Signal Processing 93 (2013) 2284–2292

Contents lists available at SciVerse ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Discriminative codebook learning for Web image search ☆

Xinmei Tian [a],*, Yijuan Lu [b],**

[a] University of Science and Technology of China, Hefei, Anhui 230027, PR China
[b] Texas State University, San Marcos, TX 78666, USA

ARTICLE INFO

Article history:
Received 16 December 2011
Received in revised form
13 April 2012
Accepted 20 April 2012
Available online 11 May 2012

Keywords:
Codebook learning
Bag-of-visual word
Subspace learning
Web image search

ABSTRACT

Given the explosive growth of the Web images, image search plays an increasingly important role in our daily lives. The visual representation of image is the fundamental factor to the quality of content-based image search. Recently, bag-of-visual word model has been widely used for image representation and has demonstrated promising performance in many applications. In the bag-of-visual-word model, the codebook/visual vocabulary plays a crucial role. The conventional codebook, generated via unsupervised clustering approaches, does not embed the labeling information of images and therefore has less discriminative ability. Although some research has been conducted to construct codebooks with the labeling information considered, very few attempts have been made to exploit manifold geometry of the local feature space to improve codebook discriminative ability. In this paper, we propose a novel discriminative codebook learning method by introducing the subspace learning in codebook construction and leveraging its power to find a contextual local descriptor subspace to capture the discriminative information. The discriminative codebook construction and contextual subspace learning are formulated as an optimization problem and can be learned simultaneously. The effectiveness of the proposed method is evaluated through visual reranking experiments conducted on two real Web image search datasets.

1. Introduction

Given the explosive growth of Web images, image search plays an increasingly important role in our daily lives. Extensive research has been conducted to improve image search quality. Text-based image search leverages mature information retrieval techniques to index and search the images' associated textual information (filename, surrounding text, URL, *etc.*). Although text-based image search approaches are efficient for large-scale image indexing, they still have their own limitations since textual information cannot describe the rich content of images comprehensively and substantially. As a consequence, techniques with visual information involved are proposed to build content-based image retrieval prototypes [1–4] or enrich the textual descriptions via automatic image annotation/concept detection [5,6]. In all of these methods, visual representation of images plays the fundamental role. In recent years, bag-of-visual-word (BOVW) model has been widely used for image visual representation and has demonstrated promising performance in image retrieval [7–9] and image categorization [10–12]. In BOVW, a visual codebook needs to be constructed first by clustering a set of local features such as SIFT [13] extracted from a training image set. Then after quantizing all local descriptors into visual words in the codebook, each image can be represented as a histogram of number of visual words count.

☆ This work was supported in part by start-up funding from the University of Science and Technology of China to X. Tian, in part by Research Enhancement Program (REP), start-up funding from the Texas State University, DoD HBCU/MI grant W911NF-12-1-0057, and NSF CRI 1058724 to Y. Lu.

* Corresponding author. Tel.: +86 183 551 02690.
** Corresponding author. Tel.: +1 512 245 6580.
E-mail addresses: xinmei@ustc.edu.cn (X. Tian), yl12@txstate.edu (Y. Lu).

0165-1684/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.sigpro.2012.04.018

Obviously, in BOVW model, the quality of codebook directly affects the performance of image search. The most popular visual codebook generation method is *K*-means clustering [8,14]. It divides a large set of training SIFT feature points in the high dimensional feature space into clusters. Each cluster corresponds to a sub-space in the feature space, and the centroid of cluster is treated as a visual word. All visual words constitute a visual codebook. Then, given a novel feature point, feature quantization assigns it the visual word ID of its closest visual word in the space. As the size of image database becomes larger, a vocabulary tree method with hierarchical *K*-means [7] is more preferred for hierarchical clustering and efficient local feature quantization. Such kind of unsupervised clustering based codebook generalization method is easy for implementation and has been widely used in many applications. However, it totally ignores the known labeling information of training images. As a consequence, when the labeling information of the training images is given, the codebook generated via unsupervised clustering cannot embed the important image category information. In other words, the semantic contexts are lost.

To address this problem, some learning-based codebook construction methods [15–27] are proposed. These methods try to build supervised visual word codebooks in different ways: (1) refine/adapt original codebook based on image semantic labels; (2) build class specific vocabularies for image categorization; (3) learn discriminative and sparse coding models for object recognition; (4) generate supervised codebook by minimizing mutual information lost; (5) unify codebook construction with classifier training to build semantic vocabulary, *etc.* Although these approaches have improved traditional codebook, most of them construct visual codebook based on the raw local features. Very few attempts have been made to exploit manifold geometry of the local feature space. Actually, manifold learning has been proven an effective way to reveal the intrinsic structure of the original space and maximize the discriminative ability of data in the learned subspace. Wu et al. [21] proposed to construct semantic preserving codebook via distance metric learning. However, this method suffers the following two disadvantages: (1) additional region-level annotation labels are required in the distance metric learning stage. However, those region-level labels are usually unavailable. (2) The codebook construction and semantic distance metric learning are conducted in separate steps, which can hardly achieve a joint optimum. To tackle these problems, in this paper, we propose a novel supervised discriminative codebook learning method which has the following advantages:

(1) Our method introduces the subspace learning in codebook construction and leverages its power to find a contextual local descriptor subspace for embedding the discriminative information. In the expected subspace, images from different classes can be discriminated well.

(2) In our method, the codebook construction and contextual subspace learning are formulated as an optimization problem and they can be learned simultaneously. First, the closed-form expression of the bag-of-visual-word histogram based on the codebook in the desired new subspace is derived. Then, the distance between histograms of images from different classes is maximized, and the distance between histograms of images from the same class is minimized. This one-step optimization avoids the accumulation of errors introduced in each separated step.

(3) In our method, the discriminative ability of the codebook is measured at the image level, *i.e.*, directly requiring the histogram representation of images be similar or dissimilar. Compared with the local feature level discriminative ability pursuing methods [21], *e.g.*, distinguishing local features from different object parts, it is more reasonable since different objects may contain some common local patches.

The rest of this paper is organized as follows. Related work is summarized in Section 2. The proposed discriminative codebook learning method is described in Section 3. The proposed discriminative codebook is applied on Web image search reranking and the experimental results are given in Section 4. And the summary and conclusion are provided in Section 5.

2. Related work

Bag-of-visual-words (BOVW) model has been widely used in large-scale content-based image search applications. In general, BOVW model contains two major components: codebook generation and image representation.

2.1. Codebook generation

Codebook generation is to generate a set of visual words so as to make BOVW model possible to represent, index and retrieve images like documents. Sivic and Zisserman [8] and Csurka et al. [14] proposed to cluster the local features using *K*-means algorithm to construct codebooks. The centroid of each cluster is treated as a visual word. Nister and Stewenius [7] further proposed a vocabulary tree to hierarchically cluster and quantize local features efficiently for large scale image datasets.

Such kind of unsupervised clustering based codebook generalization method is easy for implementation, but totally ignores the known labeling information of training images. To address this problem, some learning-based codebook construction methods are proposed. Moosmann et al. [15] built supervised visual word codebook using randomized clustering forest. In this method, the image semantic labels were adopted as stopping test in tree building. Instead of using the trees for classification, each leaf was treated as a visual word. Jurie and Triggs [16] adopted mean-shift based approach for codebook generation to deal with clustering bias problem. Zhang et al. [23] and Liu et al. [19] refined the initial code words to build class-specific vocabularies for image categorization. Perronnin et al. [17] proposed universal codebook and class codebook to describe the content of all the considered classes of images and the adaptive class-specific content respectively. Each image was then represented by a set of histograms derived from both the universal codebook and class codebook. Mairal et al. [20] proposed to learn discriminative and sparse coding models for object categorization. This method required the label of each encoded vector. Lazebnik and Raginsky [22] generated supervised codebook by minimizing mutual information lost between features and labels during the quantization step. Yang et al. [27] proposed to unify the codebook construction with classifier training, and then encode images by a sequence of visual bits that constitute the semantic vocabulary. Wu et al. [21] proposed to construct semantic preserving codebook via distance metric learning. It first segmented the objects in images into different parts and the semantic labels of those parts were tagged by users. A distance metric was learned by minimizing/maximizing the distance between SIFT features extracted from the same/different semantic parts. Then a codebook was generated by clustering SIFT features with the learned distance metric.

2.2. Image representation

To represent images in BOVW model, local feature extraction is the first step, which extracts interest points in images by interest point detection. The detected interest points should have high repeatability over various changes. Difference of Gaussian (DoG) [13], MSER [28], and Hessian affine [29] are the three most popular detectors. After interest point detection, feature description is to generate a descriptor to describe the visual appearance of the local region centered at the interest point. The most popular local feature descriptor is SIFT feature [30], which is invariant to image rotation, scale and is also robust to affine distortion, addition of noise, and illumination changes.

Usually, several hundred or thousand local features can be extracted from a single image. With visual codebook defined, these high dimensional local features can be quantized to visual words in codebook by assigning a visual word ID to each feature. Then, a compact image representation can be achieved as a “bag” of visual words. The simplest feature quantization method is to find the closest (the most similar) visual word of a given feature by linear scanning all the visual words in the codebook . However, liner scanning is very time consuming especially in large-scale image applications. Therefore, an efficient approximate nearest neighbor search based on hierarchical vocabulary tree was proposed to propagate the query feature vector from the root node down the tree by comparing the corresponding child nodes and choosing the closest one [7]. In [31], a descriptor-dependent soft assignment scheme was also proposed to quantize a feature vector to a weighted combination of several visual words.

3. Supervised discriminative codebook learning

In this paper, we focus on codebook generation, which plays a key role of the BOVW model. A supervised discriminative codebook learning method is proposed. The framework of the proposed approach is illustrated in Fig. 1, which consists of three steps. In step 1, local descriptors/features from a set of training images in L different classes are extracted. The scale-invariant feature transform (SIFT) [13] local descriptor is adopted in this paper. In step 2, a conventional unsupervised codebook can be generated first by using the K-means clustering approach in the original raw SIFT feature space. However, this unsupervised codebook has limited discriminative ability since it does not embed the labeling information in the training set. In order to utilize the known label information to build a better codebook, in step 3, a contextual subspace with $\mathbf{U}$ as the projection matrix and a discriminative codebook $\mathbf{C}$ are learned to encode the discriminative information contained in the training image set. The objective subspace and codebook in the learned subspace should satisfy two basic expectations: (1) the bag-of-visual-word histograms of images within the same class are similar; (2) the bag-of-visual-word histograms of images from different classes are dissimilar.

Suppose the training image set consists of M images $I=\{I_1,\ldots,I_M\}$from L classes. Their corresponding class labels are $\mathbf{f}=[f_1,\ldots f_M]^{\mathbf{T}}$, where $f_i \in \{1,\ldots,L\}$, $i=1,\ldots,M$. The local features are first extracted from all images in I. The whole local feature set is represented as $\mathbf{X}=[\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N]\in \mathrm{R}^{D^*N}$, where $\mathbf{x}_i$ is a local descriptor, D is the dimension of the local feature, and N is the total number of local features. In this paper, the 128 dimensional SIFT local descriptor is adopted as the local feature, thus $D=128$ here.

Our aim is to learn a contextual subspace $\mathbf{U}$ and a codebook $\mathbf{C}$ in the subspace which has the maximum discriminative power of separating images from different classes and keeping images from the same class close to each other in this new subspace. In other words, we try to find an optimal subspace projection matrix $\mathbf{U}\in \mathrm{R}^{D^*d}$ which projects the $\mathbf{X}$ in the original feature space into $\mathbf{Y}$ (the new subspace), *i.e.*, $\mathbf{Y}=\mathbf{U}^{\mathbf{T}}\mathbf{X}$. Therefore, the whole feature set in the new space can be represent as $\mathbf{Y}=[\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_N]\in \mathrm{R}^{d^*N}$, where $\mathbf{y}_i$ is a local descriptor in the learned subspace. By quantizing each local feature in the new space into the nearest visual word in the discriminative codebook $\mathbf{C}$, each image can be represented as a bag-of-visual-word histogram in the new space.

Denoting the histogram representation of image I_i in the new subspace as $\mathbf{t}_i$, our objective function is

$$\max_{\mathbf{Y}} \sum_{(I_i,I_j)\in D} \|\mathbf{t}_i-\mathbf{t}_j\|^2-\alpha\sum_{(I_i,I_j)\in S} \|\mathbf{t}_i-\mathbf{t}_j\|^2 \tag{1}$$

where α is a trade-off parameter, $S=\{(I_i,I_j)|f_i=f_j\}$ is the similar image pairs set which consists of image pairs belonging to the same class, and $D=\{(I_i,I_j)|f_i\neq f_j\}$ is the dissimilar image pairs set which consists of image pairs belonging to different classes. Problem (1) requires that images in the learned optimal subspace, which are represented in bag-of-visual-word histogram based on the learned codebook, should be close to each other if they belong to the same class and should be far away from each other if they belong to different classes. With these constraints, discriminative ability of the codebook is guaranteed.

The challenging issue in problem (1) is how to derive $\mathbf{t}_i$ with both $\mathbf{C}$ and $\mathbf{U}$ unknown. The conventional hard quantization methods (*e.g.*, the nearest neighbor) do not work here, since solution of $\mathbf{t}_i$ is required to have a closed-form

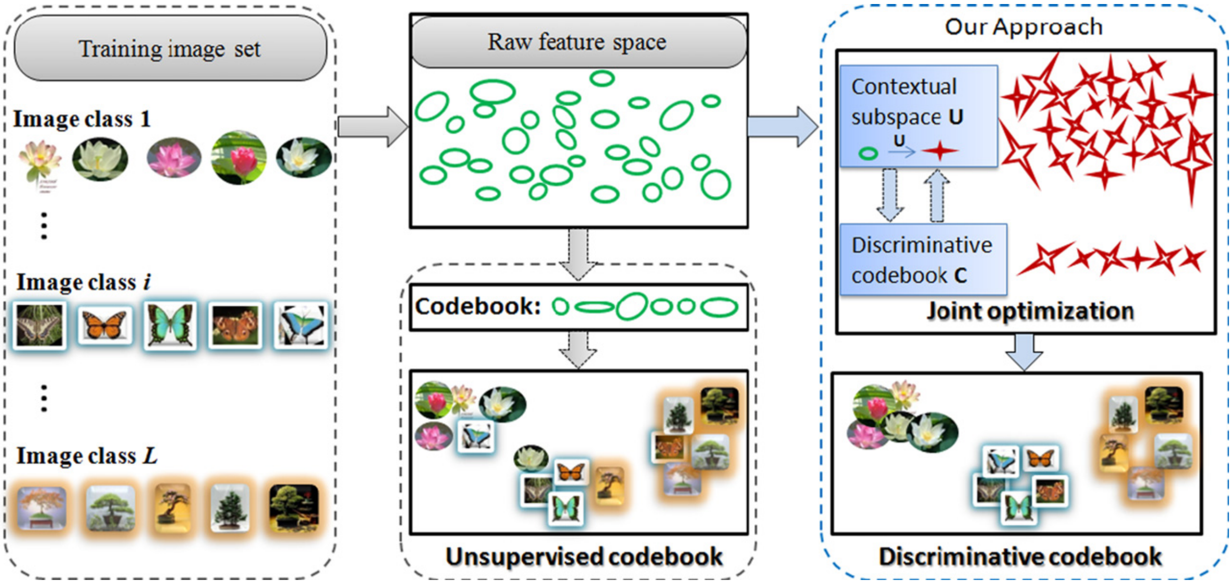

Fig. 1. Illustration of the classic unsupervised codebook and our proposed discriminative codebook learning framework.

expression regarding to $\mathbf{C}$ and $\mathbf{U}$ in problem (1). In order to solve this problem, we propose to use the soft quantization method which can express $\mathbf{t}_i$ with $\mathbf{C}$ and $\mathbf{U}$ effectively. Specifically, each local feature $\mathbf{y}\in R^{d^*1}$ in the subspace is assigned to the codebook $\mathbf{C}$ by solving the problem:

$$\min_{\boldsymbol{\beta}} \|\mathbf{y}-\mathbf{C}\boldsymbol{\beta}\|^2 \tag{2}$$

where $\boldsymbol{\beta}$ is the weighting coefficients. The solution to problem (2) with ridge regression [32] is $\boldsymbol{\beta}=(\mathbf{C}^\mathbf{T}\mathbf{C}+\lambda\mathbf{I})^{-1}\mathbf{C}^\mathbf{T}\boldsymbol{y}$, where λ is the coefficient to balance the capacity and complexity of the ridge regression model.

Then the bag-of-visual-word histogram $\mathbf{t}=[t_1,\cdots,t_H]^\mathbf{T}\in R^{H*1}$ of image I in the new subspace can be expressed as

$$\mathbf{t}=\sum\nolimits_{\mathbf{y}_i\in I}\boldsymbol{\beta}_i=\left(\mathbf{C}^\mathbf{T}\mathbf{C}+\lambda\mathbf{I}\right)^{-1}\mathbf{C}^\mathbf{T}(\mathbf{Ys}) \tag{3}$$

where $\mathbf{s}\in R^{N^*1}$ is a binary indicator vector with $\mathbf{s}_i=1$ if $\mathbf{x}_i\in I$, otherwise $\mathbf{s}_i=0$. H is the number of visual words in $\mathbf{C}$.

Based on (3), the difference between two histograms $\mathbf{t}_i$ and $\mathbf{t}_j$, which correspond to images I_i and I_j respectively, can be calculated as

$$\begin{aligned}\mathbf{t}_i-\mathbf{t}_j &= \left(\mathbf{C}^\mathbf{T}\mathbf{C}+\lambda\mathbf{I}\right)^{-1}\mathbf{C}^\mathbf{T}(\mathbf{Ys}_i)-\left(\mathbf{C}^\mathbf{T}\mathbf{C}+\lambda\mathbf{I}\right)^{-1}\mathbf{C}^\mathbf{T}(\mathbf{Ys}_j)\\ &=\left(\mathbf{C}^\mathbf{T}\mathbf{C}+\lambda\mathbf{I}\right)^{-1}\mathbf{C}^\mathbf{T}\mathbf{Y}\left(\mathbf{s}_i-\mathbf{s}_j\right)\end{aligned} \tag{4}$$

Substituting (4) into objective function (1), we obtain:

$$\begin{aligned}&\max_{\mathbf{Y}}\sum_{(i,j)\in D}\|\mathbf{t}_i-\mathbf{t}_j\|^2-\alpha\sum_{(i,j)\in S}\|\mathbf{t}_i-\mathbf{t}_j\|^2\\ &=\max_{\mathbf{Y}}\sum_{(i,j)\in D}\mathrm{tr}((\mathbf{t}_i-\mathbf{t}_j)(\mathbf{t}_i-\mathbf{t}_j)^\mathbf{T})-\alpha\sum_{(i,j)\in S}\mathrm{tr}((\mathbf{t}_i-\mathbf{t}_j)(\mathbf{t}_i-\mathbf{t}_j)^\mathbf{T})\\ &=\max_{\mathbf{Y}}\sum_{(i,j)\in D}\mathrm{tr}((\mathbf{C}^\mathbf{T}\mathbf{C}+\lambda\mathbf{I})^{-1}\mathbf{C}^\mathbf{T}\mathbf{Y}(\mathbf{s}_i-\mathbf{s}_j)((\mathbf{C}^\mathbf{T}\mathbf{C}+\lambda\mathbf{I})^{-1}\\ &\quad\times\mathbf{C}^\mathbf{T}\mathbf{Y}(\mathbf{s}_i-\mathbf{s}_j))^\mathbf{T})-\alpha\sum_{(i,j)\in S}\mathrm{tr}((\mathbf{C}^\mathbf{T}\mathbf{C}+\lambda\mathbf{I})^{-1}\mathbf{C}^\mathbf{T}\mathbf{Y}(\mathbf{s}_i-\mathbf{s}_j)\\ &\quad\times((\mathbf{C}^\mathbf{T}\mathbf{C}+\lambda\mathbf{I})^{-1}\mathbf{C}^\mathbf{T}\mathbf{Y}(\mathbf{s}_i-\mathbf{s}_j))^\mathbf{T})\\ &=\max_{\mathbf{Y}}\sum_{(i,j)\in D}\mathrm{tr}((\mathbf{C}^\mathbf{T}\mathbf{C}+\lambda\mathbf{I})^{-1}\mathbf{C}^\mathbf{T}\mathbf{Y}(\mathbf{s}_i-\mathbf{s}_j)(\mathbf{s}_i-\mathbf{s}_j)^\mathbf{T}\\ &\quad\times\mathbf{Y}^\mathbf{T}\mathbf{C}((\mathbf{C}^\mathbf{T}\mathbf{C}+\lambda\mathbf{I})^{-1})^\mathbf{T})-\alpha\sum_{(i,j)\in S}\mathrm{tr}((\mathbf{C}^\mathbf{T}\mathbf{C}+\lambda\mathbf{I})^{-1}\\ &\quad\times\mathbf{C}^\mathbf{T}\mathbf{Y}(\mathbf{s}_i-\mathbf{s}_j)(\mathbf{s}_i-\mathbf{s}_j)^\mathbf{T}\mathbf{Y}^\mathbf{T}\mathbf{C}((\mathbf{C}^\mathbf{T}\mathbf{C}+\lambda\mathbf{I})^{-1})^\mathbf{T})\\ &=\max_{\mathbf{Y}}\mathrm{tr}\left((\mathbf{C}^\mathbf{T}\mathbf{C}+\lambda\mathbf{I})^{-1}\mathbf{C}^\mathbf{T}\mathbf{Y}\left(\sum_{(i,j)\in D}(\mathbf{s}_i-\mathbf{s}_j)(\mathbf{s}_i-\mathbf{s}_j)^\mathbf{T}\text{-}\alpha\right.\right.\\ &\quad\left.\left.\times\sum_{(i,j)\in S}(\mathbf{s}_i-\mathbf{s}_j)(\mathbf{s}_i-\mathbf{s}_j)^\mathbf{T}\right)\mathbf{Y}^\mathbf{T}\mathbf{C}((\mathbf{C}^\mathbf{T}\mathbf{C}+\lambda\mathbf{I})^{-1})^\mathbf{T}\right)\\ &=\max_{\mathbf{Y}}\mathrm{tr}((\mathbf{C}^\mathbf{T}\mathbf{C}+\lambda\mathbf{I})^{-1}\mathbf{C}^\mathbf{T}\mathbf{YLY}^\mathbf{T}\mathbf{C}((\mathbf{C}^\mathbf{T}\mathbf{C}+\lambda\mathbf{I})^{-1})^\mathbf{T})\end{aligned} \tag{5}$$

where $\mathbf{L}=\sum_{(i,j)\in D}(\mathbf{s}_i-\mathbf{s}_j)(\mathbf{s}_i-\mathbf{s}_j)^\mathbf{T}\text{-}\alpha\sum_{(i,j)\in S}(\mathbf{s}_i-\mathbf{s}_j)(\mathbf{s}_i-\mathbf{s}_j)^\mathbf{T}$ and $\mathbf{Y}=\mathbf{U}^\mathbf{T}\mathbf{X}$.

The codebook $\mathbf{C}$ is approximated by the projection of an initial codebook $\tilde{\mathbf{C}}\in R^{D*H}$ obtained in original space, *i.e.*, $\mathbf{C}=\mathbf{U}^\mathbf{T}\tilde{\mathbf{C}}$. By imposing $\mathbf{U}^\mathbf{T}\mathbf{U}=\mathbf{I}$, problem (5) can be rewritten as

$$\begin{aligned}&\max_{\mathbf{Y}}\mathrm{tr}((\mathbf{C}^\mathbf{T}\mathbf{C}+\lambda\mathbf{I})^{-1}\mathbf{C}^\mathbf{T}\mathbf{YLY}^\mathbf{T}\mathbf{C}((\mathbf{C}^\mathbf{T}\mathbf{C}+\lambda\mathbf{I})^{-1})^\mathbf{T})\\ &=\max_{\mathbf{U}}\mathrm{tr}((\tilde{\mathbf{C}}^\mathbf{T}\mathbf{UU}^\mathbf{T}\tilde{\mathbf{C}}+\lambda\mathbf{I})^{-1}\tilde{\mathbf{C}}^\mathbf{T}\mathbf{UU}^\mathbf{T}\mathbf{XLX}^\mathbf{T}\mathbf{UU}^\mathbf{T}\tilde{\mathbf{C}}\\ &\quad\times[(\tilde{\mathbf{C}}^\mathbf{T}\mathbf{UU}^\mathbf{T}\tilde{\mathbf{C}}+\lambda\mathbf{I})^{-1}]^\mathbf{T})\end{aligned} \tag{6}$$

Since $(\mathbf{I}+\mathbf{AB})^{-1}\mathbf{A}=\mathbf{A}(\mathbf{I}+\mathbf{BA})^{-1}$, we obtain that

$$\begin{aligned}(\tilde{\mathbf{C}}^\mathbf{T}\mathbf{UU}^\mathbf{T}\tilde{\mathbf{C}}+\lambda\mathbf{I})^{-1}\tilde{\mathbf{C}}^\mathbf{T}\mathbf{U}&=\tilde{\mathbf{C}}^\mathbf{T}\mathbf{U}(\mathbf{U}^\mathbf{T}\tilde{\mathbf{C}}\tilde{\mathbf{C}}^\mathbf{T}\mathbf{U}+\lambda\mathbf{I})^{-1}\\ &=\tilde{\mathbf{C}}^\mathbf{T}\mathbf{U}(\mathbf{U}^\mathbf{T}\tilde{\mathbf{C}}\tilde{\mathbf{C}}^\mathbf{T}\mathbf{U}+\lambda\mathbf{U}^\mathbf{T}\mathbf{U})^{-1}\\ &=\tilde{\mathbf{C}}^\mathbf{T}\mathbf{U}(\mathbf{U}^\mathbf{T}(\tilde{\mathbf{C}}\tilde{\mathbf{C}}^\mathbf{T}+\lambda\mathbf{I})\mathbf{U})^{-1}\\ &=\tilde{\mathbf{C}}^\mathbf{T}\mathbf{UU}^\mathbf{T}(\tilde{\mathbf{C}}\tilde{\mathbf{C}}^\mathbf{T}+\lambda\mathbf{I})^{-1}\mathbf{U}\end{aligned} \tag{7}$$

Substituting (7) into problem (6), we obtain:

$$\begin{aligned}&\max_{\mathbf{Y}}\mathrm{tr}((\mathbf{C}^\mathbf{T}\mathbf{C}+\lambda\mathbf{I})^{-1}\mathbf{C}^\mathbf{T}\mathbf{YLY}^\mathbf{T}\mathbf{C}((\mathbf{C}^\mathbf{T}\mathbf{C}+\lambda\mathbf{I})^{-1})^\mathbf{T})\\ &=\max_{\mathbf{U}}\mathrm{tr}(\tilde{\mathbf{C}}^\mathbf{T}\mathbf{UU}^\mathbf{T}(\tilde{\mathbf{C}}\tilde{\mathbf{C}}^\mathbf{T}+\lambda\mathbf{I})^{-1}\mathbf{UU}^\mathbf{T}\mathbf{XLX}^\mathbf{T}\mathbf{UU}^\mathbf{T}\end{aligned}$$

$$\times[(\tilde{\mathbf{C}}\tilde{\mathbf{C}}^{\mathbf{T}}+\lambda\mathbf{I})^{-1}]^{\mathbf{T}}\mathbf{U}\mathbf{U}^{\mathbf{T}}\tilde{\mathbf{C}})$$

$$=\max_{\mathbf{U}}\ \mathrm{tr}(\tilde{\mathbf{C}}^{\mathbf{T}}\mathbf{U}\mathbf{U}^{\mathbf{T}}\mathbf{G}\mathbf{U}\mathbf{U}^{\mathbf{T}}\mathbf{X}\mathbf{L}\mathbf{X}^{\mathbf{T}}\mathbf{U}\mathbf{U}^{\mathbf{T}}\mathbf{G}^{\mathbf{T}}\mathbf{U}\mathbf{U}^{\mathbf{T}}\tilde{\mathbf{C}}) \qquad (8)$$

where $\mathbf{G}=\left(\tilde{\mathbf{C}}\tilde{\mathbf{C}}^{\mathbf{T}}+\lambda\mathbf{I}\right)^{-1}$.

Problem (8) can be solved via gradient descent algorithm as given in Algorithm 1 with

$$\Delta(\mathbf{U})=\frac{\partial\mathrm{tr}(\tilde{\mathbf{C}}^{\mathbf{T}}\mathbf{U}\mathbf{U}^{\mathbf{T}}\mathbf{G}\mathbf{U}\mathbf{U}^{\mathbf{T}}\mathbf{X}\mathbf{L}\mathbf{X}^{\mathbf{T}}\mathbf{U}\mathbf{U}^{\mathbf{T}}\mathbf{G}^{\mathbf{T}}\mathbf{U}\mathbf{U}^{\mathbf{T}}\tilde{\mathbf{C}})}{\partial\mathbf{U}}$$
$$=2\tilde{\mathbf{C}}\tilde{\mathbf{C}}^{\mathbf{T}}\mathbf{U}\mathbf{U}^{\mathbf{T}}\mathbf{G}\mathbf{U}\mathbf{U}^{\mathbf{T}}\mathbf{X}\mathbf{L}\mathbf{X}^{\mathbf{T}}\mathbf{U}\mathbf{U}^{\mathbf{T}}\mathbf{G}^{\mathbf{T}}\mathbf{U}$$
$$+2\mathbf{G}\mathbf{U}\mathbf{U}^{\mathbf{T}}\mathbf{X}\mathbf{L}\mathbf{X}^{\mathbf{T}}\mathbf{U}\mathbf{U}^{\mathbf{T}}\mathbf{G}^{\mathbf{T}}\mathbf{U}\mathbf{U}^{\mathbf{T}}\tilde{\mathbf{C}}\tilde{\mathbf{C}}^{\mathbf{T}}\mathbf{U}$$
$$+2\mathbf{G}^{\mathbf{T}}\mathbf{U}\mathbf{U}^{\mathbf{T}}\tilde{\mathbf{C}}\tilde{\mathbf{C}}^{\mathbf{T}}\mathbf{U}\mathbf{U}^{\mathbf{T}}\mathbf{G}\mathbf{U}\mathbf{U}^{\mathbf{T}}\mathbf{X}\mathbf{L}\mathbf{X}^{\mathbf{T}}\mathbf{U}$$
$$+2\mathbf{X}\mathbf{L}\mathbf{X}^{\mathbf{T}}\mathbf{U}\mathbf{U}^{\mathbf{T}}\mathbf{G}^{\mathbf{T}}\mathbf{U}\mathbf{U}^{\mathbf{T}}\tilde{\mathbf{C}}\tilde{\mathbf{C}}^{\mathbf{T}}\mathbf{U}\mathbf{U}^{\mathbf{T}}\mathbf{G}\mathbf{U} \qquad (9)$$

In (8), the $\mathbf{C}$ is approximated by the projection of the initial codebook obtained in the original space by $\mathbf{C}=\mathbf{U}^{\mathbf{T}}\tilde{\mathbf{C}}$. Based on the learned space $\mathbf{U}$, the updated codebook $\tilde{\mathbf{C}}$ can be derived by clustering local features $\mathbf{Y}$ in the new space. With the updated $\tilde{\mathbf{C}}$, a new subspace projection matrix $\mathbf{U}$ can be obtained again by solving problem (8). After repeating this iterative learning for several times, the optimal $\mathbf{U}$ and $\mathbf{C}$ will be learned.

Algorithm 1.

```
//Gradient descent for solving problem (8)
Input: $\mathbf{X}$, $\mathbf{L}$, $\tilde{\mathbf{C}}$
Initialization:
$\mathbf{U}$=random($D$,$d$)//random initialization
$\mathbf{U}$←orthogonalize($\mathbf{U}$)//orthogonalization
For $i=1:T$
$\mathbf{U}\leftarrow\mathbf{U}+\eta\Delta(\mathbf{U})$
$\mathbf{U}$←orthogonalize($\mathbf{U}$)
End for
Return $\mathbf{U}$
```

3.1. Complexity analysis

Compared with the unsupervised *K*-means codebook construction method, the extra time cost introduced in our method is spent on solving problem (8). The complexity analysis of learning problem (8) consists of three parts. The first part is the time complexity of calculating the $\mathbf{XLX^T}$ which is $O(D^2N_{\mathrm{ImgPair}}+DN_{\mathrm{Img}}N_{\mathrm{AvgSift}})$, where $N_{\mathrm{ImgPair}}=|S|+|D|$ is the total number of image pairs in similar and dissimilar sets, N_{Img} is the number of training images, and N_{AvgSift} is the average number of sift features in each training image. This part only needs to be calculated once. The second part is the time complexity of calculating $\mathbf{G}$ which is $O(D^3)$. The third part is the time complexity of gradient descent algorithm, which is $O(TD^2(H+d+D))$, where T is the iteration number in Algorithm 1 and H is the size of codebook. Since both d and D are much smaller than H, the time cost in this part can be approximated by $O(TD^2H)$. Taking all the three parts into account, the total complexity is $O(D^2TH+D^2N_{\mathrm{ImgPair}}+DN_{\mathrm{Img}}N_{\mathrm{AvgSift}})$. In practical experiments, we find that the Algorithm 1 usually converges soon after few hundreds of iterations, *i.e.*, T is about 500. N_{ImgPair} is usually smaller than TH and $N_{\mathrm{Img}}N_{\mathrm{AvgSift}}$ is smaller than DTH. Therefore the total complexity can be approximated by $O(D^2TH)$. For the unsupervised *K*-means codebook generation, its complexity is $O(T_1DHN_{Sift})$ where T_1 is the number of iterations in *K*-means and N_{Sift} is the total number of sift features used for codebook construction. To ensure the capacity of the codebook, N_{Sift} is usually larger than 10,000. We can see that, the extra time cost introduced in our proposed method is comparable to that of *K*-means. Besides, the codebook learning process can be done offline usually. Therefore, it is worth to learn more powerful codebook with moderate extra computational cost introduced.

3.2. Discussion

In the above sections, we illustrate our method with training images from L different classes. Actually our method can be extended to more general cases. According to Eq. (1), all we need is the semantic similarity information between images, *i.e.* whether two images are similar (belonging to S) or dissimilar (belonging to D). The exact class labels of these images are not essentially required. Compared to other supervised codebook construction methods which require class labels [15,17], our approach has the following two advantages: (1) our approach does not rely on exact class label information. It can be flexibly applied to many real applications, which do not provide image labels, but have the side information of similar and dissimilar image pair constraints available, such as image reranking; (2) in many applications, there may exist a special class, which consists of images not belonging to any pre-defined class. For example, in the image retrieval, a negative class exists and contains all the irrelevant images to a given query. Traditional supervised codebook construction methods either ignore the images in the special class since they do not belong to any pre-defined class, or require them to be clustered together. However, all the images in the special class are dissimilar in their own way. Therefore, it is unreasonable to force them to be close to each other. Our method can effectively deal with this problem by only requiring images from this special class be far away from the images in other classes, without pushing images within the special class to be gathered together.

4. Experiments

In this paper, we evaluate the proposed discriminative codebook learning method on Web image search reranking. The major purpose is that if the learned codebook can well capture discriminative information, more related images should be re-ranked to the top and irrelevant images should be ranked to the bottom. Other two state-of-the-art codebook generation methods are selected for comparison, including one unsupervised codebook learning method (KM) and one supervised codebook learning method (ERCF). These methods are tested on two real Web image search datasets: Web29 and MSRA-MM, and their performance on image reranking are compared.

4.1. Experimental dataset and setting

Web29 dataset contains 25,890 Web images collected from Microsoft Bing image search engine based 29 popular queries. These 29 queries were selected from a commercial image search engine query log and popular tags from Flickr. These queries cover a vast range of topics, including scene ("sky", "winter"), objects ("funny dog", "grape"), named person ("George W. Bush"), *etc*. We submitted each query to Bing, and collected at most top-1000 images returned. Some example images in this dataset are shown in Fig. 2. For each query, the relevance labels of returned images are evaluated on two levels: "relevant" or "irrelevant". In this dataset, there are 49.80% images labeled as relevant.

MSRA-MM is a public available dataset released by Microsoft Research Asia [33]. It consists of 68 representative queries which are selected based on the query log of Microsoft Bing search. For each query, about 1000 images are collected from, resulting 65,443 images in total. For each image, its relevance to the corresponding query is labeled with three levels: "very relevant", "relevant", and "irrelevant". In this paper, we do not distinguish "very relevant" and "relevant", and treat both as relevant ones. Example images in this dataset are shown in Fig. 3.

To learn the discriminative codebook, labeled images are required to form the similar image pairs set S and dissimilar image pairs set D in Eq. (1). However, in real image search applications, the image labels are unavailable. To solve this problem, we apply the popular pseudo relevance feedback assumption to get pseudo-relevant and pseudo-irrelevant images to construct the training image set. Specifically, for each query, we can get the initial ranks of the images from the search engine. Then, with the pseudo relevance feedback assumption, the top and bottom ranked images are treated as pseudo-relevant and pseudo-irrelevant images respectively. In our experiments, the top 20% returned images are pseudo-relevant images and the bottom 20% returned images are pseudo-irrelevant ones. With these label information, the discriminative codebook can be generated according to the approach presented in Section 3 and the dimension of the new space is set as 50 empirically.

The pseudo relevance feedback (PRF) reranking [34] method is adopted to reorder the images for each query. Specifically, it first trains a classification model with the aforementioned pseudo-relevant and irrelevant images. Then, all images are re-ordered according to the relevance score predicted by the trained classifier. We compare the reranking results performed on the images represented by our proposed discriminative codebook (denoted **DC**) and two other baseline codebooks. One is an unsupervised codebook generated via *K*-means clustering algorithm (denoted **KM**) and the other is a supervised codebook generated via extremely randomized clustering forests [15] (denoted **ERCF**).

For the ranking performance measurement, the non-interpolated average precision (AP) [35], which is widely used in information retrieval, is adopted. The AP averages the precision values obtained when each relevant image occurs. The AP of top-T ranked images AP@T is calculated as

$$\text{AP@}T = \frac{1}{Z_T}\sum_{i=1}^{T}[precison(i) \times rel(i)] \tag{10}$$

where $rel(i)$ is the binary function on the relevance of the i-th ranked image with "1" for relevant and "0" for irrelevant. The Z_T is a normalization constant that is chosen to guarantee that AP@T=1 for the perfect ranking result. The $precison(i)$ is the precision of top-i ranked images:

$$precison(i) = \frac{1}{i}\sum_{j=1}^{i} rel(j) \tag{11}$$

4.2. Experimental results

Figs. 4 and 5 show the reranking performance on datasets Web29 and MSRA-MM respectively, with images represented by our proposed **DC** codebook, conventional unsupervised KM codebook, and supervised ERCF codebook. The MAP, average performance of AP over all queries in the dataset, is reported. The text-based search

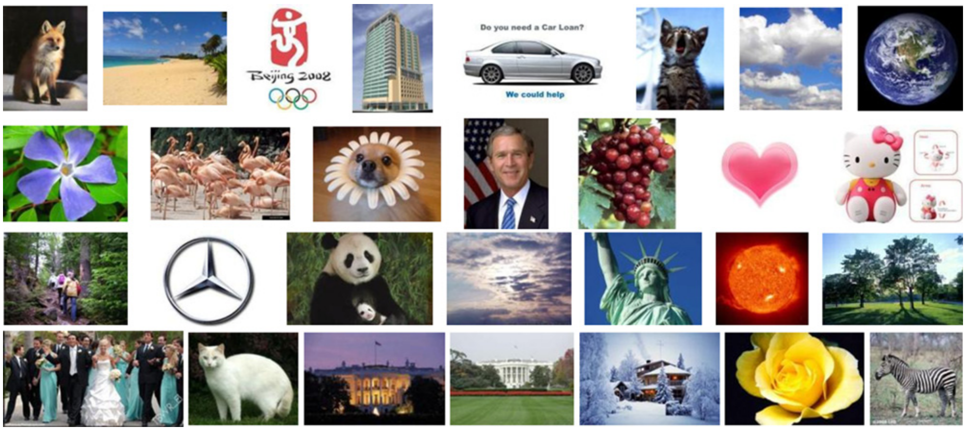

Fig. 2. Example images in the Web29 dataset.

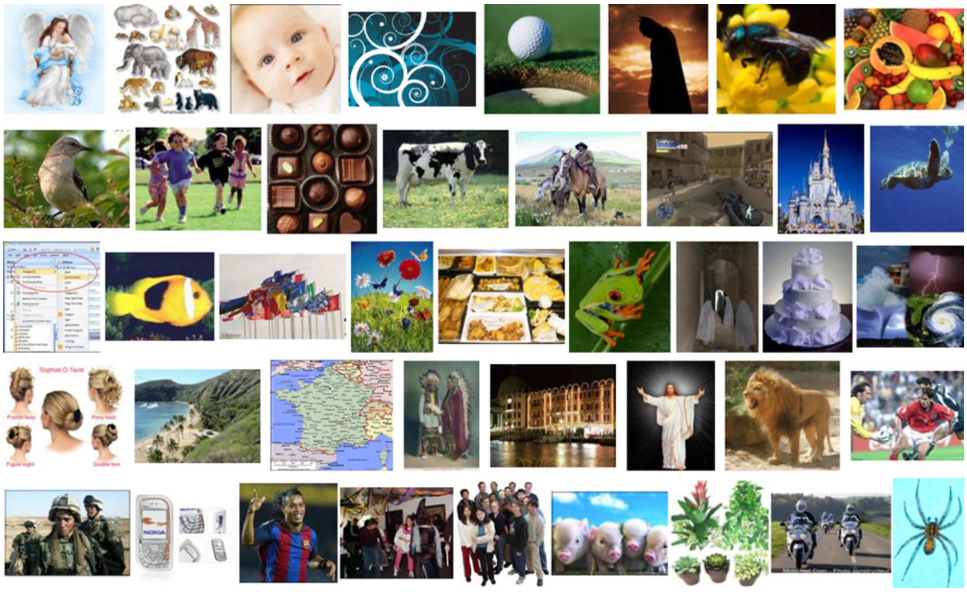

Fig. 3. Example images in the MSRA-MM dataset.

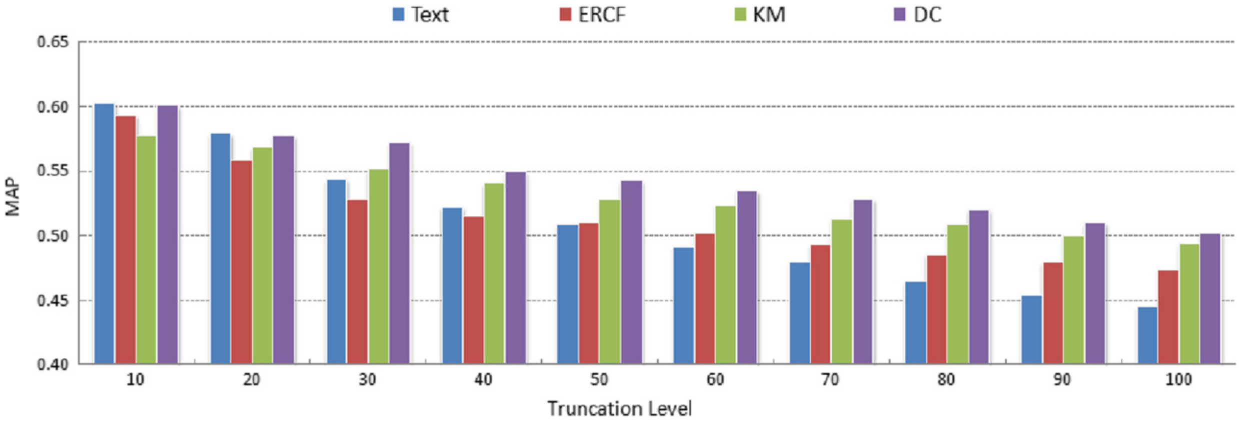

Fig. 4. Performance comparison of the text-based search baseline (Text), reranking using BOVW features generated with conventional supervised and unsupervised codebook (ERCF, KM) and the proposed discriminative codebook (DC) on Web29.

baseline (Text) of the search engine is also given for reference. Both Figs. 4 and 5 show that the reranking with our proposed DC codebook achieves better performance than the classical unsupervised and supervised codebooks consistently over various truncation levels. Taking truncation level 60 as example, on Web29, ERCF and KM improve the text-based search baseline (Text) from 0.49 to 0.50 and 0.52 respectively, and our method DC further improves it to 0.54. This phenomenon demonstrates that our proposed discriminative codebook not only outperforms unsupervised codebook, but also shows superiority to conventional supervised one.

The supervised ERCF does not perform very well here and sometimes it is even worse than unsupervised KM. The possible reason is that in the Web image search reranking problem, the pseudo class labels of relevant and irrelevant images are very noisy. ERCF requires all images from the same class should be alike, including images from the irrelevant class. However, in this application, the images in the irrelevant class have high visual appearance variance. Each of these images is irrelevant in its own way. That is why ERCF may fail. Instead, our DC does not have such a limitation. It can flexibly require all relevant images be similar, and relevant images and irrelevant images be dissimilar without knowing their exact label information. Therefore, our method has better capacity and achieves better performance.

We further analyze the sensitivity of important tradeoff parameter in our method, *i.e.*, the α in Eq. (1). This parameter has two major influences. A smaller α reflects the importance of separating relevant images from irrelevant ones. A larger α denotes that more attention is given to keep relevant images close in the new space. Fig. 6 shows the performance of DC with various αs' on Web29 dataset. The performance of Text (the text-based search baseline) and KM are also given for comparison. From Fig. 6, we find the following observations: (1) when α is large, *e.g.*, more than 0.5, the performance is

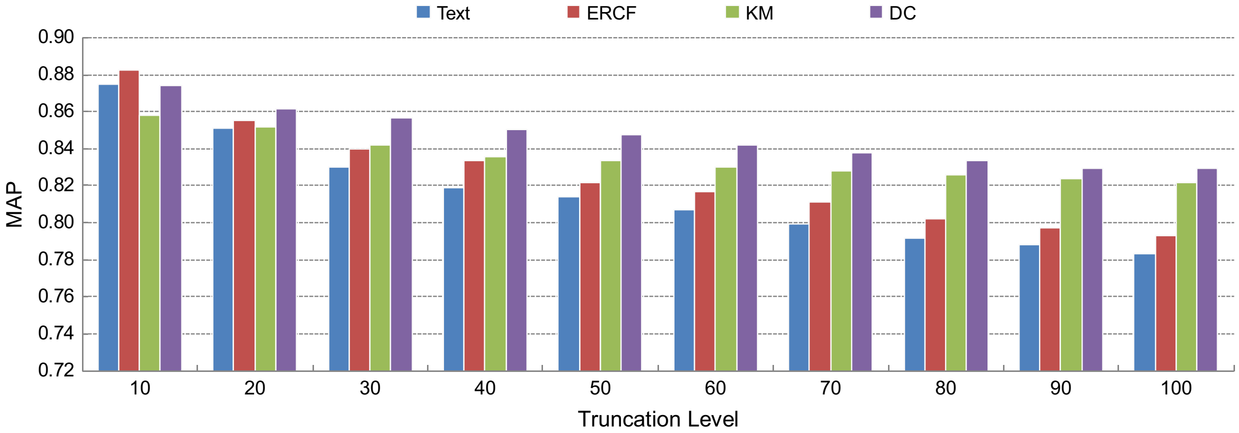

Fig. 5. Performance comparison of the text-based search baseline (Text), reranking using BOVW features generated with conventional supervised and unsupervised codebook (ERCF, KM) and the proposed discriminative codebook (DC) on MSRA-MM.

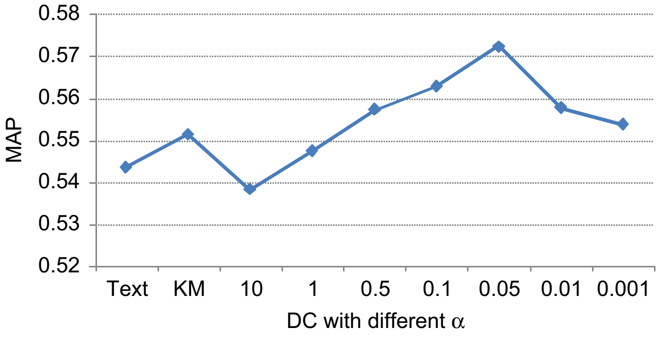

Fig. 6. Performance comparison of text-based search baseline (Text), KM and DC with different α on Web29.

unsatisfactory and even worse than KM. The major reason is that in this situation, the similar image pairs information is mainly preserved while important dissimilar image pairs (discriminative) information is less considered. This phenomenon reveals the importance of the discriminative information in codebook learning. (2) The performance of DC increases when α decreases, and reaches the optimal value at about $\alpha = 0.05$. However, when α becomes smaller than 0.05, the MAP decreases. The major reason is that in this case the discriminative information of dissimilar image pairs plays the dominant role and the similar image pairs information is ignored. It reveals that both the similar and dissimilar image pairs information reflect the discriminative information from different aspects complimentarily. A suitable combination of them is essential to achieve a good performance.

5. Conclusion

In this paper, we propose a novel supervised discriminative codebook learning method, which not only finds a contextual subspace to embed the discriminative information, but also learns the contextual subspace and discriminative codebook simultaneously. In the learned new space, images from different classes can be well separated and images from the same class are close to each other. We apply the proposed method on Web image search reranking problem and the experimental results on two real Web image search datasets have demonstrated the effectiveness of our approach and its superiority than other state-of-the-art codebook learning methods. Future work will be devoted to adapting the proposed framework to more sophisticated quantization and manifold learning methods. We are also investigating application of the proposed method to other applications such as image classification and concept detection.

References

[1] Y. Yang, Y.-T. Zhuang, F. Wu, Y.-H. Pan, Harmonizing hierarchical manifolds for multimedia document semantics understanding and cross-media retrieval, IEEE Transactions on Multimedia 10 (3) (2008) 437–446.

[2] Y. Yang, F. Nie, D. Xu, J. Luo, Y.-T. Zhuang, Y.-H. Pan., A multimedia retrieval framework based on semi-supervised ranking and relevance feedback, IEEE Transactions on Pattern Analysis and Machine Intelligence 34 (4) (2012) 723–742.

[3] L. Duan, W. Gao, W. Zeng, D. Zhao, Adaptive relevance feedback based on Bayesian inference for image retrieval, Signal Processing 85 (2) (2005) 395–399.

[4] M. Wang, K. Yang, X.-S. Hua, H.-J. Zhang, Towards a relevant and diverse search of social images, IEEE Transactions on Multimedia 2 (8) (2010) 829–842.

[5] Y. Yang, F. Wu, F. Nie, H.T. Shen, Y.-T. Zhuang, A.G. Hauptmann, Web and personal image annotation by mining label correlation

with relaxed visual graph embedding, IEEE Transactions on Image Processing 21 (3) (2012) 1339–1351.

[6] J. Tang, G.-J. Qi, M. Wang, X.-S. Hua, Video semantic analysis based on structure-sensitive anisotropic manifold ranking, Signal Processing 89 (12) (2009) 2313–2323.

[7] D. Nister, H. Stewenius, Scalable recognition with a vocabulary tree, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2006, pp. 2161–2168.

[8] J. Sivic, A. Zisserman., Video Google: a text retrieval approach to object matching in videos, in: Proceedings of the IEEE International Conference on Computer Vision, 2003, pp. 1470–1477.

[9] Z. Wu, Q.F. Ke, J. Sun, Bundling features for large-scale partial-duplicate web image search, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 25–32.

[10] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: spatial pyramid matching for recognizing natural scene categories, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2006, pp. 2169–2178.

[11] C. Wang, D. Blei, L. Fei-Fei, Simultaneous image classification and annotation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 1903–1910.

[12] Z. Si, H. Gong, Y.N. Wu, S.C. Zhu, Learning mixed templates for object recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 272–279.

[13] D.G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2) (2004) 91–110.

[14] G. Csurka, C. Dance, L. Fan, J. Willamowski, C. Bray, Visual categorization with bags of keypoints, in: Proceedings of the European Conference on Computer Vision Workshop on on Statistical Learning in Computer Vision, 2004, pp. 1–22.

[15] F. Moosmann, B. Triggs, F. Jurie, Randomized clustering forests for building fast and discriminative visual vocabularies, Advances in Neural Information Processing Systems (2007).

[16] F. Jurie, B. Triggs, Creating efficient codebooks for visual recognition, in: Proceedings of the International Conference on Computer Vision, 2005, pp. 604–610.

[17] F. Perronnin, C. Dance, G. Csurka, M. Bressan, Adapted vocabularies for generic visual categorization, in: Proceedings of the European Conference on Computer Vision, 2006, pp. 464–475.

[18] F. Perronnin, C.R. Dance, Fisher kernels on visual vocabularies for image categorization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2007.

[19] J. Liu, Y. Yang, M. Shah, Learning semantic visual vocabularies using diffusion distance, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 461–468.

[20] J. Mairal, F. Bach, J. Ponce, G. Sapiro, A. Zisserman, Supervised dictionary learning, Advances in Neural Information Processing Systems (2008) 1033–1040.

[21] L. Wu, S. Hoi, N. Yu, Semantic-preserving bag-of-words models for efficient image annotation, in: Proceedings of the ACM Workshop on Large-Scale Multimedia Retrieval and Mining, 2009, pp. 19–26.

[22] S. Lazebnik, M. Raginsky, Supervised learning of quantizer codebook by information loss minimization, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (7) (2009) 1294–1309.

[23] J. Zhang, M. Marszalek, S. Lazebnik, C. Schmid, Local features and kernels for classification of texture and object categories: a comprehensive review, International Journal of Computer Vision (2007) 213–238.

[24] L. Wang, Toward a discriminative codebook: codeword selection across multi-resolution, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2007.

[25] M. Marszalek, C. Schmid, Semantic hierarchies for visual object recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.

[26] X. Lian, Z. Li, C. Wang, L. Zhang, Probabilistic models for supervised dictionary learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 2305–2312.

[27] L. Yang, R. Jin, R. Sukthankar, F. Jurie, Unifying discriminative visual codebook generation with classifier training for object category reorganization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 461–468.

[28] J. Matas, O. Chum, U. Martin, T. Pajdla, Robust wide baseline stereo from maximally stable extremal regions, in: Proceedings of the British Machine Vision Conference, 2002.

[29] K. Mikolajczyk, C. Schmid, Scale & affine invariant interest point detectors, International Journal of Computer Vision 1 (60) (2004) 63–86.

[30] M. Calonder, V. Lepetit, C. Strecha, P. Fua, BRIEF: binary robust independent elementary features, in: Proceedings of the European Conference on Computer Vision, 2010.

[31] J. Philbin, O. Chum, M. Isard, J. Sivic, A. Zisserman, Lost in quantization: Improving particular object retrieval in large scale image databases, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2008.

[32] A.E. Hoerl, R.W. Kennard, Ridge regression: biased estimation for nonorthogonal problems, Technometrics 42 (1) (1970).

[33] M. Wang, L. Yang, X.-S. Hua., MSRA-MM: Bridging Research and Industrial Societies for Multimedia Information Retrieval. no. MSR-TR-2009-30, 16 March 2009.

[34] R. Yan, A.G. Hauptmann, R. Jin, Multimedia search with pseudo relevance feedback, in: Proceedings of the ACM International Conference on Content-based Image and Video Retrieval, 2003, pp. 238–247.

[35] Trecvid video retrieval evaluation. 〈http://wwwnlpir.nist.gov/projects/trecvid/〉.