of model to data and the retrieval performance simultaneously. The proposed method (PLSA-FK) is empirically evaluated over three datasets, and the results exhibit promising performance.

Keywords: Image retrieval, PLSA, Fisher kernel, similarity measure learning

This work was supported by NSFC 61403090.

http://dx.doi.org/10.3837/tiis.2015.04.009 **ISSN : 1976-7277**

1. Introduction

The last decade has seen the increasing popularity of digital images, especially along with the development of Internet. How to search images according to users' intention from a large number of candidates has been an important yet challenging problem [1-9]. In real application, there are two main cases. The first one is "retrieval by text", i.e., the users describe their intention through keywords and demand images matching these keywords. This is trackled by image annotation tecniques, i.e., assigning keyworks to candidate images and thus convert the problem to text matching, which is popular in commercial search engines. The second is "retrieval by image" [1, 3, 4, 5], i.e., the useres input a sample image and expect to get those images like the input one. The methods for this case is mainly so called image similarity or distance learning. In this work, we focus on the second case since it is more descriptive and requires no keyword or textural metadata to describe the image. This case is also referred to as content based image retrieval (CBIR). Note that, although user click of search engine can significantly improve the retrival performance [8-10], we in this work foucs on the visual channel and the proposed method can be extended to work with user click information.

As the most important component of CBIR systems, image similarity or distance measure greatly determines the retrieval performance [1, 5]. Noticing the fact that the similarity measure and distance measure are technically convertible and functionally equal, we in this work stick to similarity measure, but the proposed method could be directly applied to distance measure learning. In the technical perspective, similariyt measure has two important factors. The first factor is the image feature representation. A satisfied image feature is expected being robust to illumination, pose, shape and other variances, and simutaniouly being effective in capturing useful information, i.e., reaching balance between robustness and selectivity. It is worth noting that the evaluation of feature representation is task-specific. Therefore, it would be important to select or learn feature representation under the criterion of the task. The second factor is the similarity function which is a function defined over the feautre space, outputing large value for a similar pair and small value for a distinct pair. A typical kind of similarity measure is the predefined similarity measures [11], such as L1 distance and Euclidean distance. This kind of predefined similarity measures, however, are usually not adaptive enough to data distribution [1]. Alternatively, similarity learning methods [12-17] have been proposed, to improve the adaption ability to data distribution. The method proposed in this work belong to this branch.

Similarity learning methods in general fall to three typical categories, unsupervised method, semisupervised method and supervised method. Unsupervsed learning methods [18,19,20] seek to find a feature space and a similarity measure for the given data, without taking class label into account. The typical methods include subspace based methods, e.g. locally linear embedding (LLE) [18], non-negative matrix factorization (NNMF) [19], probabilistic model based methods [20, 21, 22, 23, 24, 25]. Among them, probablistic model based similarity shown promsing performance and received increasing attention. These methods formulated the feature mapping [24, 25] or similarity measure [20, 21, 22] based on the probabilistic model. Thus, they inherited the abilities of probabilistic model and exhibited adaption ability to data distribution. These methods include probability product kernels [21], Kullback Leibler divergence based similarity [22], Fisher kernel [20], free energy score space [24]. These methods are useful when the class label is missed or is expensive to otain.

Semisupervised learning methods [11, 14] make use of both labeled data and unlabeled data, laying somewhere between unsupervised learning and supervised learning. They are highly effecitive when the number of labeled data is limited and the number of unlabeled data is easy to access. On the other hand, supervised learning methods learn similarity measure by exploiting class label, seeking to find a similarity function that outputs large value for images with the same labels and outputs small value for images with distinct labels. Popular methods include large margin nearest neighborhood (LMNN) [26], local distance metric learning (LDML) [12], linear transformation based metric learning (LTML) [27] and discriminative Fisher kernel learning [28] etc. These methods however not fully exploit data distribution and hidden information which will potentially improve the adaption ability and effectiveness of similarity measure.

To fully exploit data distribution, hidden information and class label for similarity learning, we in this paper propose an approach based on probabilistic latent semantic analysis (PLSA) [29] and Fisher kernel [20]. The proposed approach, referred to as PLSA-FK, uses bag-of-words feature represention for images, where the visual words are quantified from local descriptor, and then leverage PLSA to model the distribution of the visual words. On the basis of PLSA model, it then derives the Fisher kernel which is a function over the parameters and variables of PLSA. To exploit class label, i.e., tuning the similarity measure as well as the PLSA model to have good retreival performance, we developed a supervised learning approach for the PLSA based Fisher kernel. The motivations of the proposed method are twofold. First, exploit the semantic information by means of coupling with PLSA which is able to infer topic. Second, exploit the label information which is informative for retrieval. The proposed method has three main advantages. First, probabilistic models could well adapt to data distribution. Second, PLSA can exploit the ability of Bayes inference. Third, the supervised learning method can tune the similarity and model according to the retrieval performance.

The remaining part of this paper is organized as follows. Section 2 revisits the related works of similairty learning. Section 3 proposes our approach, PLSA based similarity learning. Section 4 experimentally evaluates the proposed approach over the real databases. Section 5 draws a conclusion.

2. Related Works

There are a number of works have contributed to similarity learning [30-33, 14-17] and to content based image retrieval [4, 5, 13, 11, 1]. We in this work attempt to make a progress on the adaption ability, and thus naturally focus on supervised similarity learning approaches, and probabilistic model based approaches. For other related approaches, see references for a details. It is worth noting that, similarity learning and distance learning are essentially equal because they are convertable. Thus, we in this work treat them as the same notation.

Supervised similarity measure learning attempts to learn a similarity measure from a set of equivalence constraints for image pair within the same class, and inequivalence constraints for image pair of the different classes. The similarity measure is determined under the criterion that keeps images in equivalence constraints close and images in inequivalence constraints separated. A number of recent works attempted to cooperate relevance feedback [3, 30], dimensionality reduction [31], Bayesian inference [34] and kernel method [27]. [32] casted the problem into a constrained convex optimization problem by minimizing the pairwise distance in the same classes so that images of different classes are well separated. Discriminative component analysis (DCA) [2] incorporated equivalence constraints for similarity learning.

Large-margin nearest neighbor (LMNN) [26] took the class margin into account. SDPM [35] formulated Mahalanobis distance learning as a convex optimization problem. Distance metric learning with eigenvalue optimization (DML-eig.) [36] casted distance learning problem as a eigenvalue optimization problem. [33] learnt local perceptual distance function which is a combination of a set of local distance functions. [37] proposed to learn the Mahalanobis distance function subject to a set of pairwise constraints, i.e., must-links that associate images which must be in the same class and cannot-links that associate images which must be in different classes. [38] made ues of context information to learn similarity measures. [39, 16] leveraged discriminative learning techniques to learn similarity measure.

Probabilistic similarity methods formulate explicit feautre space or similarity measure based on the quantities of adopted probabilistic models. Probability product kernels [21] used the posterior distributions of hidden variables to characterize the samples, and define the similarity measure as the expectation of the inner product of the hidden variables, with respect to the posterior distribtuions. [22] used distributions to characterize the samples and uses Kullback–Leibler divergence over those distributions to measure the distance between samples. [23] developed a hierarchical probabilistic model to learn image representation and similarity. Fisher score (FS) [20] derived feature mapping by considering how the samples affect the model parameters, and defined the similarity, i.e., Fisher kernel, as the inner product of the feature mappings of samples. Free energy score space (FESS) [24] and posterior divergence (PD) [25] extended Fisher score by exploring more informative measures. These approaches are able to exploit information from probabilistic models, they however can be further boosted through fully exploiting the class label, by fitting the similarity measure to the retrieval performance. Dsicriminative Fisher kernel learning (DFK) [28] extents Fisher kernel to cooperate class label, where Gaussian mixture model is used to model the distribution of visual features. It does not utiliize semantic level information in an explicit way.

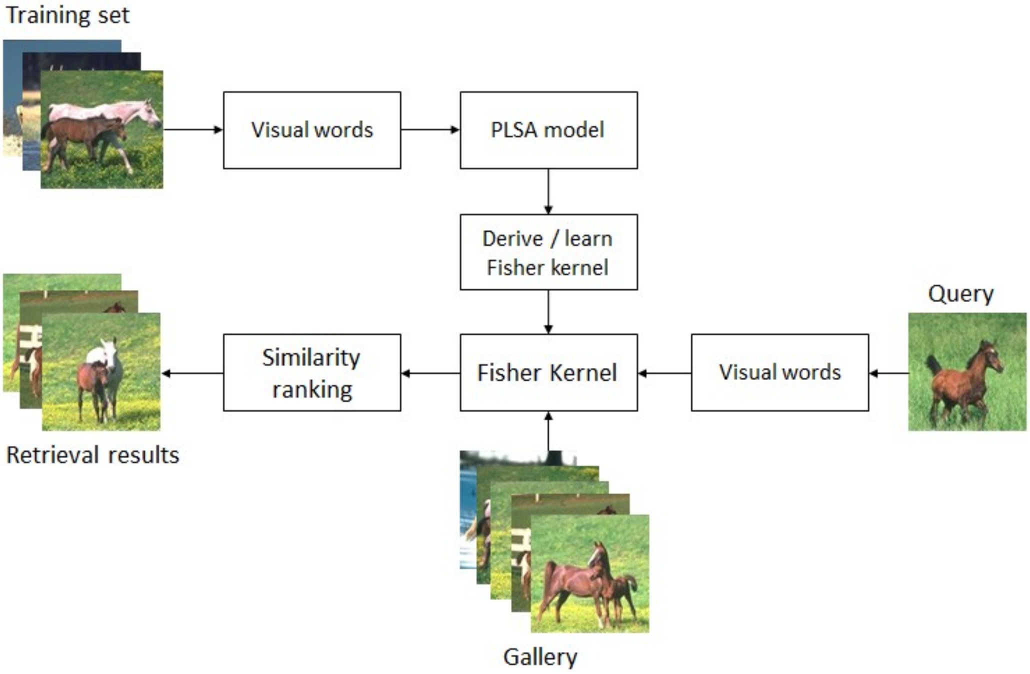

Fig. 1. The framework of our proposed approach PLSA-FK.

Recently, numerous works introduced a variety of techniques for similarity learning. [40, 6] used graph to represent data and formulated similarity learning problem as a graph learning problem. [9, 8] exploited user click information in real retrieval systems to boost the retrieval performance. [15] learned distance subject to the criterion that the semantic information is preserved. [17] proposed to learn the similarity by considering the neighborhood structure.

The work in this paper is based on the score space methods which have been validated to be very competitive in a wide range of applications [24, 25, 28]. The advantages of the proposed method are mainly twofold: (1) compared with deterministic similarity learning methods, our method fully exploits data distribution information and semantic level hidden variables by means of Bayesian inference; (2) compared with probabilistic similarity learning method, our method provides a sophisticated way to utilize class label.

3. Learning Fisher Kernel with PLSA

In this section, we will proceed to derive the Fisher kernel based on PLSA and propose a supervised learning approach for the derived kernel. First, we use PLSA to model the distribution of visual words, for its popularity and effectiveness in image modeling [41]. Then we derive the Fisher kernel [20] based on PLSA. At last, we propose a supervised learning method for Fisher kernel. See **Fig. 1** for the illustration and **Table 1** for the notations, of the proposed method.

Table 1. Mathematical notations involved in this work

Notation	Definition	Notation	Definition
d	document	$D=\{d_1,\ldots,d_N\}$	document collection
w	term	$W=\{w_1,\ldots,w_V\}$	dictionary
z	topic	$Z=\{z_1,\ldots,z_K\}$	topic set
$n(w_l,d_i)$	frequency of w_l in d_i	$\theta=\{\alpha,\beta\}$	model parameter
U	weight matrix	$y=(y_1,\cdots,y_C)$	label vector
i,j	index of sample	k,l	index of topic, word

3.1. Probabilistic Latent Semantic Analysis

In this work, we utilize Probabilistic Latent Semantic Analysis (PLSA) [29] to model the distribution of images represented in bag of visual words quantized from image features. The effectiveness of PLSA in image and text representation has been extensively verified [41].

PLSA is a probabilistic generative model orignally developped for text analysis [29]. Sepecifically, it could discover the semantic topics hidden in documents using the bag of words representation. PLSA can be also applied to image analysis, since images can also use bag of words representation where an image patch is quantified to a visual word.

Let $D=\{d_1,\ldots,d_N\}$ be a collection of N documents formed by words from a dictionary $W=\{w_1,\ldots,w_V\}$ of V terms. The data can be denoted by a $V\times N$ co-occurrence matrix where the element $n(w_l,d_i)$ is the frequency of a term w_l appeared in a document d_i. Let $z\in Z=\{z_1,\ldots,z_K\}$ be the hidden topic variable associating with each observation which is actually the occurrence of a word in a certain document. Let $P(d_i)$ denote the probability of a certain document d_i; $P(w_l \mid z_k)$ denote the conditional probability of a specific word w_l conditioned on the hidden topic z_k; $P(z_k \mid d_i)$ denote the conditional probability of a hidden topic z_k conditioned on the document d_i. Then PLSA can be expressed as follows:

(1) Choose a document d_i according to a probabilistic distribution $P(d_i)$;

(2) Choose a hidden topic variable z_k according to a probabilistic distribution $P(z_k \mid d_i)$;

(3) Generate a word w_l according to a probabilistic distribution $P(w_l \mid z_k)$. Then we have an paired observation (w_l, d_i).

The joint probability of the above model can be expressed as:

$$P(w,d,z) = P(w \mid z)P(z \mid d)P(d) \tag{1}$$

By marginalizing over the hidden variable z, it gives the marginal distribution,

$$P(w,d) = \sum_{z \in Z} P(w,d,z) = P(d)\sum_{z \in Z} P(w \mid z)P(z \mid d)$$

Note that $P\left(w,d\right) = P\left(d\right)P(w \mid d)$, we have the condition distribution $P(w \mid d)$ as follows,

$$P(w \mid d) = \sum_{z \in Z} P(w \mid z)P(z \mid d)$$

In this model, each document is modeled as a mixture of topics, and the word histogram for a certain document is composed of a mixture of the histograms corresponding to each topic. Especially, each document is a combination of the topic vector.

Learning this model involves the determination of the mixture coefficients which are specific for each document, and involves the determination of the topic parameters shared by all documents. To learn the PLSA model, we determine the conditional probabilities $P(z \mid d)$ and $P(w \mid z)$ by maximizing the following log likelihood function:

$$L = \log P(D,W) = \sum_{d \in D}\sum_{w \in W} n(w,d)\log P(w,d) \tag{2}$$

Maximizing this log likelihood function is equivalent to minimizing the Kullback-Leibler divergence between the empirical distribution and the parameterized model. This model can be effectively learned using Expectation Maximization (EM) algorithm, as described in [29].

3.2. Fisher kernel based on PLSA

Having the PLSA model, to derive the Fisher kernel, we first give the variational lower bound [42] of the log likelihood function $\log P(d,w;\theta)$, on which the derivation will be simple,

$$\log P(d,w;\theta) \geq -\mathrm{KL}(Q(z)\mathrm{P}P(d,w,z;\theta)) = -F(\theta)$$

where $Q(z)$ is the approximation of the real posterior $P(z \mid d,w)$ and usually shares the same parameterizations with $P(z)$. Then we have,

$$\begin{aligned} F(\theta) &= \mathrm{E}_{Q(z)}[\log Q(z) - \log P(d,w,z;\theta)] \\ &= \sum_{d,w} n(d,w)\sum_{z} Q(z \mid d,w)[\log Q(z \mid d,w) - \log P(w \mid z)P(z \mid d)] \\ &= \sum_{d,w} n(d,w)\sum_{k} g_{kwd}[\log g_{kwd} - \log \beta_{kw}\alpha_{kd}] \end{aligned} \tag{3}$$

Given the variational lower bound $-F(\theta)$ of the log likelihood function $\log P(d,w;\theta)$, the elements of Fisher score is its gradient with respect to the model parameters [20],

$$\frac{\partial - F(\theta)}{\partial \beta_{kw}} = -\sum_{d,w} n(d,w)\frac{g_{kwd}}{\beta_{kw}} \tag{4}$$

$$\frac{\partial - F(\theta)}{\partial a_{kd}} = -\sum\nolimits_{d,w} n(d,w)\frac{g_{kwd}}{a_{kd}} \tag{5}$$

Note that the elements of Fisher score is the expectation over a function of the observed variable d, hidden variables z and model parameters θ, where the hidden variables allow Fisher kernel to exploit hidden information and the model parameters make it adaptive to data distribution. The complete Fisher score is the combination of those gradients,

$$\Phi(d) = \left(\frac{\partial - F(\theta)}{\partial \beta_{11}}, \cdots, \frac{\partial - F(\theta)}{\partial \beta_{KV}}, \frac{\partial - F(\theta)}{\partial \alpha_{1d}}, \cdots, \frac{\partial - F(\theta)}{\partial \alpha_{Kd}} \right) \tag{6}$$

The Fisher kernel then can be defined as [20],

$$K(d_i, d_j) = \Phi(d_i)^T \mathrm{I}\Phi(d_j)$$

where $\mathrm{I}=\mathrm{E}_d[\Phi(d)\Phi(d)^T]$ is the Fisher information matrix. In order to exploit class label for similiarity learning, we here extend the kernel to the following parameterized form:

$$K(d_i, d_j) = \Phi(d_i)^T U \Phi(d_j) \tag{7}$$

where U=diag$(u_1, \cdots, u_M)$ is a diagonal matrix to be learnt, and u_m weights the importance of Φ_m, i.e. the *m*-th element of Φ, to the similarity. In particular, $u_m = 0$ indicates that Φ_m is completely non-informative. Given the above parameterized form of Fisher kernel, we will show how to determine U in next section.

3.3. Learning PLSA based Fisher kernel

Let $y_i = (y_{1i}, \cdots, y_{Ci})$ be the label vector of a sample d_i, where $y_{ci} = 1$ indicates that the c-th label of all C ones is assigned to the sample d_i and $y_{ci} = 0$ otherwise. Here we consider the criterion that sample pairs take high similarity for sample pairs with the same class label, and takes low similarity for sample pairs with different class label,

$$J(\theta, U) = \sum_i \sum_{j \neq i} y_i^T y_j K(d_i, d_j; U, \theta) \tag{8}$$

where $y_i^T y_j$ measures the similarity of two label vectors.

Given the approximate posterior $Q(\mathbf{z})$ over the hidden variable, we seek to minimize the objective function $J(\theta, U)$ using gradient descent,

$$\frac{\partial J(\theta,U)}{\partial \beta_{kw}} = \sum_i \sum_{j \neq i} y_i^T y_j \left[\left(\sum\nolimits_{d,w} n(d,w) g_{kwi} \beta_{kw}^{-2} \right) u_{m'} \Phi_{m'j} + \left(\sum\nolimits_{d,w} n(d,w) g_{kwj} \beta_{kw}^{-2} \right) u_{m'} \Phi_{m'i} \right] \tag{9}$$

$$\frac{\partial J(\theta,U)}{\partial \alpha_{kd}} = \sum_i \sum_{j \neq i} y_i^T y_j \left[\left(\sum\nolimits_{d,w} n(d,w) g_{kwi} \alpha_{kd}^{-2} \right) u_{m''} \Phi_{m''j} + \left(\sum\nolimits_{d,w} n(d,w) g_{kwj} \alpha_{kd}^{-2} \right) u_{m''} \Phi_{m''i} \right] \tag{10}$$

$$\frac{\partial J(\theta,U)}{\partial u_m} = \sum_i \sum_{j \neq i} y_i^T y_j \Phi_m(d_i) \Phi_m(d_j) \tag{11}$$

where m' indexes the element of feature mapping Φ for β_{kw} (Eq. (4)) while m'' indexes the element of feature mapping Φ for α_{kd} (Eq. (5)).

The learning procedure of the proposed approach is the iteration of the E-step and M-step (Eq. (9-11)), which is summarized in **Algorithm 1**.

Algorithm 1 Learning Fisher kernel

```
input: training set $\{(d_i, y_i)\}_{i=1}^{N}$; iteration number $T$; learning rate $\gamma > 0$
initialize parameters $\theta^{(0)}, U^{(0)}$
for $t = 1$ to $T$ do
    $P(z_k \mid d_i, w_l) = \frac{P(w_l \mid z_k) P(z_k \mid d_i)}{\sum_{k=1}^{K} P(w_l \mid z_k) P(z_k \mid d_i)}$
    $\theta^{(t)} \leftarrow \theta^{(t-1)} - \gamma \frac{\partial J(\theta, U)}{\partial \theta}$
    $U^{(t)} \leftarrow U^{(t-1)} - \gamma \frac{\partial J(\theta, U)}{\partial U}$
end for
output: $\theta^{(T)}, U^{(T)}$
```

The learnt Fisher kernel can be embeded to kernel-compatible classifier for classification, and the kernel similarity of a pair of samples d_i, d_j can be computed following **Algorithm 2**.

Algorithm 2 Computing Fisher kernel

```
input: a pair of samples $d_i, d_j$
compute posterior $P(z_k \mid d_i, w_l) = \frac{P(w_l \mid z_k) P(z_k \mid d_i)}{\sum_{k=1}^{K} P(w_l \mid z_k) P(z_k \mid d_i)}$ for $d_i$
compute posterior $P(z_k \mid d_j, w_l) = \frac{P(w_l \mid z_k) P(z_k \mid d_j)}{\sum_{k=1}^{K} P(w_l \mid z_k) P(z_k \mid d_j)}$ for $d_j$
compute the Fisher kernel similarity using Eq. (7)
output: $K(d_i, d_j)$
```

4. Experiments

In this section, we will apply the proposed method, i.e., Probabilistic Latent Semantic Analysis based Fisher Kernel (PLSA-FK) for image retrieval. The proposed method will be compared with several state-of-the-art methods on three real datasets, Corel5K [43], MIRFlickr 25,000 [44] and Corel30K [45].

4.1. Image Representation

Image feature representation is crucial for CBIR systems due to the great variance of visual contents across image datasets. In this work, we ues color SIFT descriptors as the feature for its excellent performance which has been extensively validated [46]. Specifically, following the recommendation in [46], four color SIFT descriptors (OpponentSIFT, rgSIFT, C-SIFT and RGB-SIFT) are adopted. These descriptors are extracted from the image patches given by dense sampling and Harris-Laplace point sampling, with spatial pyramid followed.

4.2 Performance Measure

Following the previous works [36, 13, 11], we evaluate the retrieval performance using leave-one-out manner. First, a query image is chosen from the test set. Then, search similar images from the candidate set according to the adopted similarity or distance measure. Mean average precision (MAP) is used to measure the performance of image retrieval. MAP is the summarization of the precision-recall curve, where precision is defined as the percentage of returned images that contain the same label with the query image in all returned images.

Let k be the rank, the precision at cut-off k can be computed as:

$$P(k)=\frac{|\{\text{relevant retrieved images of rank } k \text{ or less}\}|}{k}$$

Averaging the precision of those relevant returned images gives Average Precision (AP),

$$\mathrm{AP}=\frac{1}{N}\sum\nolimits_{k=1}^{N}P(k)\times\mathrm{rel}(k)$$

where N is the number of retrieved images, $\mathrm{rel}(k)$ is an indicator function outputting 1 if the image at the rank k is a relevant image and 0 otherwise. MAP is then given by averaging AP across all the query images,

$$\mathrm{MAP}=\frac{1}{Q}\sum\nolimits_{q=1}^{Q}\mathrm{AP}(q)$$

4.3 Experiments on Corel5K dataset

To evluate our approach PLSA-FK, we first perform an experiment on Corel5K dataset [43]. Corel5k dataset is a subset selecting from Corel Photo Gallery, being composed of 50 categories, such as beach, tile, wave, food texture, tigers, France, bears, autumn, and tropical plants, where each category contains 100 images. It contains 371 word vocabulary. The sizes of the images are normlized to 192×128 or 128×192. The sample images are shown in **Fig. 2**. In this experiment, we randomly select 70% samples to form the training set and remain the rest as the test set. The training set is used for PLSA-FK learning and the test set is used for performance evlauation. For all compared approaches, we measure the average precision for each category over the top 20 retrieved images.

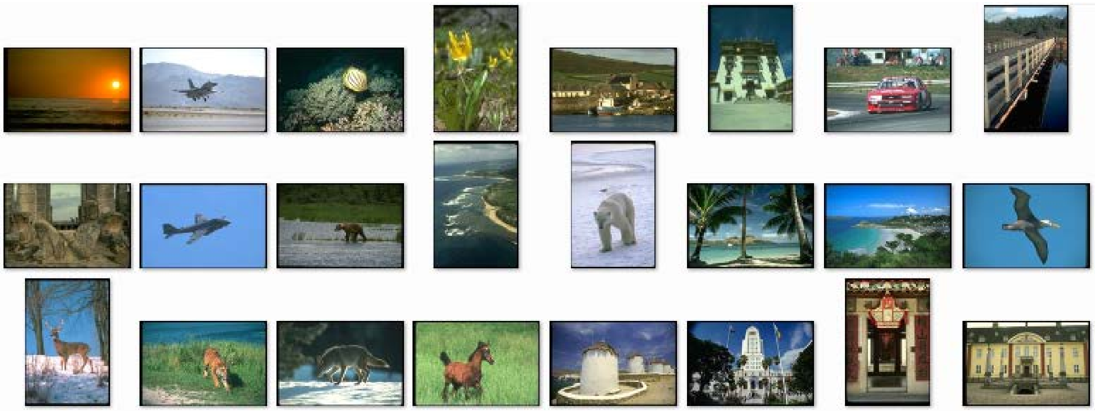

Fig. 2. Sample images from Corel5K dataset

We will compare our PLSA-FK with other similarity or distance learning methods. Xing's approach [32] casts the distance measure learning problem to a convex optimization problem. Discriminative components analysis (DCA) [2] introduces inequivalence constraints. SDPM [35] learns Mahalanobis distance through formulating it as a convex optimization problem. DML-eig. [36] learns distance by means of eigenvalue optimization. Large margin nearest neighbor (LMNN) [26] learns Mahalanobis distance under the criterion of nearest neighbor and large margin. Fisher kernel [20] and Free energy score space (FESS) [24] converts similarity learning to feature mapping learning on the basis of probabilistic generative models. Its similarity measure is the inner product of feature mapping. Discriminative Fisher kernel learning (DFK) [28] learns kernel similarity through exploiting class label.

For our approach, the number of topics for PLSA is determined through cross-validation on the test set. It is set to $K = 120$ in this experiment. For compared approaches, we referred to the results for Xing's [32], DCA [2], SDPM [35], DML-eig. [36], LMNN [26] from literatures, and implemented the algorithms of FK, FESS and DFK on the basis of authors' implementations and parameter configurations.

The experimental results are reported in **Table 2**. It can be found that, Xing's approach [32] and DML-eig [36] show competitive performance. Meanwhile, for distance measure learning approaches, DCA beat SDPM, DML-eig and LMNN. The underlying reason is that DCA introduced negative constraints which capture intrinsic structures within samples. And, the probabilistic similarity measure learning approach FK and FESS get better results due to the exploitation of probabilistic modeling of image distribution. Our approach PLSA-FK, as shown in **Table 2**, achieves the best performance against the compared approaches in most cases. Specifically, PLSA-FK approach outperforms FESS by about 2.7%. This improvement should be credited to that PLSA-FK utilizes the label information while FESS does not. Also, PLSA-FK outperforms DFK about 1.4%. The main reason is that, compared with GMM, PLSA can capture the image attributes and infer the semantic level hidden information better. These results demonstrate the effectiveness of the proposed method in image retrieval.

Table 2. Retrieval performance of all algorithms over Corel5K dataset

Algorithm	MAP (mean average precision)
Xing [32]	0.307
DCA [2]	0.325
SDPM [35]	0.315
DML-eig [36]	0.309
LMNN [26]	0.310
FK [20]	0.314
FESS [24]	0.316
DFK [28]	0.329
PLSA-FK (ours)	**0.343**

Fig. 3 shows an example of our approach PLSA-FK on Corel5k dataset. Given the query image (left-top), it returns relevant images and lists the top 11 images in the figure. We could find that most results are relevant, and even incorrect results exhibit similarity in both shape

and color, which indicates that our proposed approach could potentially capture multiple kinds of information and comprehensively contributes to the retrieval.

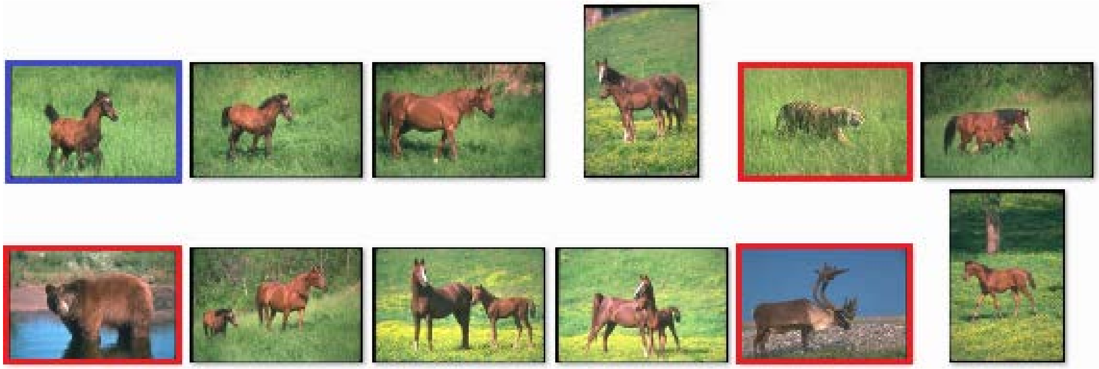

Fig. 3. Retrieval results of our approach PLSA-FK. The query image is marked by blue box and the incorrect results of top 11 ones are marked by red box.

4.4 Experiments on MIRFlickr dataset

In real applications, the dataset is usually very large, which requires that the similarity measure (1) is scalable and computationally efficient; (2) is able to characterize the semantic similarity between images, given the large variance. To evaluate performance of our method on large dataset, we experimented on MIRFlickr dataset [44]. The MIRFlickr-25000 dataset contains 25,000 samples with high-resolution images and text annotations, collected from Flickr which is an online photo-sharing website. The size of the images are normalized to 500×height where height<500 or width×500 where width<500. See **Fig. 4** for sample images. For fair comparison, we follow the typical experimental scheme. The dataset is split into two parts, 15,000 images for training and the rest 10,000 images for test. We randomly chosen 1,000 images from the test dataset as queries and remained the rest 24,000 images as the gallery. In the gallery, 15,000 images are with text annotations.

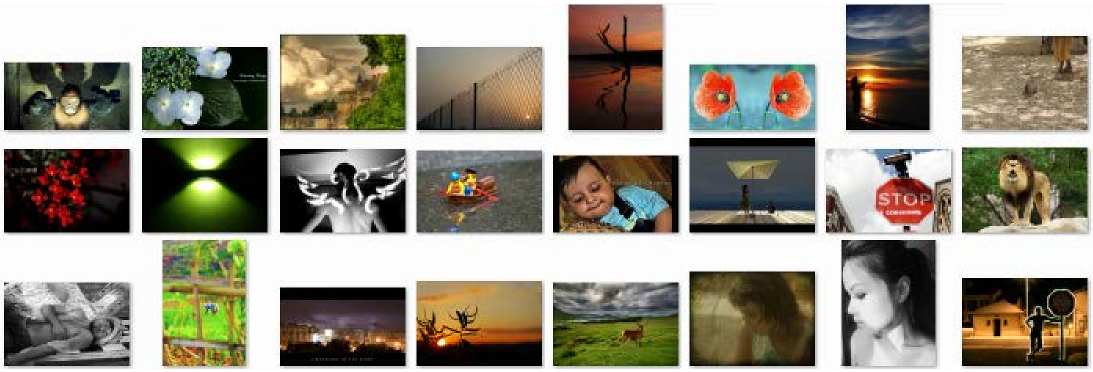

Fig. 4. Sample images from MIRFlickr dataset

We compare our proposed method PLSA-FK with several related methods: non-negative matrix factorization (NNMF) [19], large margin nearest neighbor (LMNN) [26], free energy score space (FESS) [24] and posterior divergence (PD) [25]. NNMF is a state-of-the-art image retrieval method on the basis of matrix factorization. LMNN is a supervised distance learning method under the criterion of large margin. FESS and PD are probabilistic similarity learning methods closely related to our method. For all these compared methods, we used the authors' suggested settings. For our method, the number of topics of PLSA is set to K=160 according to cross validation.

Table 3. The retrieval performance of compared algorithms on MIRFlickr dataset.

Algorithm	MAP (mean averge precision)
NNMF [19]	0.583
LMNN [26]	0.586
FK [20]	0.588
FESS [24]	0.590
PD [25]	0.593
DFK [28]	0.606
PLSA-FK (ours)	0.619

The experimental results are reported in **Table 3**. It can be clearly found that, FK, FESS and PD outperform NNMF and LMNN. The underlying reason is that, compared with NNMF and LMNN, FK, FESS and PD exploit the data distribution more sophisticatedly. Further, our method PLSA-FK shows superiority over FESS and PD. The reason accounting for this is that, PLSA-FK exploits class label through tuning similarity measure according to performance. Again, PLSA-FK outperforms DFK, which benefits from the semantic level information given by PLSA. **Fig. 5** presents the retrieval results of our method for a query image "flower". For the query image on the left-top, our method retrieves the relevant images and presents the top 11 ones in the figure. It can be seen that, 9 retrieved images are relevant. It is interesting to find that the 2 unrelevant images are similar to the query image in shape and pattern, which suggests that the results of our approach could give reasonable retrieval results.

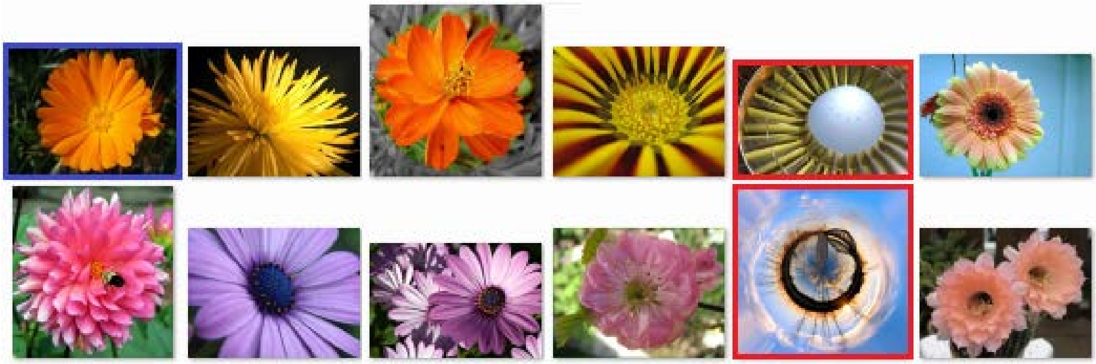

Fig. 5. Retrieval results of our proposed approach PLSA-FK. The query image "flower" is highlighted by blue box and the incorrect results in top 11 relevant images are highlighted by red box.

4.5 Experiments on Corel30K dataset

To further evaluate the performance of our proposed approach PLSA-FK on large dataset, we experimented on Corel30k [45] for image retrieval. It contains annotated 31,695 images (28,525 training and 3,170 testing) with annotations from 950 words. It is worth noting that, only a few works, up to now, have experimented on this dataset [47]. This experiment compared our proposed approach PLSA-FK with PLSA-WORD [48] and GM-PLSA [47] on a total of 950 keyword sets. PLSA-WORD quantified visual features into discrete words and exploited PLSA model [29] for distribution modeling and image retrieving. GM-PLSA is also based on PLSA but further cooperate with other models. We refer to the results of PLSA-WORD and GM-PLSA. For Fisher kernel [20], free energy score space (FESS) [24] and posterior divergence (PD) [25], we implemented them according to authors' suggestions. For our PLSA-FK, the number of topics in PLSA is set to K=180 according to cross validation.

The overall experimental results are summerized in **Table 4**. It shows that, for two evaluation criterion, our approach PLSA-FK outperforms PLSA-WORD significantly, and shows competitive performance with GM-PLSA. It is worth noting that, both PLSA-WORD and GM-PLSA are based on PLSA and thus capture some semantic level information. Also, our PLSA-FK shows superority over three score methods, FK, FESS and PD. Although they exploit semantic level information through PLSA, they do not benefit from the class label information. These results are firmly consistent with the above two experiments, which support the advantage of our approach that it can adapt to data distribution and fully exploit high-level semantic information hidden in the images.

Table 4. The retrieval performance on Corel30K

Algorithms	MAP (All words)	MAP (Words with recall > 0)
PLSA-WORD [48]	0.14	0.17
GM-PLSA [47]	0.23	0.28
FK [20]	**0.23**	**0.27**
FESS [24]	0.24	0.26
PD [25]	0.22	0.25
PLSA-FK (ours)	**0.26**	**0.30**

4.6 Discussions

The learning procedure (**Algorithm 1**) of the proposed method is the iterations of the inference step (E-step) and parameter estimation step (M-step). This procedure is relatively time consuming to reach convergence. However, this procedure can be greatly sped up by means of pretraining, i.e., train PLSA first and use the trained parameter as the initial value of **Algorithm 1**. The computation procedure (**Algorithm 2**) is highly effective, since it only involves two inference steps, and can be realtime. The limitation of applying the method for larger dataset, e.g. ImageNet, is the learning procedure. That is, to learn over larger dataset, the method should be compatible with incremental learning. In our future work, we will seek to develop the incremental algorithm and apply it for larger dataset.

5. Conclusions

In this paper, we exploited probabilistic latent semantic analysi (PLSA) for similarity measure learning towards content based image retrieval (CBIR). The proposed approach (PLSA-FK) derived Fisher kernel based on PLSA and learnt the kernel similarity subject the criterion that image pairs with same label have large value and image pairs with different labels have small value. Because PLSA models the distribution of visual words, our approach can well adapt to data distribution. Further, PLSA infers the topic, thus our method exploits high level semantic information of retrieval. The proposed method is applied to image retrieval. The experimental results over three datasets approve the competitive performance of our method as well as its scaleablity to large datset. The method, however, can be further optimized for large dataset, which remains in the future work.

References

[1] A. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain, “Content-based image retrieval at the end of the early years,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1349-1380, 2000. Article (CrossRef Link)

[2] S. Hoi, W. Liu, M. Lyu, W. Ma, “Learning distance metrics with contextual constraints for image retrieval,” in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2072-2078, 2006. Article (CrossRef Link)

[3] S. Hoi, M. Lyu, R. Jin, “A unified log-based relevance feedback scheme for image retrieval,” in *Proc. of IEEE Transactions on Knowledge and Data Engineering*, 18(4):509-524, 2006. Article (CrossRef Link)

[4] F. Faria, A. Veloso, H. Almeida, E. Valle, R. Torres, M. Gonc¸alves, W. Meira Jr, “Learning to rank for content-based image retrieval,” in *Proc. of ACM conference on Multimedia Information Retrieval*, pp. 285-294, 2010. Article (CrossRef Link)

[5] M. Arevalillo-Herr´aez, F. Ferri, J. Domingo, “A naive relevance feedback model for content-based image retrieval using multiple similarity measures,” *Pattern Recognition*, vol. 43, no. 3, pp. 619-629, 2010. Article (CrossRef Link)

[6] M. Wang, H. Li, D. Tao, K. Lu, “Multimodal graph-based reranking for web image search,” in *Proc. of IEEE Transactions on Image Processing*, vol. 21, no. 11, pp. 4649-4661, 2012. Article (CrossRef Link)

[7] M. Wang, K. Yang, X.S. Hua, H.J. Zhang, “Towards a relevant and diverse search of social images,” in *Proc. of IEEE Transactions on Multimedia*, vol. 12, no. 8, pp. 829-842, 2010. Article (CrossRef Link)

[8] J. Yu, D. Tao, M. Wang, Y. Rui, “Learning to rank using user clicks and visual features for image retrieval,” in *Proc. of IEEE Transactions on Cybernetics*, vol. 99, pp.2168-2267, 2014. Article (CrossRef Link)

[9] J. Yu, Y. Rui, B. Chen, “Exploiting click constraints and multi-view features for image re-ranking,” in *Proc. of IEEE Transactions on Multimedia*, vol. 16, no. 1, pp. 159-168, 2014. Article (CrossRef Link)

[10] J. Yu, Y. Rui, D. Tao, “Click prediction for web image reranking using multimodal sparse coding,” in *Proc. of IEEE Transactions on Image Processing*, vol. 23, no. 5, pp. 2019-2032, 2014. Article (CrossRef Link)

[11] S. Hoi, W. Liu, S. Chang, “Semi-supervised distance metric learning for collaborative image retrieval and clustering,” in *Proc. of ACM Transactions on Multimedia Computing, Communications, and Applications*. vol. 6, no. 3, 2010. Article (CrossRef Link)

[12] L. Yang, R. Jin, R. Sukthankar, Y. Liu, “An efficient algorithm for local distance metric learning,” in *Proc. of the National Conference on Artificial Intelligence*, 2006. Article (CrossRef Link)

[13] L. Yang, R. Jin, L. Mummert, R. Sukthankar, A. Goode, B. Zheng, S. Hoi, M. Satya-narayanan, “A boosting framework for visuality-preserving distance metric learning and its application to medical image retrieval,” in *Proc. of IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 1, pp. 30-44, 2010. Article (CrossRef Link)

[14] J. Yu, M. Wang, D. Tao, “Semisupervised multiview distance metric learning for cartoon synthesis,” in *Proc. of IEEE Transactions on Image Processing*, vol. 21, no. 11, pp. 4636-4648, 2012. Article (CrossRef Link)

[15] J. Yu, D. Tao, J. Lic, J. Cheng, “Semantic preserving distance metric learning and applications,” *Information Sciences*, vol. 281, pp. 674-686, 2014. Article (CrossRef Link)

[16] B. Liu, M. Wang, R. Hong, Z.J. Zha, X.S. Hua, "Joint learning of labels and distance metric," in *Proc. of IEEE Transactions on Systems, Man and Cybernetics*, vol. 40, no. 3, pp. 973-978, 2010. Article (CrossRef Link)

[17] M. Wang, X.S. Hua, J. Tang, R. Hong, "Beyond distance measurement: constructing neighborhood similarity for video annotation," in *Proc. of IEEE Transactions on Multimedia*, vol. 11, no. 3, pp. 465-476, 2009. Article (CrossRef Link)

[18] S. Roweis, L. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323-2326, 2000. Article (CrossRef Link)

[19] J.C. Caicedo, J. BenAbdallah, F.A. González, O. Nasraoui, "Multimodal representation, indexing, automated annotation and retrieval of image collections via non-negative matrix factorization," *Neurocomputing*, vol. 76, no.1, pp. 50-60, 2012. Article (CrossRef Link)

[20] T. Jaakkola and D. Haussler, "Exploiting generative models in discriminative classifiers," *NIPS*, pp. 487-493, 1999.

[21] T. Jebara, R. Kondor, A. Howard, "Probability product kernels," in *Proc. of Journal of Machine Learning Research*, vol. 5, pp. 819-844, 2004.

[22] N. Vasconcelos, "On the efficient evaluation of probabilistic similarity functions for image retrieval," in *Proc. of IEEE Transactions on Information Theory*, vol. 50, no. 7, pp.1482-1496, 2004. Article (CrossRef Link)

[23] C. Schmid, "Constructing models for content-based image retrieval," *CVPR*, 2001. Article (CrossRef Link)

[24] A Perina, M. Cristani, U. Castellani, V. Murino, N. Jojic, "Free energy score spaces: using generative information in discriminative classifiers," in *Proc. of IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011. Article (CrossRef Link)

[25] X. Li, T.S. Lee, Y. Liu, "Hybrid generative-discriminative classification using posterior divergence," *CVPR*, 2011. Article (CrossRef Link)

[26] K. Weinberger, L. Saul, "Distance metric learning for large margin nearest neighbor classification," *The Journal of Machine Learning Research*, vol. 10, pp. 207-244, 2009.

[27] P. Jain, B. Kulis, J. Davis, I. Dhillon, "Metric and kernel learning using a linear transformation," *The Journal of Machine Learning Research*, vol. 13, pp. 519–547, 2012.

[28] B. Wang, X. Li, Y. Liu, "Learning discriminative Fisher kernel for image retrieval," in *Proc. of KSII Transaction on Internet and Information System*, vol. 7, no. 3, 2013.

[29] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Machine Learning*, vol. 42, pp. 177-196, 2001. Article (CrossRef Link)

[30] J. Su, W. Huang, P. Yu, V. Tseng, "Efficient relevance feedback for content-based image retrieval by mining user navigation patterns," in *Proc. of IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 3, pp. 360-372, 2011. Article (CrossRef Link)

[31] H. Cai, K. Mikolajczyk, J. Matas, "Learning linear discriminant projections for dimensionality reduction of image descriptors," in *Proc. of IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 2, pp. 338-352, 2011. Article (CrossRef Link)

[32] E. Xing, A. Ng, M. Jordan, S. Russell, "Distance metric learning, with application to clustering with side-information," *NIPS*, pp. 505-512, 2002.

[33] A. Frome, Y. Singer, J. Malik, "Image retrieval and classification using local distance functions," *NIPS*, 2007.

[34] L. Yang, R. Jin, R. Sukthankar, "Bayesian active distance metric learning," *arXiv preprint arXiv* vol. 1206, no. 5283, 2012.

[35] J. Kim, C. Shen, L. Wang, "A scalable algorithm for learning a Mahalanobis Distance Metric," *ACCV*, 2010. Article (CrossRef Link)

[36] Y. Ying, P. Li, "Distance metric learning with eigenvalue optimization," *The Journal of Machine Learning Research*, vol. 13, pp. 1-26, 2012.

[37] S. Xiang, F. Nie, C. Zhang, "Learning a mahalanobis distance metric for data clustering and classification," *Pattern Recognitio*, vol. 41, no. 12 pp. 3600-3612, 2008. Article (CrossRef Link)

[38] H. Becker, M. Naaman, L. Gravano, "Learning similarity metrics for event identification in social media," in *Proc. of ACM international conference on Web search and data mining*, pp. 291-300, 2010. Article (CrossRef Link)

[39] S. Cao, N. Snavely, "Learning to match images in large-scale collections," in *Proc. of ECCV Workshops and Demonstrations*, Springer, pp. 259–270, 2012. Article (CrossRef Link)

[40] M. Wang, X.S. Hua, R. Hong, J. Tang, "Unified video annotation via multigraph learning," in *Proc. of IEEE Transactions on Circuits and Systems for Video Technology*, vol. 19, no. 5, pp. 733-746, 2009. Article (CrossRef Link)

[41] A. Bosch, A. Zisserman, X. Muoz, "Scene classification using a hybrid generative discriminative approach," in *Proc. of IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 4, 2008. Article (CrossRef Link)

[42] M. Jordan, Z. Ghahramani, T. Jaakkola, and S. Lawrence, "Introduction to variational methods for graphical models," *Machine Learning*, vol. 37, pp.183-233, 1999. Article (CrossRef Link)

[43] P. Duygulu, K. Barnard, N. de Freitas, and D. Forsyth, "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," *ECCV*, 2002. Article (CrossRef Link)

[44] M. J. Huiskes, M. S. Lew, "The MIR Flickr retrieval evaluation," in *Proc. of ACM International Conference on Multimedia Information Retrieval*, 2008. Article (CrossRef Link)

[45] G. Carneiro, A. B. Chan P. J. Moreno, and N. Vasconcelos, "Supervised learning of semantic classes for image annotation and retrieval," in *Proc. of IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 3, pp. 394-410, 2006. Article (CrossRef Link)

[46] K. Van De Sande, T. Gevers, C. Snoek, "Evaluating color descriptors for object and scene recognition," in *Proc. of IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1582-1596, 2010. Article (CrossRef Link)

[47] Z. Li, Z. Shi, X. Liu and Z. Shi, "Modeling continuous visual features for semantic image annotation and retrieval," *Pattern Recognition Letters*, vol. 32, no. 3, pp. 516-523, 2010. Article (CrossRef Link)

[48] Z. Li, Z. Shi, X. Liu, Z. Li and Z. Shi, "Fusing semantic aspects for image annotation and retrieval," *Journal of Visual Communication and Image Representation*, vol. 21, no. 8, pp. 798-805, 2010. Article (CrossRef Link)

Xiong Li received the PhD degree in pattern recognition and intelligence system from Shanghai Jiao Tong University, China, in 2013. He is currently an engineer in National Computer Network Emergency Response Technical Team, China. His research interests include hybrid generative discriminative learning and probabilistic graphical model.

Qi Lv received the BS and MS degrees in Flight Vehicle Propulsion Engineering from Nanjing University of Aeronautics and Astronautics, China, in 2002 and 2005 respectively, and PhD degree in mathematics from Zhengzhou University, China, in 2010. His research interests include pattern recognition, statistics and Internet public opinion.

Wenting Huang received his BS degree in computer science and technology from Jilin University, China, in 2004, and the MS degree in computer science from Computer Network Information Center of the Chinese Academy of Sciences in 2007. He is currently an engineer in National Computer network Emergency Response technical Team of China. His research interests include multimedia