Learning Discriminative Fisher Kernel for Image Retrieval

Bin Wang, Xiong Li *, Yuncai Liu
Department of Automation, Shanghai Jiao Tong University
Shanghai 200240, China
[e-mail: lixiong@sjtu.edu.cn]
*Corresponding author: Xiong Li

Received December 10, 2012; revised February 8, 2013; accepted March 4, 2013; published March 29, 2013

Abstract

Content based image retrieval has become an increasingly important research topic for its wide application. It is highly challenging when facing to large-scale database with large variance. The retrieval systems rely on a key component, the predefined or learned similarity measures over images. We note that, the similarity measures can be potential improved if the data distribution information is exploited using a more sophisticated way. In this paper, we propose a similarity measure learning approach for image retrieval. The similarity measure, so called Fisher kernel, is derived from the probabilistic distribution of images and is the function over observed data, hidden variable and model parameters, where the hidden variables encode high level information which are powerful in discrimination and are failed to be exploited in previous methods. We further propose a discriminative learning method for the similarity measure, i.e., encouraging the learned similarity to take a large value for a pair of images with the same label and to take a small value for a pair of images with distinct labels. The learned similarity measure, fully exploiting the data distribution, is well adapted to dataset and would improve the retrieval system. We evaluate the proposed method on Corel-1000, Corel5k, Caltech101 and MIRFlickr 25,000 databases. The results show the competitive performance of the proposed method.

Keywords: Image retrieval, Fisher kernel, supervised similarity learning, GMMs

This work was supported by National Basic Research Program of China 2011CB302203, China NSFC Key Program 60833009, NSFC Program 60975012.

http://dx.doi.org/10.3837/tiis.2013.03.007

1. Introduction

Retrieving images according to users' interests from database is particular valuable for search engines. Approaches based on textual metadata or keywords face with a number of challenges [1]. First, it is inherently difficult to describe the content of images using textual metadata, and the description is made even harder for the diversification and rapid growth of image datasets. Second, manual annotation for large image datasets is prohibitively expensive, while the surrounding or contextual text is unreliable. Under such circumstances, content based image retrieval (CBIR) is particularly promising since it does not need textual metadata or keywords to describe the image content.

The CBIR systems [2] are composed of two core components: the feature representation of images and the similarity measures over image features (we do not distinguish similarity measure and distance measure because they are convertible). For *feature representation*, CBIR usually represents each image using a set of descriptors which are expected to be well descriptions of the semantic content of images. The *similarity measures* are defined over the features and expected to reflect the similarity of the semantic content of images. The common target [3] of the two components is to merge the so called semantic gap between low-level features and high-level semantic content. In this paper, we focus on the similarity measures. Some CBIR systems adopt predefined similarity measures [4], e.g., Euclidean distance, Gaussian kernel and L1 distance. However, predefined distance measures share some limitations. Especially they fail to adapt the data distribution [5] which varies along databases. To alleviate the semantic gap in CBIR, a much more promising way, is to learn the similarity measure from the database. Considerable machine learning techniques [3,6,7,8,9] have been used to learn the similarity metric from data over the past few years. According to the usage of label information, algorithms for similarity metric learning can be categorized into two classes: unsupervised methods and and supervised methods [3].

Unsupervised methods attempt to find an underlying low dimensional embedding or a similarity measure from high dimensional input data, under certain criterion. An effective criterion is to keep the geometric relationships among most of the observed data. Approaches under this criterion include principle component analysis (PCA) [10], locality constrained linear coding (LLC) [6], locally linear embedding (LLE) [7]. Although these methods are good at utilizing unlabeled data which is easy to obtain, they fail to exploit class label which is very informative in similarity learning. *Supervised methods* aim to learn similarity metrics by keeping the data within the same classes close and keeping the data of different classes separated. Representative approaches include neighborhood components analysis (NCA) [8], large margin nearest neighborhood classification (LMNN) [11], local distance metric learning (LDML) [9] and linear transformation based metric learning (LTML) [33]. As a common insight from the above discussions, learning based metrics are adaptive to data much better. Those approaches, however, do not fully exploit the distribution information which are shown to be very information in image representation [28].

A probabilistic branch of methods, probabilistic similarity [29,39-43], recently received increasing attention. These methods derive the explicit feature mapping or similarity measure based on the probabilistic distribution over the data. Consequently, they are able to exploit the abilities of probability models, e.g. dealing with structured data and exploiting hidden variables. The representative methods include probability product kernels [39], Kullback

Leibler divergence based similarity [40], Fisher kernel [29], free energy score space [42]. These methods can be unsupervised or supervised. However, of them, unsupervised methods are unable to exploit label information, while supervised methods are not flexible enough, i.e., unable to embed into specified classifiers.

In this paper, we propose a similarity learning algorithm, discriminative Fisher kernel (DFK) , for CBIR. It is able to exploit label information and embed to any classifier. First, we employ Gaussian Mixture Models (GMMs) to model the distribution of image features, which has been validated to be an effective way [28]. Second,we derive Fisher kernel [29] based on the GMMs, where Fisher kernel is a similarity function over the observed variable (image features), hidden variables (indicators of mixture centers) and model parameters. Third, to exploit label information, we propose a supervised learning method for Fisher kernel. There are two advantages of the proposed method: (1) the probabilistic modeling allows us to exploit hidden information and well adapt to data distribution; (2) the discriminative learning method fully utilizes label information in a computationally effective way.

The remainder of the paper is organized as follows. Section 2 reviews the related works. Section 3 presents the details of Fisher kernel based similarity learning approach. Our approach is compared with the state-of-the-art approaches over three popular image databases in Section 4. Section 5 draws a conclusion.

2. Related Works

In this paper, we focus on supervised similarity learning based on probabilistic similarity (see the categorization of similarity learning methods in Section 1). We in this section review the supervised similarity learning and probabilistic similarity methods, leaving other methods out.

Supervised similarity metric learning approaches attempt to learn a similarity metric from a set of equivalence constraints (for image pair within the same class) and inequivalence constraints (for image pair of the different classes) between images. The optimal distance metric is eventually found by keeping images in equivalence constraints close and images in inequivalence constraints well separated. Xing et al. [12] formulates the task into a constrained convex optimization problem by minimizing the distance between images in the same classes such that images from different classes are well separated. Relevant Components Analysis (RCA) [13] makes use of side information to learn a Mahalanobis distance from the equivalence constraints. Discriminative Component Analysis (DCA) and kernel DCA [14] extend RCA by incorporating equivalence constraints and exploring nonlinear transformation from context. Neighborhood Component Analysis (NCA) [8] extends nearest neighbor classifiers to component analysis, and is further extended to Large-Margin Nearest Neighbor (LMNN) [11] by consider margin. [16] learns local perceptual distance functions for image retrieval and classification, where the distance function is a combination of several local distance functions. [34] aims to learn a Mahalanobis distance metric from pairwise constraints in the form of must-links (i.e. links indicating the pair of data points must in the same class) and cannot-links (i.e. links indicating the pair of data points must in different classes). [35] takes context information associated with media content into consideration when learning similarity metrics. [36] proposes an image matching approach which leverages discriminative learning techniques to compute a better similarity metric for predicting whether two images are similar. Many recent studies [17-22,33] focus on the cooperation of metric learning, relevance feedback, dimensionality reduction, Bayesian learning and kernel learning.

Probabilistic similarity methods start from the modeling of the data distribution which encodes the interior information of the data. Mathematically, the probabilistic similarity is a function over the quantities of the data distribution. Probability product kernels [39] treat the posteriors for given samples as the representation of samples, and define the similarity as the expectation of the product of the two posterior. [40] represents samples as some distributions and uses Kullback–Leibler divergence to measure the similarity of samples. [41] builds a hierarchical probabilistic model to learn image representation and similarity. Fisher kernels (FK) [29] derive explicit feature mapping for samples by considering how the samples affect the model parameters, and define the similarity as the inner product of the feature mappings of any pair of samples. Free energy score space (FESS) [42] and posterior divergence (PD) [43] extend FK by considering more informative measures. Although these methods are able to utilize label information in terms of class conditional modeling, their abilities can be enhanced by means of joint learning of similarity and probabilistic models. As the current researches [42,43] show the highly competitive performance of score space methods, we in this paper will extend FK to a discriminative learning paradigm.

The advantages of our method are twofold: (1) compared with non-probabilistic similarity learning method, our method can fully exploit data distribution information (e.g., hidden variables); (2) compared with probabilistic similarity learning method, our method provides a sophisticated way to utilize label information and can be embedded into any classifier.

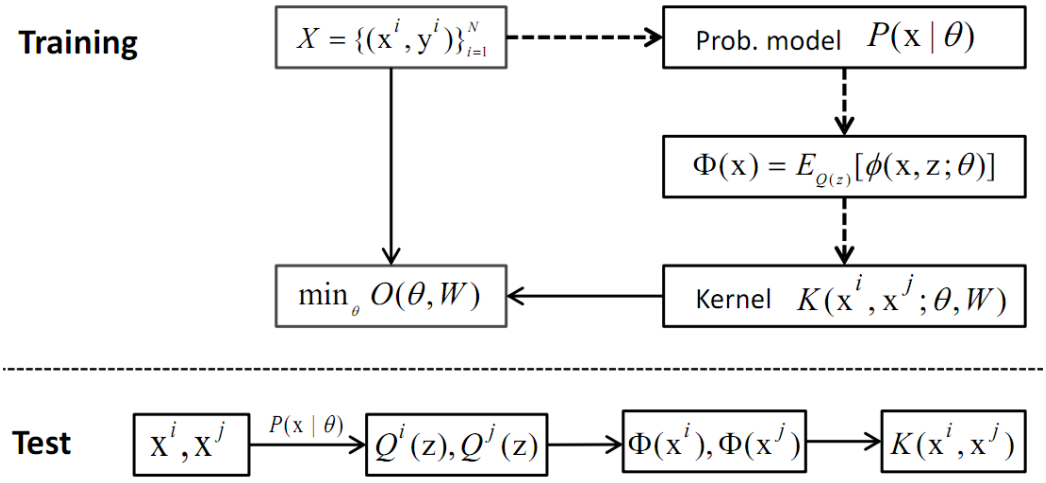


Fig. 1. The graphical illustration of learning discriminative Fisher kernel.

3. Learning Discriminative Fisher Kernels

In this section, we proceed to derive the Fisher kernel and propose a discriminative learning approach for the kernel. We first employ Gaussian Mixture Models (GMMs) to model the distribution because of its effectiveness in image feature modeling [28]. Then we derive the Fisher kernel [29] based on GMMs. At last, we further propose a discriminative learning method for Fisher kernel. See **Fig. 1** for the illustration of the proposed method.

3.1. Gaussian Mixture Models

We here use Gaussian Mixture Models (GMMs) to model the distribution of the image features for its effectiveness in image feature modeling [28]. Let $\mathbf{x} \in \mathrm{R}^D$ be the observed variable (image feature) and $\mathbf{z} = \{z_1, \cdots, z_K\}$ be a set of hidden variables (indicator) following the Multinomial distribution over K possible events,

$$P(\mathbf{z}) = \prod_{k=1}^{K} \alpha_k^{z_k}, \quad where \;\; z_k = \{0,1\}, \sum\nolimits_k z_k = 1, \alpha_k \geq 0, \sum\nolimits_k \alpha_k = 1$$

where $P(z_k) = \alpha_k$. Note that $\mathbf{z}$ is an indication vector. For GMMs, $\mathbf{z}$ indicates that which mixture center is selected to generate the sample of $\mathbf{x}$. A mixture center here is a Gaussian distribution. Then the conditional distribution, given $\mathbf{z}$, is,

$$P(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} \left[\frac{1}{(2\pi)^{D/2} |\Sigma_k|^{1/2}} \exp\left\{ -\frac{1}{2} (\mathbf{x} - \mathbf{u}_k)^T \Sigma_k^{-1} (\mathbf{x} - \mathbf{u}_k) \right\} \right]^{z_k}$$

where $\mathbf{u}_k$ and Σ_k are the mean and covariance matrix of the k-th component. Then the joint distribution of GMMs can be expressed as,

$$\begin{aligned} P(\mathbf{x}, \mathbf{z}; \theta) &= P(\mathbf{x}|\mathbf{z}) P(\mathbf{z}) = \prod_{k=1}^{K} N(\mathbf{x}; \mathbf{u}_k, \Sigma_k)^{z_k} \prod_{k=1}^{K} \alpha_k^{z_k} \\ &= \prod_{k=1}^{K} \left[\frac{1}{(2\pi)^{D/2} |\Sigma_k|^{1/2}} \exp\left\{ -\frac{1}{2} (\mathbf{x} - \mathbf{u}_k)^T \Sigma_k^{-1} (\mathbf{x} - \mathbf{u}_k) \right\} \right]^{z_k} \prod_{k=1}^{K} \alpha_k^{z_k} \end{aligned} \tag{1}$$

where Let $\theta = \{\mathbf{u}_k, \Sigma_k\}_{k=1}^{K}$. For computational efficiency, we assume that the covariance matrixes Σ_k are diagonal, i.e., $\Sigma_k = \mathrm{diag}(\sigma_{k1}^2, \cdots, \sigma_{kD}^2)$. Note that this assumption would not bring degeneration to the performance in practice [28]. The marginal distribution of GMMs is the integration of $P(\mathbf{x}, \mathbf{z}; \theta)$ (Eq. (1)) over $\mathbf{z}$,

$$P(\mathbf{x}; \theta) = \sum_{\mathbf{z}} P(\mathbf{x}, \mathbf{z}; \theta) = \sum_{k=1}^{K} \frac{\alpha_k}{(2\pi)^{D/2} |\Sigma_k|^{1/2}} \exp\left\{ -\frac{1}{2} (\mathbf{x} - \mathbf{u}_k)^T \Sigma_k^{-1} (\mathbf{x} - \mathbf{u}_k) \right\} \tag{2}$$

The GMMs can be learned using Expectation-Maximization (EM) algorithm [32] that alternatively maximizes the log likelihood function with respect to the posterior of $\mathbf{z}$ (E-step or inference step) and the parameters θ (M-step or parameter estimation step). As suggested in [43], we let $Q^i(\mathbf{z}) = \prod_{k=1}^{K} (g_k^i)^{z_k}$ be the posterior of hidden variables, conditioned on $\mathbf{x}^i$. The E-step updates the posterior of the hidden variable, for each observed sample,

$$g_k^i = \frac{\alpha_k N(\mathbf{x}^i; \theta_k)}{\sum_k \alpha_k N(\mathbf{x}^i; \theta_k)} \tag{3}$$

where $g_k^i = P(z_k | \mathbf{x}^i)$ is the probability assigning the sample $\mathbf{x}^i$ to the mixture center k. The M-step updates the parameters of GMMs,

$$\alpha_k = \frac{1}{N} \sum_{i=1}^{N} g_k^i, \quad \mathbf{u}_k = \sum\nolimits_{i=1}^{N} g_k^i \mathbf{x}_i, \quad \sigma_{kd}^2 = \frac{1}{N} \sum_{i=1}^{N} g_k^i (x_d^i - u_{kd})^2. \tag{4}$$

The learning algorithm for GMMs is the iteration of the E-step and M-step.

3.2. Normalized Fisher kernel

Although the analytic form of the $P(\mathbf{x};\theta)$ is available in Eq. (2), the differential operation over $\log P(\mathbf{x};\theta)$ is pretty complex because $P(\mathbf{x};\theta)$ takes the form of summation. We resort to the variational lower bound [30] of the log likelihood function, on which the differential operation will be very simple,

$$\begin{aligned}\log P(\mathbf{x};\theta) &\geq -\mathrm{KL}(Q(\mathbf{z}) \| P(\mathbf{x},\mathbf{z};\theta)) = F(\theta) \\ &= -\mathrm{E}_{Q(\mathbf{z})}[\log Q(\mathbf{z}) - \log P(\mathbf{x},\mathbf{z};\theta)] \\ &= -\mathrm{E}_{Q(\mathbf{z})}\left[\sum\nolimits_k z_k \log\frac{\alpha_k}{g_k} + \sum\nolimits_k z_k\left(\log(2\pi)^{D/2}|\Sigma_k|^{1/2} + \frac{1}{2}(\mathbf{x}-\mathbf{u}_k)^T\Sigma_k^{-1}(\mathbf{x}-\mathbf{u}_k)\right)\right] \\ &= -\sum\nolimits_k g_k \log\frac{\alpha_k}{g_k} - \sum\nolimits_k g_k\left(\log(2\pi)^{D/2}|\Sigma_k|^{1/2} + \frac{1}{2}(\mathbf{x}-\mathbf{u}_k)^T\Sigma_k^{-1}(\mathbf{x}-\mathbf{u}_k)\right)\end{aligned} \tag{5}$$

where $Q(\mathbf{z})$ is the approximate posterior of $\mathbf{z}$ and takes the same parameterizations with $P(\mathbf{z})$. It is worth noting that using the lower bound will not loss generality, because the lower bound equals to the real log likelihood and $Q(\mathbf{z})$ equals to the real posterior when using exact inference. Having the lower bound $F(\theta)$ of $\log P(\mathbf{x};\theta)$, the elements of Fisher score is its gradient with respect to model parameters [29],

$$\begin{aligned}\frac{\partial F(\theta)}{\partial u_{kd}} &= \sum\nolimits_k g_k \Sigma_k^{-1}(x_d - u_{kd}) \\ \frac{\partial F(\theta)}{\partial \sigma_{kd}} &= -\sum\nolimits_k g_k\left(\frac{1}{\sigma_{kd}} - \frac{1}{\sigma_{kd}^{-3}}(x_d - u_{kd})^2\right) \\ \frac{\partial F(\theta)}{\partial a_k} &= -\frac{g_k}{\alpha_k}\end{aligned} \tag{6}$$

Note that the elements of Fisher score is the expectation over a function of the observed variable $\mathbf{x}$, hidden variables $\mathbf{z}$ and model parameters θ, where the hidden variables allow Fisher kernel to exploit hidden (high level) information and model parameters allow it to adapt to data distribution. The complete Fisher score is the combination of those gradients,

$$\Phi(\mathbf{x}) = \mathrm{vec}\left(\left\{\frac{\partial F(\theta)}{\partial u_{kd}}, \frac{\partial F(\theta)}{\partial \sigma_{kd}}, \frac{\partial F(\theta)}{\partial \alpha_k}\right\}_{kd}\right) \tag{7}$$

The normalized Fisher kernel then can be defined as [31],

$$K(\mathbf{x}^i, \mathbf{x}^j) = \frac{\exp(-(\Phi^i - \Phi^j)^T W(\Phi^i - \Phi^j))}{\sum_{k\neq i}\exp(-(\Phi^i - \Phi^k)^T W(\Phi^i - \Phi^k))} \tag{8}$$

where $W \succ 0$ is a weight matrix and assumed to be diagonal, i.e., $W = \mathrm{diag}(w_1, \cdots, w_{D_\Phi})$ where D_Φ is the number of dimension of Φ. Functionally, w_d weights the domination of the d-th dimension of Φ to the similarity, i.e., a dimension with large weight dominates much

than a dimension with small weight. In particular, $w_d = 0$ indicates that the d-th dimension is completely uninformative. Now we have the parameterized Fisher kernel. In the next section, we will present the method to determine these parameters.

3.3. Discriminative learning of Fisher kernel

Let $\mathrm{y}^i = (y_1^i, \cdots, y_C^i)$ be the label vector of the sample x^i, where $y_c^i = 1$ iif the c-th label of all C labels belongs to the sample x^i and $y_c^i = 0$ otherwise. We consider the 1-NN classifier that favors high similarity for samples with same classes. Besides, we also expect that samples which are from different classes have low similarities.

$$O(\theta, W) = \sum_i \sum_{j \neq i} s(\mathbf{y}^i, \mathbf{y}^j) K(\mathbf{x}^i, \mathbf{x}^j) \tag{9}$$

where $s(\mathbf{y}^i, \mathbf{y}^j)$ is a similarity measure over the two label vectors, and takes a positive value (encourage) if they have common labels and takes negative value (inhibition) if they have no common label. We choose the sigmoid based function:

$$s(\mathbf{y}^i, \mathbf{y}^j) = \frac{2}{1 + \exp(2\langle \mathrm{y}^i, \mathrm{y}^j \rangle + 1)} - 1$$

Given the posterior $Q(\mathbf{z})$ over the hidden variable, we minimize the objective function $O(\theta, W)$ using gradient descent,

$$\begin{aligned} \frac{\partial O(\theta, W)}{\partial \theta} &= \sum_i \sum_{j \neq i} c(\mathbf{y}^i, \mathbf{y}^j) K(\mathbf{x}^i, \mathbf{x}^j)(\phi^i - \phi^j) \circ \left(\frac{\partial \phi^i}{\partial \theta} - \frac{\partial \phi^j}{\partial \theta}\right) \\ &\quad - \sum_i \left(\sum\nolimits_j K(\mathbf{x}^i, \mathbf{x}^j)\right) \sum_{j \neq i} K(\mathbf{x}^i, \mathbf{x}^j)(\phi^i - \phi^j) \circ \left(\frac{\partial \phi^i}{\partial \theta} - \frac{\partial \phi^j}{\partial \theta}\right) \end{aligned} \tag{10}$$

$$\frac{\partial O(\theta, W)}{\partial W} = -2W \sum_i \sum_{j \neq i} \left[c(\mathbf{y}^i, \mathbf{y}^j) + \left(\sum\nolimits_j K(\mathbf{x}^i, \mathbf{x}^j)\right) \right] K(\mathbf{x}^i, \mathbf{x}^j) \| \phi^i - \phi^j \|^2 \tag{11}$$

where $Q(\mathbf{z})$ is the pairwise multiplication.

The complete learning procedure is the iteration of the E-step (Eq. (3)) and M-step (Eq. (10) and Eq. (11)), until it converges, which is summarized in Algorithm 1.

Algorithm 1 Discriminative learning of Fisher kernel

1: Input: training set $X=\{(\mathrm{x}^i,\mathrm{y}^i)\}_{i=1}^N$; iteration number T; learning rate γ

2: initialize parameters θ^0, W^0

3: for $t=1$ to T do

4: $$g_k^i=\frac{\alpha_k N(\mathbf{x}^i;\theta_k)}{\sum_k \alpha_k N(\mathbf{x}^i;\theta_k)}$$

5: $$\theta^t \leftarrow \theta^{t-1}-\gamma\frac{\partial O(\theta,W)}{\partial \theta}$$

6: $$W^t \leftarrow W^{t-1}-\gamma\frac{\partial O(\theta,W)}{\partial W}$$

7: end for

8: Output: θ^T, W^T

In the test step, the Fisher kernel similarity of any pair of images $\mathrm{x}^i, \mathrm{x}^j$ can be computed using Algorithm 2.

Algorithm 2 Computing the Fisher kernel similarity

1: Input: a pair of images $\mathrm{x}^i, \mathrm{x}^j$

2: infer the posterior parameters $g_k^i=\frac{\alpha_k N(\mathbf{x}^i;\theta_k)}{\sum_k \alpha_k N(\mathbf{x}^i;\theta_k)}$ for x^i (Eq.(3))

3: infer the posterior parameters $g_k^j=\frac{\alpha_k N(\mathbf{x}^j;\theta_k)}{\sum_k \alpha_k N(\mathbf{x}^j;\theta_k)}$ for x^j (Eq.(3))

4: compute the Fisher kernel similarity using Eq.(6)-Eq.(8)

5: Output: $K(\mathbf{x}^i,\mathbf{x}^j)$

4. Experiments

In this section, we proceed to evaluate the proposed method, i.e., discriminative learning of Fisher kernel, on four datasets for image retrieval. We compare the proposed method with several related methods and several state-of-the-art methods in image retrieval.

4.1. Databases

We evaluate the proposed approach on four benchmark image databases: Corel-1000 [14] database, Corel5k [23] database, Caltech101 database [24], and MIRFlickr 25,000 database [44]. Some sample images of the dabases used in our experiments are shown in **Fig. 2**.

Fig. 2. Sample images from the databases used in our experiments

Corel-1000 and Corel5k are two real-world image databases, both of which are subsets of the Corel Photo Gallery. The Corel-1000 database includes 10 categories of images, such as roses, cats, horses, eagles, etc. Each category has a different semantic meaning and comprises of 100 images. There are totally 1,000 images in the Corel-1000 database. The Corel5k database is composed of 5,000 images from 50 categories and each category contains 100 images. Each category represents certain semantic content such as beach, tile, wave, food texture, tigers, France, bears, autumn, and tropical plants, etc. The Caltech101 database is used for larger scale experiments, which contains 9,196 images. These images are classified into 101 categories, such as chair, barrel, anchor and dolphin, etc. Different from Corel-1000 and Corel5k databases, the number of images varies along category in Catech101 database. The MIRFlickr-25000 database is comprised of 25,000 images downloaded from the online photo-sharing service Flickr [44] with high-resolution images and text annotations. Those images were collected from the web directly to provide a realistic large scale database for image retrieval research. The way to use images with semantic categories can help us to evaluate the image retrieval performance in an automatical manner, which reduces the subjective error induced by manual evaluation.

4.2. Image Representation

Feature representation is very important for a CBIR system due to the diverse visual contents of image databases. Here, we represent images using color SIFT descriptors for its excellent discrimination power [25]. The performance of color SIFT descriptors have been validated to be effective in image annotation and retrieval. In our experiments, four color SIFT descriptors (OpponentSIFT, C-SIFT, rgSIFT and RGB-SIFT) recommended by [25] are used to represent the visual content of images. To combine these color SIFT descriptors, we simultaneously use dense sampling and Harris-Laplace point sampling, which are followed by leveraging spatial pyramid.

4.3 Performance Measure

We use a standard performance measure, mean average precision (MAP), in our comparative experiments. MAP, widely used in image retrieval, gives a summarized measure of the precision-recall curve. Precision for image retrieval can be defined as the percentage of images whose ground truth annotations contain the same label as the query image. Average precision (AP) focuses on ranking relevant images higher [23], and is the average of the precision values at the ranks where relevant items occurs. MAP is then given by averaging AP over all the query keywords.

4.4 Compared Methods and Experimental Setting

We will compare our approach, discriminative Fisher kernels, with the baseline method (i.e., predefined Euclidean distance) as well as other similarity learning methods. Specifically, compared approaches are listed as follows:

Euclidean. A baseline method without using metric learning.

Xing. Learning the distance metric with nonlinear optimization [12], which allows to derive efficient, local-optima free algorithms.

DCA. Discriminative components analysis [14], which makes an improvement of RCA by using inequivalence constraints.

SDPM. A fast and scalable algorithm to learn a Mahalanobis distance [37], which formulates the distance metric learning as a convex optimization problem

DML-eig. A metric learning method based on an eigenvalue optimization framework [38].

LMNN. Large margin nearest neighbor classification [11] learns a Mahalanobis distance metric for kNN classification from labeled samples.

FESS. Free energy score space [42], which is a probabilistic similarity approach via extending FK by considering more information measures.

PD. Posterior divergence similarity measure derived from the probabilistic models over images [43], by measuring how samples affect the model parameters.

DFK. The proposed discriminative Fisher-kernel based similarity learning approach, which is much more adaptive to the database by considering the data distribution.

We evaluate the retrieval performance under the evaluation criteria in the leave-one-out manner. Specifically, a query image is chosen from the test database, and the rest images form the gallery. Then, the test image is queried by the above evaluated distance metrics. The retrieval performance is evaluated by the mean average precision(mAP). It is worth noting that, parameters of Fisher kernel, except the number of mixture centers K in Eq. (1), are learned from the dataset. We set $K = 60$ throughout the experiments. We will discuss the effect of the parameter in Section 4.6. With the learned kernel, for a query, the retrieval algorithm returns those images with high similarity with respect to the query.

Table 1. Average Precision for Corel-1000 database on different algorithms

Category	Euclid.	Xing	DCA	SDPM	DML-eig	LMNN	FESS	PD	DFK
Butterfly	0.310	0.345	0.390	0.380	0.385	0.378	0.384	0.390	**0.410**
Mountain	0.505	0.570	0.635	0.592	0.604	0.625	0.609	0.611	**0.645**
Dogs	0.420	0.390	0.500	0.420	0.435	0.437	0.442	0.450	**0.502**
Horses	0.775	0.830	0.850	0.835	0.821	**0.852**	0.809	0.807	0.820
Cats	0.495	0.640	0.600	0.610	0.615	0.605	0.621	0.630	**0.651**
Eagles	0.575	0.665	0.590	0.600	0.570	0.645	0.614	0.625	**0.670**
Roses	0.505	0.545	0.610	0.552	0.565	0.599	0.585	0.600	**0.630**
Sunset	0.570	0.560	0.395	0.551	0.561	0.486	0.561	0.565	**0.584**
Balloon	0.260	0.265	0.240	0.245	0.250	0.246	0.264	**0.266**	0.242
Penguins	0.215	0.260	0.470	0.306	0.329	0.315	0.330	0.364	**0.473**
mAP	0.463	0.507	0.528	0.509	0.514	0.519	0.522	0.531	**0.563**

4.5 Experiments Results

To validate the effectiveness of the proposed approach, we firstly perform an experiment on Corel-1000 database [14]. In this experiment, we randomly split each dataset into 70% for training and 30% for testing, with the former subset for the discriminative Fisher kernel learning and the latter subset serving as performance test dataset. For each compared approach, we measure the average precision for each category on the top returned images. More specifically, we perform comparison for each category on the top 20 returned images. The experimental results are summarized in **Table 1**. The values of mAP, obtained by averaged average precision, are used to evaluate the performance of each algorithm towards the whole Corel-1000 database. We find that, compared with the baseline approach, both Xing's approach and SDPM obtain a significant improvement. DCA, LMNN and DML-eig show competitive performance and outperform Xing's approach and SDPM. Meanwhile, FESS and PD, which are two probabilistic similarity measure learning approaches most close to our method, get better results due to the consideration of probabilistic modeling of image distribution. Our proposed DFK approach, as shown in **Table 1**, achieves the best performance among these compared approaches in most cases. Specifically, the DFK approach outperforms FESS and PD by 4.1% and 3.2% respectively. The reason accounting for this improvement is that DFK fully utilizes the label information while FESS and PD do not. These convincing results demonstrate the effectiveness of the proposed method for the image retrieval.

Table 2. Retrieval performance evaluation on Corel5k database using different algorithms

Algorithm	**mAP**
Euclidean	0.269
Xing	0.307
DCA	0.325
SDPM	0.315
DML-eig	0.309
LMNN	0.310
FESS	0.316
PD	0.320
DFK	**0.342**

Table 3. Retrieval performance evaluation on Caltech101 database using different algorithms

Algorithm	**mAP**
Euclidean	0.155
Xing	0.175
DCA	0.186
SDPM	0.182
DML-eig	0.179
LMNN	0.180
FESS	0.186
PD	0.187
DFK	**0.204**

We then conduct experiments on two larger databases: Corel5k and Caltech101 to further verify the ability of DFK in adapting to databases. These two datasets share the same experimental setting as Corel-1000. The mAP values evaluated on different metric learning approaches over Corel5k database are summarized in **Table 2**. We can see that Xing's approach and DML-eig show competitive performance, both achieving a significant improvement over the baseline Euclidean metric. Meanwhile, DCA performs the best in comparison with three other distance metric learning approaches, SDPM, DML-eig and LMNN. This is because DCA introduces negative constraints and captures complex nonlinear relationships among samples, which can be informative. FESS and PD outperform most of the above approaches, where FESS and PD respectively achieve 4.7% and 5.1% improvements over the baseline Euclidean approach. Obviously, DFK achieves the best performance among all. These results should be again credited to its ability in exploiting data distribution information and class label information in content based image retrieval. The experimental results for the Caltech101 database are shown in **Table 3**. We can see that, on this challenging database, the performance is similar to that on Corel-1000 and Corel5k. On the whole, our proposed approach again outperforms other methods. This demonstrates the fact that the proposed approach can adapt to databases via probabilistic modeling of images and utilize label information by the discriminative learning of Fisher kernel. Specifically, when compared with the baseline approach, Xing's approach, DCA, SDPM, DML-eig and LMNN show superiority. Simultaneously, SDPM shows competitive performance with DCA. Also, probabilistic approaches (FESS and PD) exhibit superiority over Xing's approach, SDPM, DCA, DML-eig and LMNN. While, as shown in **Table 3**, our DFK outperforms FESS and PD up to 1.8%. This is due to the successful exploitation of label information.

When large-scale image collections come into the view, the underlying similarity function should be able to characterize the content-level similarity between images with large variance. To demonstrate the effectiveness of the proposed approach on large-scale image database, we evaluate our DFK on MIRFlickr database [44]. In this experiment, we follow the conventional training-validation scheme. Specifically, 15,000 images are used for training and the rest 10,000 images are used for test. We randomly select 1000 images from the test dataset as queries and use the rest 24,000 images as the gallery. Among them, 15,000 images are with text annotations while the remaining 9,000 images are not. This is a relatively realistic setup. We compare DFK with several approaches: Euclidean (the baseline approach), non-negative matrix factorization (NNMF) [44], LMNN [11], FESS [42], and PD [43], where NNMF is a state-of-the-art image retrieval approach; LMNN is a popular supervised distance metric learning method; FESS and PD are two probabilistic similarity learning approaches closely related to our proposed method. The results are summarized in **Table 4**. We observe that NNMF and LMNN both make significant improvements over the baseline approach. The performances of FESS and PD are superior to NNMF and LMNN, because they exploit the data distribution in a more sophisticated way than NNMF and LMNN. Compared with FESS and PD, our DFK fully exploits label information which in very informative in content based image retrieval. Overall, the proposed DFK achieves the best performance among the approaches compared with.

Table 4. Retrieval performance of different algorithms on MIRFlickr database.

Algorithm	**mAP**
Euclidean	0.455
NNMF	0.583
LMNN	0.586

FESS	0.590
PD	0.595
DFK	**0.619**

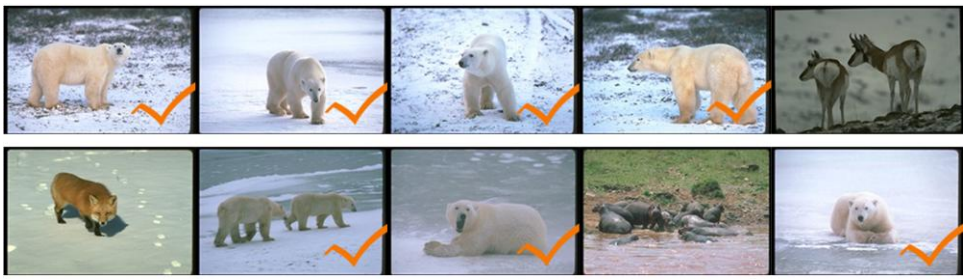


Fig. 3. Retrieval results for the "bear" query. The figure shows the top 10 images returned by our approach. The first image is the query image and the relevant images are marked with a tick symbol.

Fig. 3 shows an example of the image retrieval performed by our DFK approach on Corel5k dataset. Given the query image, the system automatically returns the relevant images. We have chosen the top ten images here. We can observe that most of the returned images are relevant with the query image. Similar results can be found in **Fig. 4** and **Fig. 5**. Note that in most cases, the returned images are meaningful. In sum, the overall experimental results demonstrate that the proposed approach are empirically much more effective to learn good quality similarity metric than the previous approaches for the performance improvement of image retrieval.

Fig. 4. Retrieval results for the "rose" query. The figure shows the top 10 images returned by our approach. The first image is the query image and the relevant images are marked with a tick symbol.

Fig. 5. Retrieval results for the "airplane" query. The figure shows the top 10 images returned by our approach. The first image is the query image and the relevant images are marked with a tick symbol.

4.6 Discussions

In this section, we discuss the effects of the parameter K (Eq. (1)) and the weight function $s(\mathbf{y}^i, \mathbf{y}^j)$ (Eq. (9)) to the retrieval performance. In the above experiments, our observation on K is consistent with [43]. That is, Fisher kernel shows robustness to the number of mixture centers K. However, for the discriminative learning approach in this paper, we find that a relative small number $K = 60$ can produce satisfied results. This is different with [28] that sets $K = 256$. A benefit of setting a small K is that it can effectively reduce the computational cost. For the weight function $s(\mathbf{y}^i, \mathbf{y}^j)$, we find that (1) a negative value for $\left\langle \mathbf{y}^i, \mathbf{y}^j \right\rangle = 0$ can significantly improve the performance; (2) the method prefers a relative small value for $\left\langle \mathbf{y}^i, \mathbf{y}^j \right\rangle = 1$ and relative large value for others.

6. Conclusions

In this paper, we propose a data-distribution-aware similarity metric learning approach for content-based image retrieval. The approach is based on the Fisher kernel, which is derived from the probabilistic distribution of the datasets. The Fisher kernel is a function over the observed variable (image feature), hidden variables (indicator of mixture centers) and model parameters, and allows to fully exploit hidden information and well adapt to data distribution. The discriminative learning approach for Fisher kernel incorporates the label information and output a kernel which would improve the retrieval performance. Extensive experiments are conducted on four benchmark datasets. The convincing experimental results demonstrate the effectiveness of our proposed similarity metric learning approach for the image retrieval task.

References

[1] F. Faria, A. Veloso, H. Almeida, E. Valle, R. Torres, M. Gonc¸alves and W. Meira Jr, "Learning to rank for content-based image retrieval," in *Proc. of the international conference on Multimedia information retrieval, ACM*, pp. 285–294, 2010. Article (CrossRef Link)

[2] M. Arevalillo-Herr´aez, F. Ferri and 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)

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

[4] S. Hoi, W. Liu and S. Chang, "Semi-supervised distance metric learning for collaborative image retrieval and clustering," *ACM Transactions on Multimedia Computing, Communications and Applications (TOMCCAP)*, vol. 6, no. 3, 18, 2010. Article (CrossRef Link)

[5] A. Smeulders, M. Worring, S. Santini, A. Gupta and 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)

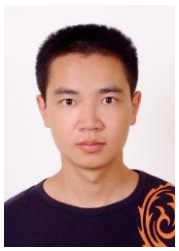
[6] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang and Y. Gong, "Locality-constrained linear coding for image classification," in Proc. *of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3360–3367, 2010. Article (CrossRef Link)

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

[8] J. Goldberger, S. Roweis, G. Hinton and R. Salakhutdinov, "Neighborhood components analysis."
[9] L. Yang, R. Jin, R. Sukthankar and Y. Liu, "An efficient algorithm for local distance metric learning," in *Proc. of the National Conference on Artificial Intelligence*, Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press. vol. 21, pp. 543, 2006.
[10] A. Webb, "Statistical pattern recognition," Wiley, 2003.
[11] K. Weinberger and L. Saul, "Distance metric learning for large margin nearest neighbor classification," *The Journal of Machine Learning Research*, vol. 10, pp. 207–244, 2009.
[12] E. Xing, A. Ng, M. Jordan and S. Russell, "Distance metric learning, with application to clustering with side-information," *Advances in neural information processing systems*, vol. 15, pp.505–512, 2002.
[13] A. Bar-Hillel, T. Hertz, N. Shental and D. Weinshall, "Learning distance functions using equivalence relations," in *Proc. of Machine Learning-International Work Shop Then Conference-*, vol. 20, pp. 11,2003.
[14] S. Hoi, W. Liu, M. Lyu and W. Ma, "Learning distance metrics with contextual constraints for image retrieval," in *Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 2, pp. 2072–2078, 2006. Article (CrossRef Link)
[15] T. Kim, S. Wong, B. Stenger, J. Kittler and R. Cipolla, "Incremental linear discriminant analysis using sufficient spanning set approximations," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2007. Article (CrossRef Link)
[16] A. Frome, Y. Singer and J. Malik, "Image retrieval and classification using local distance functions," in *19:* in *Proc. of the 2006 Conference Advances in Neural Information Processing Systems*, vol. 19, MIT Press, pp. 417, 2007.
[17] J. Su, W. Huang, P. Yu and V. Tseng, "Efficient relevance feedback for content-based image retrieval by mining user navigation patterns," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 3, pp. 360–372, 2011. Article (CrossRef Link)
[18] C. Ferreira, J. Santos, R. da S Torres, M. Gonc¸alves, R. Rezende and W. Fan, "Relevance feedback based on genetic programming for image retrieval," *Pattern Recognition Letters*, vol. 32, no. 1, pp. 27–37, 2011. Article (CrossRef Link)
[19] H. Cai, K. Mikolajczyk and J. Matas, "Learning linear discriminant projections for dimensionality reduction of image descriptors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 2, pp. 338–352, 2011. Article (CrossRef Link)
[20] J. Dillon, Y. Mao, G. Lebanon and J. Zhang, "Statistical translation, heat kernels and expected distances," *arXiv preprint arXiv*:1206.5248, 2012.
[21] L. Yang, R. Jin and R. Sukthankar, "Bayesian active distance metric learning," *arXiv preprint arXiv*:1206.5283, 2012.
[22] H. Chang and D. Yeung, "Kernel-based distance metric learning for content-based image retrieval," *Image and Vision Computing* 25 (5) (2007), pp. 695–703. Article (CrossRef Link)
[23] J. Caicedo, J. BenAbdallah, F. González and 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)
[24] R. Vieux, J. Benois-Pineau and J. Domenger, "Content based image retrieval using bag of-regions," *Advances in Multimedia Modeling,* pp. 507–517, 2012.
Article (CrossRef Link)
[25] K. Van De Sande, T. Gevers, C. Snoek, "Evaluating color descriptors for object and scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1582–1596, 2010. Article (CrossRef Link)
[26] A. Bar-Hillel, T. Hertz, N. Shental and D. Weinshall, "Learning a Mahalanobis metric from equivalence constraints," *Journal of Machine Learning Research*, vol. 6, no. 1, pp. 937, 2006.
[27] S. Hoi, M. Lyu, R. Jin, "A unified log-based relevance feedback scheme for image retrieval," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 4, pp. 509–524, 2006.
Article (CrossRef Link)

[28] K. Chatfield, V. Lempitsky, A. Vedaldi and A. Zisserman, "The devil is in the details: an evaluation of recent feature encoding methods," in *Proc. of British Machine Vision Conference*, 2011.
[29] T. Jaakkola and D. Haussler, "Exploiting generative models in discriminative classifiers," In *NIPS*, pp. 487–493, 1999.
[30] M. Jordan, Z. Ghahramani, T. Jaakkola and S. Lawrence, "Introduction to variational methods for graphical models," *Machine Learning*, 37, pp. 183-233, 1999.
Article (CrossRef Link)
[31] der Maaten and Laurens Van, "Learning discriminative fisher kernels," In *ICML*, pp. 217–224, 2011.
[32] J. Friedman, T. Hastie, and R. Tibshirani, "The Elements of Statistical Learning," *Springer*, 2008.
[33] P. Jain, B. Kulis, J. Davis and I. Dhillon, "Metric and kernel learning using a linear transformation," *The Journal of Machine Learning Research,* vol. 13, pp. 519–547, 2012.
[34] S. Xiang, F. Nie, C. Zhang, "Learning a mahalanobis distance metric for data clustering and classification," *Pattern Recognition*, vol. 41, no. 12, pp. 3600–3612, 2008. Article (CrossRef Link)
[35] H. Becker, M. Naaman and L. Gravano, "Learning similarity metrics for event identification in social media," in *Proceedings of the third ACM international conference on Web search and data mining*, ACM, pp. 291–300, 2010.
Article (CrossRef Link)
[36] S. Cao and N. Snavely, "Learning to match images in large-scale collections," in *Computer Vision–ECCV 2012. Workshops and Demonstrations*, Springer, pp. 259–270, 2012.
Article (CrossRef Link)
[37] J. Kim, C. Shen and L. Wang, "A scalable algorithm for learning a Mahalanobis Distance Metric," in *ACCV*, 2010. Article (CrossRef Link)
[38] Y. Ying and P. Li, "Distance metric learning with eigenvalue optimization," *The Journal of Machine Learning Research* , vol. 13, pp. 1–26, 2012.
[39] T. Jebara, R. Kondor, A. Howard, "Probability product kernels," *Journal of Machine Learning Research*, vol. 5, pp. 819-844, 2004.
[40] N. Vasconcelos, "On the efficient evaluation of probabilistic similarity functions for image retrieval," *IEEE Trans. on Information Theory*, vol. 50, no. 7, pp. 1482-1496, 2004.
Article (CrossRef Link)
[41] C. Schmid, "Constructing models for content-based image retrieval," In *CVPR* 2001.
Article (CrossRef Link)
[42] A Perina, M. Cristani, U. Castellani, V. Murino, N. Jojic, "Free energy score spaces: using generative information in discriminative classifiers," *IEEE Trans. on PAMI*, 2011.
Article (CrossRef Link)
[43] X. Li, T.S. Lee, Y. Liu, "Hybrid generative-discriminative classification using posterior divergence," In *CVPR* 2011. Article (CrossRef Link)
[44] 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)

Bin Wang received her BE degree in information and computional science from Shandong Normal University, Ji'nan, China, in 2006; and the MS degree in applied mathmatics from University of Science and Technology Beijing, Beijing, China. She is currently a PhD candidate at Department of Automation, Shanghai Jiao Tong University, Shanghai, China. Her research interests include computer vision, machine learning, image processing, multimedia analysis.



Xiong Li received his BE degree in Aircraft Power Engineering from Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2004; and the MS degree in pattern recognition and intelligent system from Southeast University, Nanjing, China. He currently is a PhD candidate at Department of Automation, Shanghai Jiao Tong University, Shanghai, China. His research interests include probabisltic graphical models and hybrid generative discriminative learning.

Yuncai Liu received the Ph.D. degree in the Department of Electrical and Computer Science Engineering in 1990 from the University of Illinois at Urbana-Champaign (UIUC), and worked as an associate researcher at the Beckman Institute of Science and Technology from 1990 to 1991. Since 1991, he had been a system consultant and then a chief consultant of research in Sumitomo Electric Industries, Ltd., Japan. In October 2000, he jointed the Shanghai Jiao Tong University as a distinguished professor. His research interests are in image processing and computer vision.