Geometric Optimum Experimental Design for Collaborative Image Retrieval

Lining Zhang, *Student Member, IEEE,* Lipo Wang, *Senior Member, IEEE,* and Weisi Lin, *Senior Member, IEEE*

Abstract

Relevance feedback (RF) schemes have been widely designed to improve the performance of content-based image retrieval (CBIR). Despite the success, it is not appropriate to require the user to label a large number of samples in RF. Collaborative image retrieval (CIR) aims to reduce the labeling efforts of the user by resorting to the auxiliary information. Support vector machine (SVM) active learning can select ambiguous samples as the most informative ones for the user to label with the help of the optimal hyperplane of SVM, and thus alleviate the labeling efforts of conventional RF. However, the optimal hyperplane of SVM is usually unstable and inaccurate with small-sized training data, and this is always the case in image retrieval since the user would not like to label a large number of feedback samples and cannot label each sample accurately all the time. In this paper, we propose a novel active learning method, i.e., geometric optimum experimental design (GOED), to select multiple representative samples in the database as the most informative ones for the user to label. Especially, GOED can alleviate the small-sized training data problem by leveraging the geometric structure of unlabeled samples in the reproducing kernel Hilbert space (RKHS) and thus further enhance the performance of image retrieval. Different from the conventional manifold regularization framework, the new method can effectively select the most informative samples for the user to label in image retrieval. By minimizing the expected average prediction variance on the test data, GOED has a clear geometric interpretation to select a set of the most representative samples in the database iteratively with the global optimum. Compared with the popular SVM active learning, our method is label-independent and can effectively avoid various potential problems caused by insufficient and inexactly labeled samples in RF, and is more appropriate and useful for image retrieval. Extensive experiments on both synthetic datasets and a real-world image database have been conducted to show the advantages of the proposed GOED for CIR.

Index Terms

content-based image retrieval, relevance feedback, active learning, optimum experimental design, manifold regularization

L. Zhang and L. Wang are with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. L. Zhang is also with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore. Email: liningzh@gmail.com, elpwang@ntu.edu.sg.

W. Lin is with the School of Computer Engineering, Nanyang Technological University, Singapore. Email: wslin@ntu.edu.sg.

I. INTRODUCTION

Content-based image retrieval (CBIR) has attracted much attention during the past decades [1], [2], [3]. However, the gap between low-level visual features and high-level semantic concepts usually leads to poor performance for CBIR. Relevance feedback (RF) [4] is one of the most powerful tools to narrow down this semantic gap by letting the user label semantically relevant and irrelevant images, which are positive and negative feedback samples, respectively.

Recently, a variety of RF approaches have been widely designed based on different assumptions for the positive and negative feedback samples to improve the performance of CBIR. In [5], one-class support vector machine (SVM) can estimate the density of positive feedback samples. Regarding the positive and negative feedback samples as two different classes, RF can be considered as an online binary classification problem. Two-class SVM was widely used to construct the RF schemes due to its good generalization ability [6], [7], [8], [9], [10]. With the observation that "all positive samples are alike and each negative sample is negative in its own way", RF was formulated as a biased subspace learning problem, in which there is an unknown number of classes, but the user is only concerned about the positive one [11], [12], [13], [14].

However, it is boring and tedious for the user to be asked to label a large number of samples in RF. To reduce the labeling efforts of conventional RF, a large amount of studies have attempted to accelerate this procedure by leveraging various auxiliary information. In this work, the paradigm of leveraging the auxiliary information for image retrieval can be referred to as "collaborative image retrieval (CIR)". In view of the characteristics of the research in this category, we briefly classify the studies on CIR into two categories. The first group of research intends to improve the performance of conventional RF by resorting to the user historical feedback log data or the large-scale web data [15], [16]. In [15], Hoi et al. proposed a log-based RF method, which can integrate the user historical feedback log data into the conventional RF and learn the correlation between low-level visual features and high-level semantic concepts. In [16], Liu et al. proposed an RF method for personal image retrieval via a cross-domain learning scheme, and it can effectively alleviate the labeling efforts of the user by leveraging a large number of loosely labeled web images. The second group of research attempts to select a set of the most informative samples from the database, which should be labeled by the user in RF and used as the training data to define an effective similarity metric for CIR.

SVM active learning (SVMactive) is one of the most popular methods in this category, which has been widely used to identify the ambiguous samples as the most informative ones by leveraging the optimal hyperplane of SVM [17], [18], [19]. To explain the mechanism of SVMactive, a simple toy example is illustrated in Fig.1. There are two labeled samples (i.e., the red solid dot "$\bullet$" for the positive feedback sample while the green cross point "$\times$" for the negative feedback sample) and several unlabeled ones (i.e., open circles "$\circ$"). The six samples distribute along a line and the coordinates on the horizontal axis indicate their positions. By using the SVM, the optimal hyperplane of the classifier f, which separates the two labeled feedback samples with a maximum margin, crosses position 0 as shown in Fig.1 with the dashed line. According to the most ambiguous criterion, i.e., the samples closest

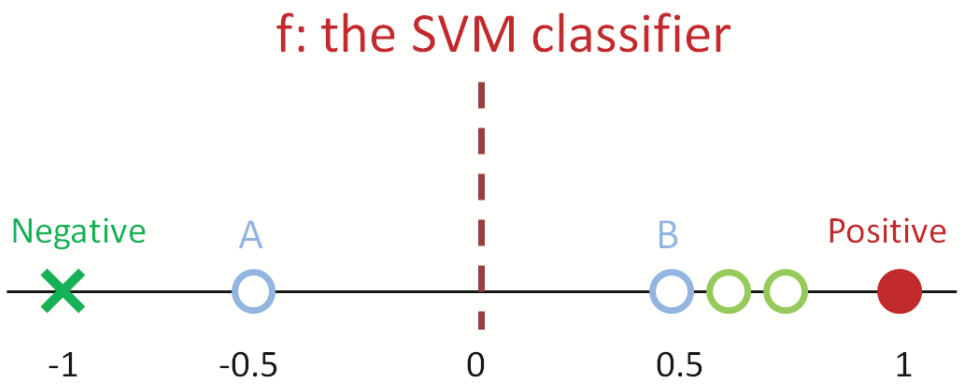


Fig. 1. A toy example to show the mechanism of SVMactive in selecting the most informative unlabeled samples for the user to label. The red solid dot, green cross point and open circles indicate the positive feedback sample, negative feedback sample and unlabeled samples, respectively. SVMactive can identify the ambiguous unlabeled samples "A" and "B" for their low prediction confidence but cannot select the more representative unlabeled sample "B".

to f having the maximum ambiguity, we can get that "A" and "B" have the maximum and identical ambiguity because they have the same distance, i.e., 0.5 for both, to the optimal hyperplane. Therefore, "A" and "B" should be identified by the user and used as the training data in RF. If we can choose only one sample for labeling, it is more reasonable to label "B" than "A" since more unlabeled samples are distributed around "B" and thus "B" is more effective than "A" to represent the distribution of unlabeled samples in the database. However, SVMactive can only select the ambiguous samples for the user to label although labeling representative ones may bring more useful information for achieving much better performance. Moreover, the optimal hyperplane of SVM is always unstable with small-sized training data [9], [10], i.e., this hyperplane is always sensitive when the size of the training data is small. Generally, in RF, the user would only label a small number of samples and cannot label each sample accurately all the time. Therefore, the optimal hyperplane of SVM cannot always be accurate with insufficient and inexactly labeled feedback samples. During the last decade, a large number of studies have been carried out to further enhance the performance of SVMactive for image retrieval [20], [21], [22], [23]. Nevertheless, these SVMactive methods always require an initial classifier model to identify the most informative samples and thus cannot be directly applied when there are no training data, which is however not appropriate for CIR [17], [20], [21], [22], [23].

In the machine learning community, there has been a long tradition of research on active learning [24], [25], [26], [27], [28]. In general, discriminative models aim to choose the most ambiguous samples, e.g., SVMactive methods [17], and generative models tend to select the most representative samples [27]. In statistics, active learning can be typically referred to as experimental design. The sample is referred to as experiment, and its label is referred to as measurement. The study of optimum experimental design (OED) [29] is concerned with the selection of the most informative experiments to measure given that conducting an experiment is expensive. Classical OED methods include A-OED, D-OED and E-OED, which are to maximize the confidence in a given model by minimizing the size of the estimated parameter covariance [29]. However, classical OED methods cannot characterize the quality

of predictions on test data if the test data are given beforehand. Transductive experimental design (TED) with different solutions [30], [31] was proposed to directly assess the quality of prediction on the given test data, which thus has a clear geometric interpretation to select the most representative samples in the database and has yielded impressive results for text categorization compared with classical OED methods (i.e., A-OED, D-OED and E-OED). Nevertheless, most of the experimental design methods can only evaluate the labeled samples while ignore abundant unlabeled samples although the unlabeled samples can also provide useful information and further enhance the performance of image retrieval. In RF, more often than not, the effort of labeling samples is generally laborious and vast amounts of unlabeled samples are readily available in the database. Various semi-supervised learning approaches under such a scenario have been widely developed to improve the generalization ability of supervised learning by leveraging the geometric structure of unlabeled samples [32], [33], [34], [35], [36], [37], [38]. In addition, conventional experimental design studies can only select one sample after another with a greedy strategy [39] or involve a semidefinite procedure with high computational burden [40] when dealing with abundant unlabeled samples, which will significantly limit their potential applications to CIR.

In this paper, we propose a novel active learning method, i.e., geometric optimum experimental design (GOED), by leveraging the geometric structure of unlabeled samples in the reproducing kernel Hilbert space (RKHS) to simultaneously select multiple representative samples in the database as the most informative ones for the user to label in CIR. The proposed scheme is largely inspired by the recent manifold regularization principle [38], [41], which is usually the key in semi-supervised learning to significantly improve the performance of supervised learning. However, different from the conventional manifold regularization framework [38], our method can effectively select the most informative samples for the user to label in image retrieval. In general, the new method can be divided into three independent stages. It first learns a data-dependent kernel function from both the labeled and unlabeled samples, which can be constructed by a conventional kernel function and then warped by a data-dependent norm to reflect the geometric structure of unlabeled samples in RKHS, and thus alleviates the small-sized training data problem in RF [37]. By minimizing the expected average prediction variance on the test data as suggested by TED [30], [31], GOED has a clear geometric interpretation to select representative samples in the database as the most informative ones. Finally, this kernel function can be used to identify a set of the most informative samples iteratively with the global optimum for the user to label. Different from the popular SVMactive [17], [20], [21], [22], [23], our method can select representative samples in the database as the most informative ones for the user to label, which is actually label-independent and thus can effectively avoid potential sensibility problem caused by insufficient and inexactly labeled feedback samples in SVMactive methods for image retrieval. Extensive experiments on both synthetic datasets and a real-world image database have demonstrated the advantages of the proposed GOED for CIR.

The rest of this paper is organized as follows: Section II briefly reviews the related work. In Section III, we propose GOED for CIR. A CIR system is introduced in Section IV. In Section V, we first give the experimental results on both synthetic datasets and a real-world image database, and then show the analysis of the important parameter in GOED. Section VI concludes this paper.

II. Related Work

To describe our method clearly, let us first review two areas of research that are closely related to our work in this paper, i.e., (1) CIR and (2) OED.

A. Review of CIR

To reduce the labeling efforts of the user in image retrieval, a variety of research work has been conducted regarding the paradigm of CIR [15], [16], [42], [43], [44].

Some of the studies have attempted to address the challenges encountered by conventional RF by resorting to the user historical feedback log data or the large-scale web data [15], [16], [44]. In [15], Hoi et al. proposed a log-based RF scheme with SVM by engaging the user historical feedback log data in a conventional online RF task. In [16], a textual query-based personal image retrieval system was proposed, which can significantly alleviate the labeling efforts of the user in RF by leveraging millions of loosely labeled web images.

Active learning is well-known for getting the necessary information by labeling as few samples as possible. SVMactive is one of the most popular techniques in this category for CIR, which has attracted much attention during the last decade [17], [20], [21], [22], [23]. To alleviate the small-sized training data problem, Wang et al. proposed to modify SVMactive with a transductive SVM by engaging unlabeled samples in the database [20]. In [21], Hoi et al. combined some semi-supervised learning techniques with the traditional SVMactive, which can also effectively exploit the information of unlabeled samples [21]. Despite the vast research work, SVMactive methods always require an initial optimal hyperplane to identify the most informative samples. However, this optimal hyperplane will not always be accurate with insufficient and inexactly labeled feedback samples, which is always the case in image retrieval.

Besides the aforementioned methods, some other research efforts have also been devoted to CIR [45], [39], [46]. In [45], a batch model active learning framework was proposed to employ the Fisher information matrix as an ambiguous measurement to select the most informative samples, which is fundamentally based on a probabilistic framework of the kernel logistic regression model. The author of [39] employed an experimental design criterion to identify one sample after another with a greedy strategy, which does not have a clear interpretation to the selected samples and is usually not the optimal solution to select multiple informative samples for the user to label.

B. Optimum Experimental Design

Given a set of unlabeled samples $X = \{x_1, \ldots, x_n\}$ in R^d, active learning aims to find a subset $Z = \{z_1, \ldots, z_l\} \subseteq X$ which contains the most informative samples. That is, if the samples $z_i (i = 1, \ldots, l)$ are labeled and used as the training data, we can predict the labels of unlabeled samples more precisely.

In statistics, active learning is typically referred to as experimental design and considers a linear regression model

$$y = w^T x + \varepsilon \tag{1}$$

where y is the observation, $w \in R^d$ is the parameter vector, $x \in R^d$ is the variable and ε is an unknown error with zero mean and constant variance σ^2. We define $f(x) = w^T x$ to be the prediction on an input variable x and the parameter vector w. Given a set of labeled samples $(z_1, y_1), \ldots, (z_l, y_l)$, a popular way to learn the prediction function f is to estimate w by minimizing the following objective function, i.e.,

$$J(w) = \sum_{i=1}^{l} \left(w^T z_i - y_i\right)^2 + \gamma_1 ||w||^2 \tag{2}$$

where $\gamma_1 > 0$ and $||\cdot||^2$ is the vector 2-norm. Let $Z = [z_1, \ldots, z_l]$ and $\mathrm{y} = [y_1, \ldots, y_l]^T$. The optimal solution to Eq.(2) is given by

$$\hat{w} = (ZZ^T + \gamma_1 I)^{-1} Z\mathrm{y} \tag{3}$$

Note that the introduced regularization term can improve the numerical stability of the solution since $\mathrm{ZZ}^T + \gamma_1 I$ is of full-rank. It has been shown that $\hat{w}$ is an unbiased estimation of w and its covariance matrix is given by $\sigma^2 C_w$, where C_w is the inverted Hessian of $J(w)$:

$$C_w = (\frac{\partial^2 J(w)}{\partial w \partial w^T})^{-1} = (ZZ^T + \gamma_1 I)^{-1} \tag{4}$$

Conventional OED methods can select the most informative samples, i.e., $z_1, \ldots, z_l$, by minimizing the size of the estimated parameter covariance, i.e., Eq.(4). The typical criteria are trace of C_w (i.e., A-OED), determinant of C_w (i.e., D-OED) and maximum eignevalue of C_w (i.e., E-OED) [29].

Conventional OED methods (i.e., A-OED, D-OED and E-OED) do not have a clear interpretation to the selected informative samples. To address this problem, TED aims to select representative samples in the database as the most informative ones by directly minimizing the prediction variance on the test data [30], [31]. For simplicity, we assume that the test data are given by $X = [x_1, \ldots, x_n]$, and thus the average prediction variance on the test data X is given by

$$\begin{aligned} &\tfrac{1}{n} \sum_{i=1}^{n} x_i^T Cov(\hat{w}) x_i \\ &= \tfrac{\sigma^2}{n} \sum_{i=1}^{n} x_i^T (ZZ^T + \gamma_1 I)^{-1} x_i \\ &= \tfrac{\sigma^2}{n} tr(X^T (ZZ^T + \gamma_1 I)^{-1} X) \end{aligned} \tag{5}$$

In order to minimize Eq.(5), we attempt to find a subset $Z \subseteq X$ which can minimize $tr(X^T (ZZ^T)^{-1} X)$. In [30], after some mathematical derivations, the problem of Eq.(5) can be reformulated as follows:

$$\min_{\alpha_\mathrm{i} \in R^l} \sum_{i=1}^{n} ||x_i - Z\alpha_\mathrm{i}||^2 + \gamma_1 ||\alpha_\mathrm{i}||^2 \tag{6}$$

The first term in Eq.(6) shows that the samples Z selected by TED are the most representative samples in the test data X. That is, the selected samples $z_i (i = 1, \ldots, l)$ can be used to reconstruct the data X most precisely

[30]. In [30], Yu et al. proposed a sequential greedy method to select one sample z_i at a time. However, the greedy solution to this problem is always suboptimal.

To obtain the global optimum, a convex relaxation of Eq.(6) was carefully designed in [31]. By introducing auxiliary variables $\beta = (\beta_1, \ldots, \beta_n)$ as a data selection coefficient to control the inclusion of samples into the training data, the optimization problem of Eq.(6) can be reformulated as follows:

$$\begin{aligned} &\min_{\alpha_{\mathrm{i}}, \beta \in R^n} \sum_{i=1}^{n} \left(||x_i - X\alpha_{\mathrm{i}}||^2 + \sum_{j=1}^{n} \frac{\alpha_{i,j}^2}{\beta_j} \right) + \lambda ||\beta||_1 \\ &s.t. \beta_i \geq 0, i = 1, \ldots, n \end{aligned} \tag{7}$$

where $\alpha_{\mathrm{i}} = (\alpha_{i,1}, \ldots, \alpha_{i,n})^T$, λ is the sparse regularization parameter and $||\cdot||_1$ denote the l_1 norm. As indicated by the well-known LASSO method [47], the minimization of the l_1 norm $||\beta||_1$ will lead to a sparse β. It is easy to check that, if $\beta_j = 0$, then all $\alpha_{1,j}, \ldots, \alpha_{n,j}$ must be zero; otherwise the objective function goes to infinity. Therefore, the jth sample will not be selected. It has been verified that the optimization of Eq.(7) is convex, and thus the global optimal solution can be obtained. For more details, please refer to [31].

Conventional experimental design methods (i.e., A-OED, D-OED, E-OED and TED) are designed based on supervised learning models (i.e., least-square ridge regression (LSRR)), which are usually not appropriate for CIR, since the effort of labeling samples is generally laborious and vast amounts of unlabeled samples are readily available in the database.

III. GOED FOR CIR

In this section, we first formulate the conventional RF in image retrieval as an active learning problem, and then propose a novel active learning method, i.e., GOED, to select multiple representative samples in the database as the most informative ones for the user to label. GOED can alleviate the small-sized training data problem by leveraging the geometric structure of unlabeled samples in RKHS. Compared with the popular SVMactive, our method is label-independent and can effectively avoid various potential problems caused by insufficient and inexactly labeled feedback samples, and is more effective and useful for CIR.

A. Problem Definition

Let $\phi : R^d \rightarrow H_K$ denote a feature mapping function from an input space R^d to an RKHS H_K, and $\phi(x)$ denote the sample x in RKHS. Regarding the positive and negative feedback samples as two different classes, RF can be considered as an online binary classification problem and a classifier is expected to learn for image retrieval, which can thus be used as a similarity metric to measure the similarity between a given sample x and the query image via

$$y = sign(f(x)) \tag{8}$$

where the classifier is assumed to be $f(x) = w^T\phi(x)$ in this paper. Here, if $y = 1$, then x belongs to the positive class; otherwise, it belongs to the negative class. It should be noted that a bias term can be incorporated into this form by expanding the weight and sample vector as $w \leftarrow [w, b]$ and $\phi(x) \leftarrow [\phi(x), 1]$. Generally, given a sample x, a small value of $|f(x)|$ means that the sample x is close to the optimal hyperplane and thus its corresponding prediction confidence will be low, and vise versa.

Suppose we have a set of labeled feedback samples $(z_1, y_1), \ldots (z_l, y_l)$, where y_i is the label of a sample z_i, a common principle for learning a good classifier f is to minimize the following structural risk in RKHS [48], [18], [19]

$$\hat{f} = \arg\min_{f \in H_K} \left(J(w) = \sum_{i=1}^{l} L(y_i, f(z_i)) + \gamma_1 ||f||_K^2 \right) \tag{9}$$

where $L(\cdot, \cdot)$ is the loss function (e.g. hinge-loss in SVM, squared-loss in LSRR), $\gamma_1 > 0$ and $||f||_K$ is an induced norm in an appropriately chosen RKHS H_K defined over a kernel Gram matrix K [49].

In general, the generic problem of active learning for RF is the following. Given a set of unlabeled samples $X = \{x_1, \ldots x_n\} \in R^d$ in the database, we aim to find a subset $Z = \{z_1, \ldots, z_l\} \subseteq X$, which contains the most informative ones for the user to label. In other words, the samples $z_i (i = 1, \ldots, l)$ can improve the performance of image retrieval the most if they are labeled and used as the training data to acquire a more effective similarity metric for image retrieval.

With the observation that the closer to the optimal hyperplane of f a sample is, the lower its prediction confidence is, SVMactive methods have been widely designed to select the unlabeled samples closest to the optimal hyperplane of f as the most informative ones so as to achieve maximal refinement on the hyperplane between the two classes [17], [20], [21], [22], [23]. However, the optimal hyperplane of f is label-dependent and often sensitive to the small-sized training data, which is always the case in image retrieval since the user would not like to label a large number of feedback samples and also cannot label each sample accurately all the time.

B. GOED for CIR

Conventional active learning methods (e.g., SVMactive and OED) are developed based on supervised learning models [17], [29], [30], [31], i.e., Eq.(9). However, in RF, the effort of requiring the user to label a large number of samples is generally laborious, although vast amounts of unlabeled samples are readily available in the database and can also provide useful information to enhance the performance of image retrieval. Semi-supervised learning under such a scenario is often designed to significantly improve the generalization ability of supervised learning by leveraging abundant unlabeled samples in the database. The common motivation of semi-supervised learning methods [50], [38], [34], [37], [32], [33], [41] is to exploit the intrinsic geometric structure of unlabeled samples by restricting the inductive prediction to comply with this geometry. The manifold regularization principle [38], [41], one of the most representative techniques, assumes that the geometry of the intrinsic data probability distribution is supported on a low-dimensional manifold. The manifold approximation term and supervised learning models are

combined together under the conventional regularization framework [48], which can smoothen the model output along the manifold estimated from the unlabeled samples [38], [41].

To alleviate the small-sized training data problem in RF, in this subsection, we propose a novel active learning framework by leveraging the geometric structure of unlabeled samples in RKHS to simultaneously select multiple informative samples for the user to label. The proposed scheme is largely inspired by the recent manifold regularization principle in the machine learning community [38], [41], which is usually the key in semi-supervised learning to significantly improve the performance of supervised learning, i.e.,

$$\hat{f} = \arg\min_{f \in H_K} \left(J(w) = \sum_{i=1}^{l} L(y_i, f(z_i)) + \gamma_1 ||f||_K^2 + \gamma_2 ||f||_I^2 \right) \tag{10}$$

where the additional $||f||_I^2$ is a smooth penalty term to reflect the intrinsic geometric structure of unlabeled samples. Parameters γ_1 and γ_2 are used to balance between the loss function and two regularization terms. The manifold regularization term $||f||_I^2$ plays a key role in semi-supervised learning and models the classifier output smoothness along the manifold estimated from both the labeled and unlabeled samples in the database [38], [41].

The proposed active learning framework shares a similar objective function with that of the manifold regularization framework in [38]. However, our method can effectively select the most informative samples for the user to label. Different from conventional active learning methods (i.e., SVMactive and OED), the new scheme can effectively alleviate the small-sized training data problem by exploiting the geometric structure of abundant unlabeled samples as in various semi-supervised learning methods [32], [33], [37], [34], [38], [35], [36]. In CIR, the system can effectively select the most informative samples for the user to label, which is an active learning problem. And then, if the most informative samples are labeled by the user, the system can use both the labeled and unlabeled samples to learn a classifier f and thus to measure the similarity between a given sample and the query image; this is actually a semi-supervised learning problem.

Given the above active learning framework for CIR, the key issue to attack this problem is first to design an appropriate RKHS induced by a kernel Gram matrix K, and thereafter find an effective way to identify the most informative samples for the user to label. In the following, we will study to define the kernel Gram matrix K and also to identify the most informative samples $z_i (i = 1, \ldots, l)$ for the user to label.

Kernel trick is widely applied in the hope of discovering the nonlinear structure of the data by mapping the original data into a higher dimensional RKHS [51]. The popular kernel functions include Gaussian and Polynomial ones. However, the nonlinear structure captured by a data-independent kernel function may not be consistent with the intrinsic geometric structure of data [37], [38]. To capture the geometric structure of data, in this work, we employ a data-dependent kernel function, which can be constructed by a conventional kernel function (e.g., Gaussian or Polynomial ones) from both the labeled and unlabeled samples with a kernel deformation principle [37]. For clarity, below we briefly review the kernel deformation principle introduced in [37].

Let H_K denote the original RKHS induced by a kernel Gram matrix K, and $\tilde{H}_{\tilde{K}}$ denote the new deformed RKHS induced by a new kernel Gram matrix $\tilde{K}$. In [37], the authors assume the following relationship between

the two spaces, i.e.,

$$< f, g >_{\tilde{H}_{\tilde{K}}} =< f, g >_{H_K} + \mathrm{f}^T M \mathrm{g} \tag{11}$$

where $\mathrm{f} = (f(x_1), \ldots, f(x_n))$ and $\mathrm{g} = (g(x_1), \ldots, g(x_n))$ are evaluate functions on both the labeled and unlabeled samples X, and M is used to capture the geometric structure of the data X. The deformation term, i.e., $\mathrm{f}^T M \mathrm{g}$, is introduced to assess the geometric relationship between the two functions f and g based on both the labeled and unlabeled samples. According to [37], a new kernel function $\tilde{k}(\cdot, \cdot)$ can thus be derived associated with a new deformed RKHS $\tilde{H}_{\tilde{K}}$ by

$$\tilde{k}(x_i, x_j) = k(x_i, x_j) - \gamma \mathrm{k}_{x_i}^T (I + MK)^{-1} M \mathrm{k}_{x_j} \tag{12}$$

where $k(\cdot, \cdot)$ is the original kernel function (e.g., Gaussian or Polynomial ones) defined in H_K with its corresponding kernel Gram matrix $K = [k(x_i, x_j)]_{n \times n}$, k_{x_i} is defined as $\mathrm{k}_{x_i} = [k(x_i, x_1), \ldots, k(x_i, x_n)]^T$, γ is the kernel deformation parameter and used to balance the original kernel function and the deformation term, and I is an identity matrix. The key issue now is the choice of M, which should be designed with respect to the intrinsic geometric structure of data X.

To capture the geometric structure of the data, we can turn to the graph Laplacian L as suggested by [37], [38]. Here, the graph Laplacian L is defined as $L = D - W$ or $L = D^{-\frac{1}{2}}(W - D)D^{-\frac{1}{2}}$ if normalized. The matrix $W \in R^n \times R^n$ is the data adjacency graph, wherein each element W_{ij} is an edge weight between two samples x_i and x_j. In the diagonal matrix $D \in R^n \times R^n$, the ith entry $D_{ii} = \sum_{j=1}^{n} W_{ij}$. Various extensions of the graph Laplacian have been proposed in [52]. A simple one is the following

$$W_{ij} = \begin{cases} 1, if x_i \in N(x_j) or x_j \in N(x_i) \\ 0, otherwise \end{cases} \tag{13}$$

where $N(x_i)$ denotes the k nearest neighbors of the sample x_i. The graph Laplacian provides the following smoothness penalty on the graph

$$\mathrm{f}^{\mathrm{T}} \mathrm{L} \mathrm{f} = \sum_{i=1}^{n} (f(x_i) - f(x_j))^2 W_{ij} \tag{14}$$

According to He and Niyogi [52], a definition in Eq.(14) corresponds to the approximation of $\int_M ||\nabla f(x)||^2$, the manifold on which the data X reside. As indicated in [37], by setting $M = L$, the modified kernel Gram matrix $\tilde{K}$ allows us to reconstruct algorithms for semi-supervised classification and reinterpret them within the standard framework of supervised learning [37], [38]. In particular, Laplacian SVM and Laplacian LSRR will become the standard SVM and LSRR by using the new kernel function (i.e., Eq.(12)), respectively. In this work, by adopting the squared-loss function as in experimental design methods [29], we can reformulate the active learning framework (i.e., Eq.(10)) as a supervised learning model in the new deformed RKHS, i.e.,

$$\hat{f} = \underset{f \in \tilde{H}_{\tilde{K}}}{\arg\min} \left\{ J(w) = \sum_{i=1}^{l} \left(w^T \tilde{\phi}(z_i) - y_i \right)^2 + \gamma_1 ||w||^2 \right\} \tag{15}$$

where $\tilde{\phi}(z_i)$ denotes the sample z_i in $\tilde{H}_K$, which can reflect the geometric structure of unlabeled samples in the database [37]. From the representation theorem [51], we know that w can be represented as a linear combination of $\tilde{\phi}(z_i), i = 1, \ldots, l$, i.e.,

$$w = \sum_{i=1}^{l} v_i \tilde{\phi}(z_i) = \tilde{\phi}(Z)\mathrm{v} \tag{16}$$

where $\mathrm{v} = [v_1, \ldots v_l]^T \in R^l$ is the expansion coefficient vector. Bring Eq.(16) back into Eq.(15), we can derive the following objective function

$$\hat{f} = \underset{f \in \tilde{H}_K}{\arg\min} \left\{ J(\mathrm{v}) = ||\tilde{K}_z \mathrm{v} - \mathrm{y}||^2 + \gamma_1 \mathrm{v}^T \tilde{K}_z \mathrm{v} \right\} \tag{17}$$

where $\mathrm{y} = [y_1, \ldots, y_l]^T$, and $\tilde{K}_z \in R^{l \times l}$ is defined only on the labeled set $\tilde{\phi}(Z) = [\tilde{\phi}(z_1), \ldots, \tilde{\phi}(z_l)]$ with its entries computed as in the kernel Gram matrix $\tilde{K}$. By setting $\frac{\partial J(\mathrm{v})}{\partial \mathrm{v}} = 0$, we can get the optimal solution to Eq.(17) as follows

$$\hat{\mathrm{v}} = (\tilde{K}_z + \gamma_1 I)^{-1} \mathrm{y} \tag{18}$$

For an input sample x, we can predict its label by

$$f(x) = \sum_{i=1}^{l} \tilde{k}(x, z_i) \hat{v}_i \tag{19}$$

where $\tilde{k}(\cdot, \cdot)$ is the new kernel function defined in Eq.(12). Therefore, Eq.(19) can be used to measure the similarity between an input sample x and the query image. Then, the next problem is how to identify a set of the most informative samples $\tilde{\phi}(z_i), i = 1, \ldots, l$ for the user to label. Let us consider the active learning framework with a supervised learning model in $\tilde{H}_{\tilde{K}}$, i.e., Eq.(15). We define $\tilde{\phi}(X) = [\tilde{\phi}(x_1), \ldots, \tilde{\phi}(x_n)]$ as the set of all samples and $\tilde{\phi}(Z) = [\tilde{\phi}(z_1), \ldots, \tilde{\phi}(z_l)]$ as the selected most informative ones in $\tilde{H}_{\tilde{K}}$. Motivated by TED [30], [31], we attempt to minimize the expected average prediction variance on the test data $\tilde{\phi}(X)$ in $\tilde{H}_{\tilde{K}}$. Similar to Eq.(6), the optimization problem can thus be reformulated accordingly as follows

$$\min_{\alpha_\mathrm{i} \in R^l} \sum_{i=1}^{n} ||\tilde{\phi}(x_i) - \tilde{\phi}(Z)\alpha_\mathrm{i}||^2 + \gamma_1 ||\alpha_\mathrm{i}||^2 \tag{20}$$

Consequently, GOED tends to select the representative samples $\tilde{\phi}(Z) = [\tilde{\phi}(z_1), \ldots, \tilde{\phi}(z_l)]$ that can span a linear space to retain most of the information of $\tilde{\phi}(X)$ in $\tilde{H}_{\tilde{K}}$, which thus has a clear geometric interpretation to the selected informative samples $\tilde{\phi}(Z)$ as TED [30], [31]. Moreover, different from SVMactive methods, GOED does not depend on the labels $y_i (i = 1, \ldots, l)$, but only on the training data $\tilde{\phi}(Z) = [\tilde{\phi}(z_1), \ldots, \tilde{\phi}(z_l)]$, which can effectively avoid various potential problems caused by insufficient and inexactly labeled samples in RF.

Similarly, we can introduce a convex relaxation of Eq.(20) to obtain the global optimum, i.e.,

$$\begin{aligned}&\min_{\alpha_{\mathrm{i}},\beta\in R^n}\sum_{i=1}^{n}\left(||\tilde{\phi}(x_i)-\tilde{\phi}(X)\alpha_{\mathrm{i}}||^2+\sum_{j=1}^{n}\frac{\alpha_{i,j}^2}{\beta_j}\right)+\lambda||\beta||_1\\&s.t.\beta_j\geq 0,j=1,\ldots,n\end{aligned}\tag{21}$$

where $\alpha_{\mathrm{i}}=(\alpha_{i,1},\ldots,\alpha_{i,n})^T$, λ is the sparse regularization parameter and $||\cdot||_1$ denotes the l_1 norm. The minimization of the l_1 norm $||\beta||_1$ can lead to a sparse coefficient β. Further, when $\beta_j=0$, all $\alpha_{1,j},\ldots,\alpha_{n,j}$ must be zero and the jth sample will not be selected as the most representative one. Therefore, β can be used as the data selection coefficient. As has been stated earlier, this optimization problem (i.e., Eq.(21)) is convex, and therefore the global optimum can be obtained [31]. In the following, we will discuss how to solve this problem.

Let D_β be a diagonal matrix with entries $\beta_1,\ldots,\beta_n$ and we have

$$\sum_{j=1}^{n}\frac{\alpha_{i,j}^2}{\beta_j}=\alpha_{\mathrm{i}}^T D_\beta^{-1}\alpha_{\mathrm{i}}\tag{22}$$

With some simple algebraic steps, we get

$$\begin{aligned}&\sum_{i=1}^{n}\left(||\tilde{\phi}(x_i)-\tilde{\phi}(X)\alpha_{\mathrm{i}}||^2+\sum_{j=1}^{n}\frac{\alpha_{i,j}^2}{\beta_j}\right)+\lambda||\beta||_1\\&=\sum_{i=1}^{n}\left(\begin{array}{l}(\tilde{\phi}(x_i)-\tilde{\phi}(X)\alpha_{\mathrm{i}})^T(\tilde{\phi}(x_i)-\tilde{\phi}(X)\alpha_{\mathrm{i}})\\+\alpha_{\mathrm{i}}^T D_\beta^{-1}\alpha_{\mathrm{i}}\end{array}\right)+\lambda||\beta||_1\\&=\sum_{i=1}^{n}\left(\begin{array}{l}\tilde{\phi}(x_i)^T\tilde{\phi}(x_i)-2\alpha_{\mathrm{i}}^T\tilde{\phi}(X)^T\tilde{\phi}(x_i)\\+\alpha_{\mathrm{i}}^T\tilde{\phi}(X)^T\tilde{\phi}(X)\alpha_{\mathrm{i}}+\alpha_{\mathrm{i}}^T D_\beta^{-1}\alpha_{\mathrm{i}}\end{array}\right)+\lambda||\beta||_1\end{aligned}\tag{23}$$

And then, taking the derivative of Eq.(23) with respect to α_{i} and requiring it to be zero, we have

$$-2\tilde{\phi}(X)^T\tilde{\phi}(x_i)+2\tilde{\phi}(X)^T\tilde{\phi}(X)\alpha_{\mathrm{i}}+2{D_\beta}^{-1}\alpha_{\mathrm{i}}=0\tag{24}$$

Thus, we can get α_{i}:

$$\alpha_{\mathrm{i}}=\left(D_\beta^{-1}+\tilde{\phi}(X)^T\tilde{\phi}(X)\right)^{-1}\tilde{\phi}(X)^T\tilde{\phi}(x_i)\tag{25}$$

Let $\tilde{K}_i$ be ith column vector of $\tilde{K}$, i.e.,

$$\begin{aligned}\tilde{K}_i&=\left(\tilde{\phi}(x_1)^T\tilde{\phi}(x_i),\ldots,\tilde{\phi}(x_n)^T\tilde{\phi}(x_i)\right)^T\\&=\tilde{\phi}(X)^T\tilde{\phi}(x_i)\end{aligned}\tag{26}$$

Since $\tilde{\phi}(X)^T\tilde{\phi}(X)=\tilde{K}$, Eq.(25) can be rewritten as follows:

$$\alpha_{\mathrm{i}}=(D_\beta^{-1}+\tilde{K})^{-1}\tilde{K}_i\tag{27}$$

TABLE I
GOED FOR CIR

Input: The image database X with n unlabeled samples, the number of selected most informative samples l, the number of nearest neighbors k, the regularization parameter γ_1, the data-dependent kernel deformation parameter γ, the sparse parameter λ and the kernel type
Step 1: Construct a nearest neighbor graph with the weight matrix W as in Eq.(13) on the data X and calculate the graph Laplacian $L = D - W$
Step 2: Compute the conventional kernel Gram matrix K with the input kernel type and let $M = L$
Step 3: Compute the data-dependent kernel Gram matrix $\tilde{K}$ according to Eq.(12)
Step 4: Initialize $\alpha_{i,j} = 1$, and iteratively compute β_j and α_{i} according to Eq.(29) and Eq.(27), respectively, until convergence
Step 5: Rank the samples in X according to $\beta_j (j = 1, \ldots, n)$ in a descending order and return the top l samples as the most informative ones Z
Step 6: The l selected most informative samples Z should be labeled by the user and used as the training data to obtain a classifier f according to Eq.(17) and Eq.(18)
Output: The classifier f, which can be used as a similarity metric to measure the similarity between a given sample x and the query image, i.e., Eq.(19)

Once α_{i} is calculated, we can fix α_{i} and find the minimum solution for β_j. Taking the derivative of Eq.(21) with respect to β_j and requiring the derivative to be zero, we can have

$$\sum_{i=1}^{n} (-\frac{\alpha_{i,j}^2}{\beta_j^2}) + \lambda = 0 \tag{28}$$

Finally, we get the data selection coefficient

$$\beta_j = \sqrt{\frac{\sum_{i=1}^{n} \alpha_{i,j}^2}{\lambda}} \tag{29}$$

$\alpha_{i,j}$ and β_j can be computed iteratively according to Eq.(27) and Eq.(29), respectively. The objective function of Eq.(21) is convex, and thus the global optimum can be obtained after iterations.

We rank the samples in X according to the data selection coefficient β in a descending order and the top l samples are selected as the most informative ones Z. The system requires the user to label the relevant and irrelevant images in Z as the positive and negative feedback samples in RF, respectively. Finally, these samples can be used to train a classifier f according to Eq.(17) and Eq.(18) and then the classifier can be used as a similarity metric to measure the similarity between a given sample x and the query image based on the output of f, i.e., Eq.(19).

The GEOD for CIR can be summarized in Table I.

The computational complexity of constructing the k nearest neighbor graph in Step 1 is $O(kn^2)$, where n is the number of unlabeled samples. The computational complexity of computing the conventional kernel Gram matrix K in Step 2 is $O(n^2)$. The computational complexity of computing the data-dependent kernel Gram matrix $\tilde{K}$ in Step 3 is $O(n^3)$ and the Step 4 requires $O(tn^3)$, where t is the iteration times. In our experiments, GOED converges very fast and t is usually a small number of iterations. Therefore, the overall computational complexity of GOED is $O(n^3)$.

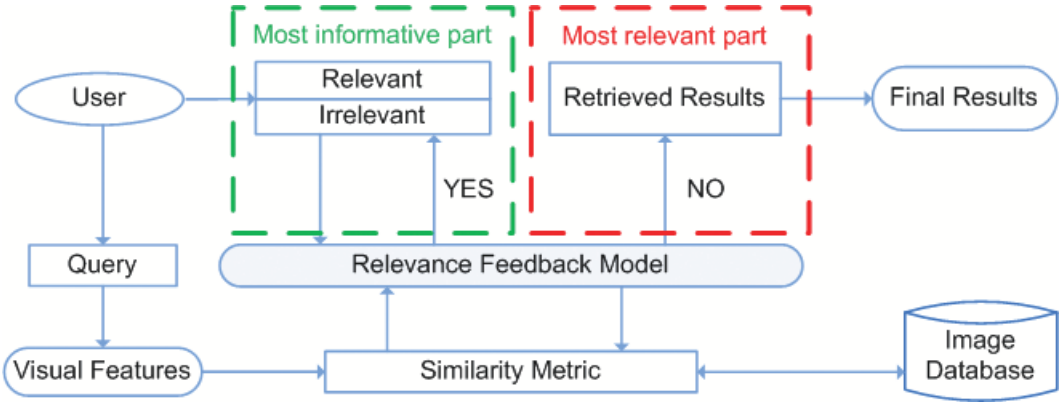


Fig. 2. The framework of our CIR system. Different from conventional CBIR systems, our system can select the most informative images in the database for the user to label and return the most semantically relevant images as the final results.

IV. The Collaborative Image Retrieval System

A. Overview of our CIR Framework

In this section, we first give an overview of our CIR system. As shown in Fig.2, when a query image is provided, the low-level visual features are first extracted. Then, all images in the database are sorted based on a predefined similarity metric. If the user is satisfied with the results, the image retrieval process is ended. However, most of the time, RF is actually required because of the poor performance of the system. The CIR requires the user to label some top informative images as the positive and negative feedback samples. Using these labeled samples as the training data, an RF model can be obtained based on certain machine learning techniques. The similarity metric can thus be updated together with the RF model. Then, all images are resorted based on the recalculated similarity metric. If the user is satisfied with the refined results, RF is no longer required and the system gives the final results, which are most semantically relevant to the query image. Otherwise, RF will be performed iteratively.

From Fig.2, it can be noticed that our CIR system is different from conventional CBIR systems, which can only present relevant and irrelevant images in the top returned results for the user to label and is not proper since the top returned results will not always be the most informative ones. The CIR system can accomplish a retrieval task with fewer iterations than conventional RF schemes based CBIR systems and alleviate the labeling efforts of the user with the help of the most informative samples in the database.

B. Corel Image Database and Image Representations

To perform empirical evaluation of the proposed method, first we should provide a reliable image database with semantic groups. The Corel photo gallery is a professionally catalogued image database and has been widely used to evaluate the performance of CBIR during the past a few years [12], [43], [44], [10]. To validate the effectiveness of the proposed method, we group the images into a number of classes based on the ground truth. The original Corel photo gallery includes many semantic categories, each of which contains 100 or more images. However, some of the categories are not suitable for image retrieval, since some images with different concepts are in the

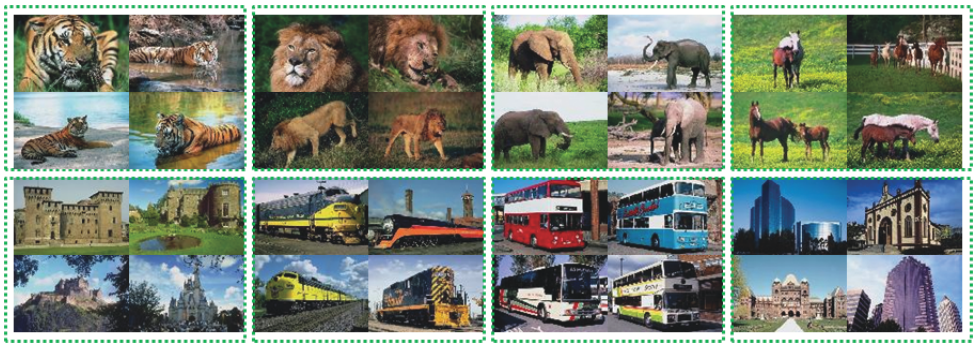


Fig. 3. Some example images in the Corel image database.

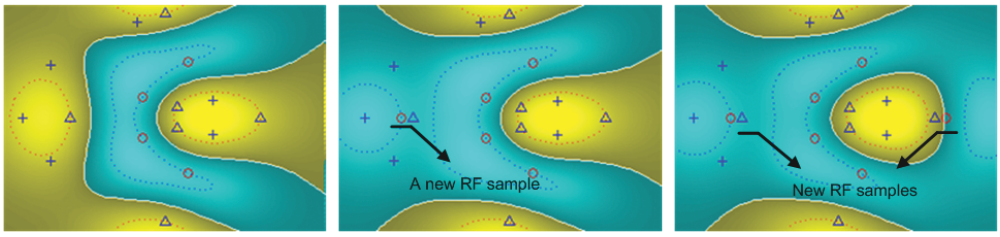
same category. Therefore, existing categories of the original Corel photo gallery are ignored and the images are reorganized into 80 conceptual classes based on the ground truth, such as lion, castle, bus, aviation, dinosaur and horse. Note that each class of the Corel photo gallery has a clearly distinct concept and the quality of the images can be considered very high. Finally, the Corel image database comprises totally 10,763 real-world images. This way of using the images with semantic categories can help evaluate the performance automatically, which can significantly reduce the subjective error compared with the manual evaluation. Some example images in the Corel image database are shown in Fig.3.

To represent images in the database, we use three different sets of low-level visual features in a 631-D space, i.e., 256-D RGB color histogram, 75-D edge distribution histogram and 300-D Bag of Words (BOW) [53]. For the generation of visual words, we firstly apply the Difference-of-Gaussian filter [54] on the grayscale image to detect a set of salient points; then we compute the Scale-Invariant-Feature-Transform (SIFT) [55] features over the local areas defined by the detected salient points; finally we perform the vector quantization on the descriptors to construct the visual vocabulary by using the k-means clustering approach. In this work, 300 clusters are generated and thus the dimension of BOW features is 300. All feature components are normalized to a normal distribution with zero mean and one standard deviation to represent the images.

V. Experimental results

A. Experiments on synthetic datasets

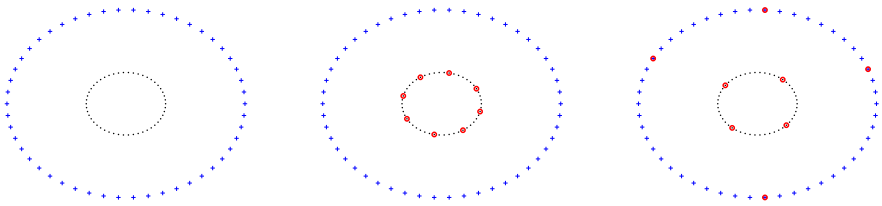
1) SVM is unstable with small-sized training feedback samples: SVMactive is one of the most popular methods to identify the most informative samples for CIR, which requires the user to label the unlabeled samples closest to the optimal hyperplane of SVM because of their low predication confidence. However, the optimal hyperplane of SVM is always unstable with small-sized training feedback samples, i.e., this optimal hyperplane is always sensitive to training data when the size of the training data is small, which is always the case in image retrieval. To visualize the unstable problem with small-sized training feedback samples, firstly we use a toy problem to illustrate the sensitivity of the optimal hyperplane of SVM in RF. In Fig.4, we use the open circles, triangles and crosses to denote the positive feedback samples, negative feedback samples and unlabeled samples in the database,



(a) (b) (c)

Fig. 4. Illustration of the unstable problem of the optimal hyperplane of SVM when dealing with small-sized training feedback samples. The open circles, triangles and crosses denote the positive feedback samples, negative feedback samples and unlabeled samples, respectively. The white solid line indicates the optimal hyperplane of SVM, which separates the positive and negative feedback samples. The unlabeled samples closest to the optimal hyperplane of SVM will be identified as the most informative ones.

respectively. Fig.4 (a) shows the optimal hyperplane of SVM, which is trained by the original feedback samples and adopted to identify the most informative unlabeled samples for the user to label. The unlabeled samples closest to the optimal hyperplane of SVM will be identified as the most informative ones for the user to label. However, as shown in Fig.4 (b) and (c), much different optimal hyperplanes will be trained by the original training data with one and two incremental positive feedback samples, respectively. Consequently, the optimal hyperplane of SVM is usually unstable and it is not appropriate to directly use this hyperplane to identify the most informative samples when dealing with small-sized training feedback samples.



(a) (b) (c)

Fig. 5. Performance comparison between TED and GOED in selecting the most representative samples. The black dots and blue crosses indicate the relevant and irrelevant unlabeled samples in the database, respectively. The red open circles are the most representative samples selected by TED and GOED in the database.

2) GOED can select the most representative samples in the database: GOED aims to select representative samples in the database as the most informative ones for the user to label, which is fundamentally based on TED with a convex solution in the new deformed RKHS [31]. The difference between the two algorithms is whether the geometric structure of unlabeled samples is fully exploited. To visualize the effectiveness of GOED in selecting the most representative samples, we give a simple toy example to compare the effectiveness of the two algorithms. In Fig.5, we use two sets of samples on two circles to indicate the relevant and irrelevant unlabeled samples with

the query image in the database. Especially, the black dots on the small circle represent the relevant unlabeled samples and the blue crosses on the big circle denote the irrelevant unlabeled samples, respectively. The eight most informative samples in the database are required to be selected by TED and GOED for the user to label. Both of the two algorithms use the Gaussian kernel function. As we can see, all samples selected by TED are from the set of relevant samples on the small circle, while GOED can select four relevant samples on the small circle and four irrelevant samples on the big circle, respectively. Clearly, the informative samples selected by GOED can better represent the original database (i.e., two circles). It should be noted that we do not compare our method with SVMactive because SVMactive cannot be applied in this case due to the lack of training data.

B. Experiments on a real-world image database

In this subsection, we will evaluate the effectiveness of the proposed GOED in selecting representative samples in the database as the most informative ones for the user to label in image retrieval based on two experiments: first, we investigate the GOED method for CIR by comparing it with a conventional RF scheme based CBIR system and thus show the potential of CIR in improving the performance of image retrieval; then, we show the performance of our CIR system by selecting the most informative samples for the user to label and compare it with some popular RF schemes for image retrieval based on a large real-world image database.

We adopt widely used average precision (AP), standard deviation (SD) and average recall (AR) to evaluate the performance of the compared algorithms. AP is the major evaluation criterion, which evaluates the effectiveness of the compared algorithms. In experiments, we empirically set the regularization parameter γ_1 as $1e-3$ and the sparse regularization parameter as 100 in experiments. For all kernel-based methods, we choose the Gaussian kernel function. Considering the computable efficiency, we do not use all unlabeled images in the database but only the top 500 returned results in the previous round of RF. The number of nearest neighbors k is empirically set as 4 according to [52]. It should be noted that both the SVM and SVMactive methods adopt the hinge-loss function, and all other methods (i.e., GOED, TED and LSRR) use the squared-loss function. In experiments, the data-dependent kernel deformation parameter γ is empirically set as 0.1, which will be analyzed carefully in the next subsection.

1) Performance evaluation on a small-scale image database: In this part, we intend to examine how effective the proposed GOED is when selecting the most informative samples for the user to label in image retrieval, and thus evaluate the potential of CIR compared with the conventional RF based CBIR systems, which can only select the top returned samples for the user to label. Basically, both GOED and TED are designed based on the LSRR model in RKHS. However, LSRR can only require the user to label the top returned images. Therefore, we compare GOED and TED with LSRR and show the performance difference of these algorithms. The experiments are conducted on a small-scale image database, which includes 3,139 images with 30 different categories. All 3,139 images in the 30 categories are used as the query images to evaluate the compared algorithms. In this experiment, we select an equal number of relevant and irrelevant images as the positive and negative feedback samples, respectively. For GOED, the first 3 relevant images and first 3 irrelevant images in the 20 most informative samples selected by the algorithms are labeled, whereas for LSRR, the first 3 relevant images and first 3 irrelevant images in the top 20

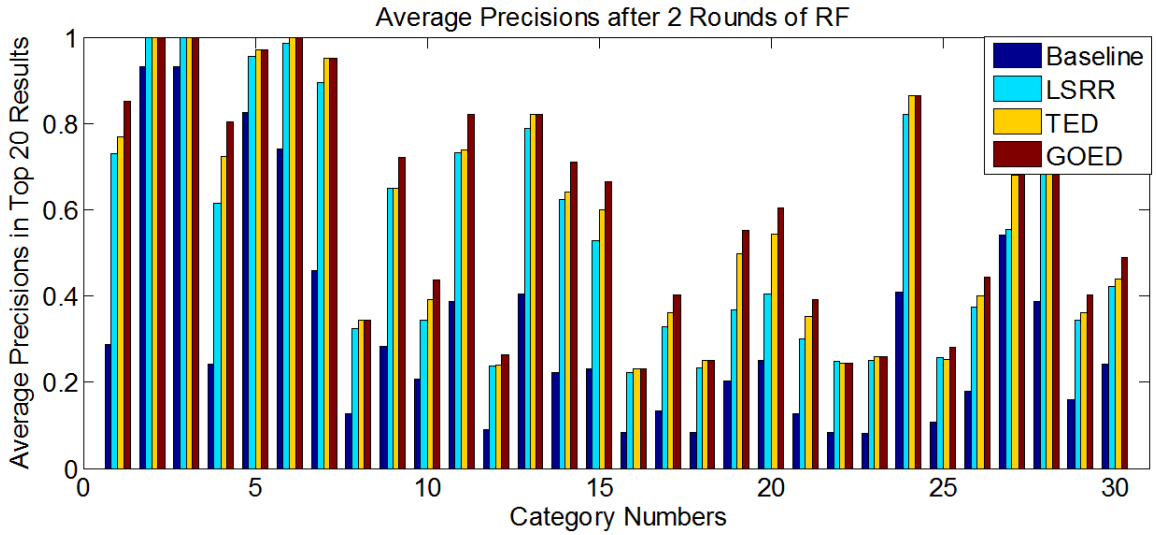
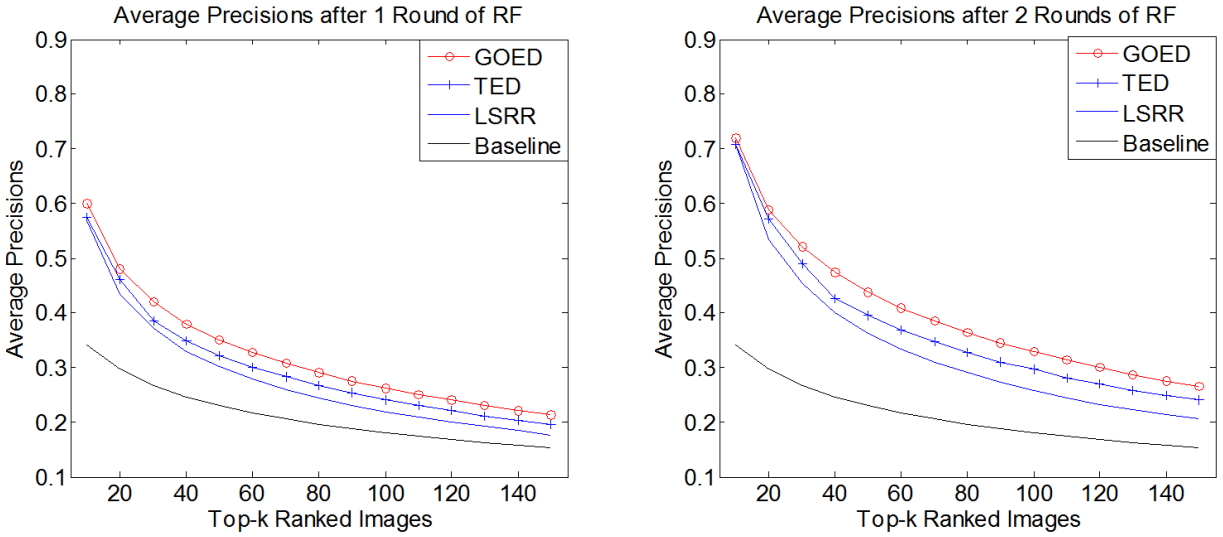


Fig. 6. APs with 30 categories in the top 20 results of the compared algorithms (i.e., GOED, LSRR, TED and Baseline) after the second round of RF.



(a) (b)

Fig. 7. APs of the compared algorithms (i.e., GOED, TED, LSRR and Baseline) after the first and second round of RF. (a) the first round of RF (b) the second round of RF.

returned samples are marked as the feedback samples.

Fig.6 shows APs in the top 20 retrieved results after the second round of RF for all 30 categories. The baseline describes the initial retrieved results without any RF information. Especially, at the beginning, the Euclidean distance metric in the high-dimensional feature space are used to measure the similarity between the query image and the images in the database. After the user provides the RF information, these methods can be applied to train a classifier

and thus define a new similarity metric to resort the images in the database. As shown in Fig.6, the performance of the three algorithms varies with different categories. For some easy categories, all of these three algorithms can perform well (e.g., Categories 2 and 3). For some hard categories, all of these three algorithms poorly perform (e.g., Categories 16, 18, 22 and 23). After the second round of RF, all of these three algorithms can show much better performance than the baseline and therefore significantly improve the performance of image retrieval. Compared with LSRR, both GOED and TED can perform much better for most of the 30 categories (e.g., Categories 1, 4, 19, 20 and 27), which indicates that both GOED and TED are more effective for image retrieval since these two methods can automatically select the most informative samples for the user to label. Compared with TED, GOED can show much better performance for most of the results, since GOED tries to find the most informative samples for the user to label by leveraging the geometric structure of abundant unlabeled samples.

Fig.7 (a) and (b) show APs of the compared algorithms after the first and second round of RF, respectively. As shown in Fig.7, we can notice that all the three algorithms (i.e., GOED, TED and LSRR) can show much better performance than the baseline on the entire scope, particularly after the second round of RF. This is mainly because by iteratively cumulating the RF information, more feedback samples will be used as the training data and therefore significantly enhance the performance of the system. Compared with LSRR, both GOED and TED can show much better performance on the entire scope, since both GOED and TED can effectively select the most informative samples for the user to label while LSRR can only label the top returned samples.

Based on the aforementioned results, we notice that both GOED and TED are more effective than LSRR thus show the potential of CIR by selecting the most informative samples for the user to label in improving the performance of image retrieval. We should highlight that, in the first round of RF, there are no training data and thus conventional SVMactive cannot be applied to select the most informative samples for the user to label. Compared with SVMactive, GOED is label-independent and thus will be more appropriate for CIR.

2) Performance evaluation on a large-scale image database: In this part, we design a scheme to model the real-world image retrieval process. In a real-world image retrieval system, a query image is usually not in the database. To simulate such an environment, we use a fivefold cross validation database to evaluate the compared algorithms. More precisely, we divide the whole image database into five subsets with equal size. Therefore, there are 20 percent per category in each subset. At each run of cross validation, one subset is selected as the query set, and the other four subsets are used as the database for image retrieval. Then, 500 query images are randomly selected from the query set, and RF is automatically conducted by the system. For each query image, the system retrieves and ranks the images in the database. Finally, 9 rounds of RF are automatically carried out by the system.

To show the effectiveness of the proposed GOED, we compare it with TED, LSRR, SVMactive, SVM, biased discriminant analysis (BDA)[11] and generalized BDA (GBDA)[14]. Out of these seven algorithms, GOED, TED and SVMactive are active learning methods, whereas LSRR and SVM are standard classification algorithms. Different from classification-based RF schemes, both BDA and GBDA are discriminant analysis-based RF schemes. BDA is one of the most promising RF approaches to deal with the feedback samples imbalance problem for CBIR. However, the singular problem of the positive within-class scatter and the Gaussian distribution assumption for

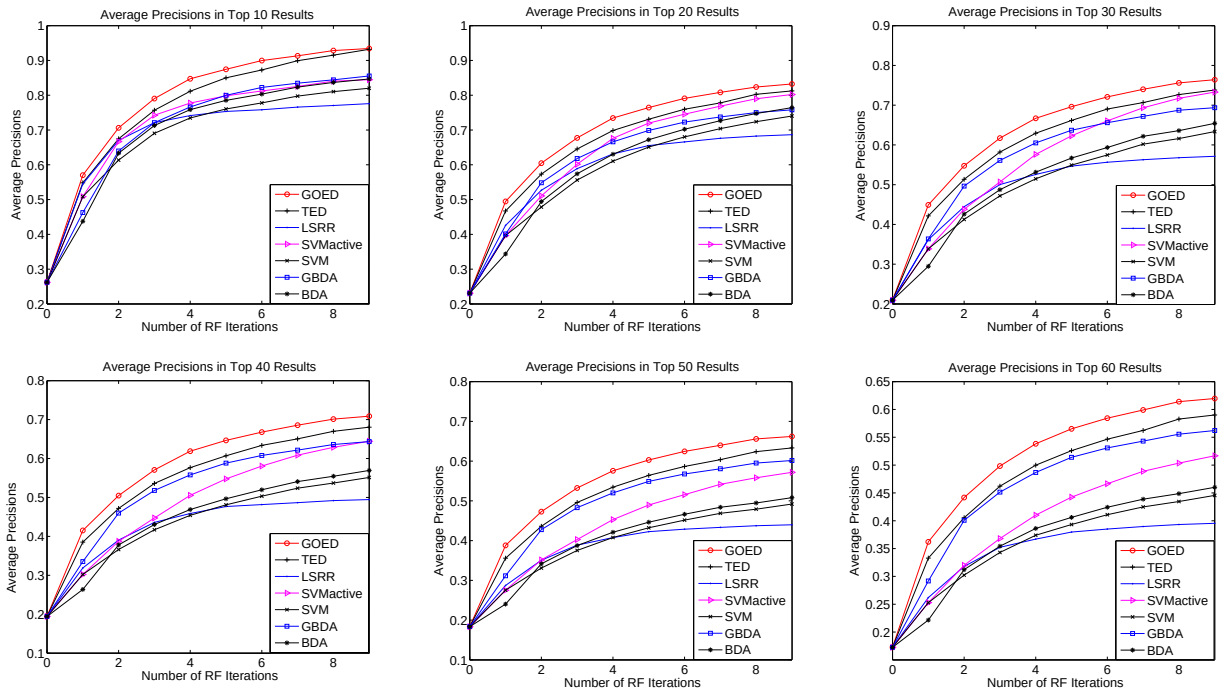


Fig. 8. APs of GOED compared with the RF methods, i.e., TED, SVMactive, SVM, LSRR, GBDA and BDA. All the methods are evaluated over 9 rounds of RF. 0-round of RF refers to the retrieved results based on the Euclidean distance metric without any RF information.

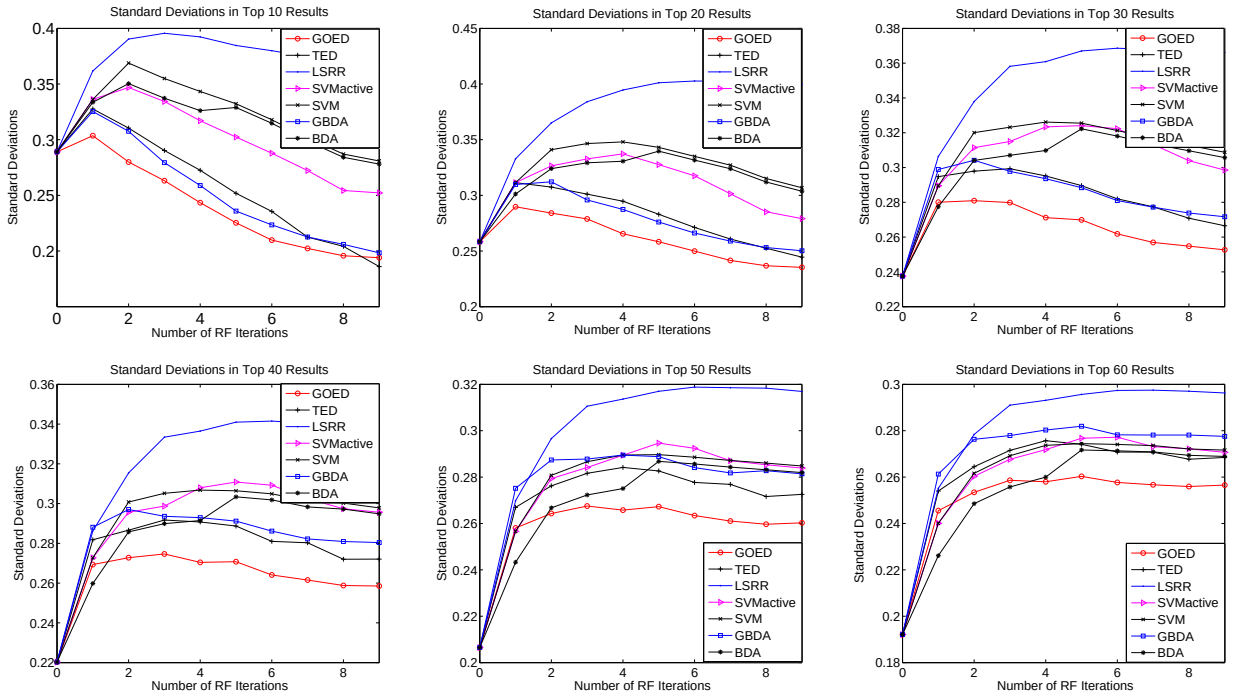


Fig. 9. SDs of GOED compared with the RF methods, i.e., TED, SVMactive, SVM, LSRR, GBDA and BDA. All the methods are evaluated over 9 rounds of RF. 0-round of RF refers to the retrieved results based on the Euclidean distance metric without any RF information.

positive samples are two main obstacles impeding the performance of BDA RF for CBIR. GBDA can avoid these two drawbacks of BDA within one framework and thus significantly improve the performance of BDA RF for CBIR. In each round of RF, 20 images are picked from the database and examined serially to mark as the positive or negative feedback samples. In general, in a real-world image retrieval system, the irrelevant images usually largely outnumber the relevant ones. To simulate such a case in the system, the first 3 relevant images are labeled as the positive feedback samples, and all other irrelevant images in the 20 images are automatically labeled as the negative feedback samples. The images that have been selected in previous RF iterations are excluded from later selections. It should be noted that for active learning-based RF methods (i.e., GOED, TED and SVMactive), the 20 images are selected from the database by the algorithms themselves, whereas for conventional classification-based RF methods (i.e., SVM and LSRR) and discriminant analysis-based RF methods (i.e., GBDA and BDA), the 20 images are composed of the top 20 returned images in the previous round of RF, which is the most popular way to select the feedback samples in the existing research of CBIR. In this experiment, we calculate the APs over the 500 query images at different positions from the top 10 to top 40 to obtain the APs, and all experimental results are computed from the fivefold cross validation.

Fig.8 and Fig.9 show the APs and SDs of the compared algorithms, respectively. As shown in Fig.8, GOED consistently outperforms all the other compared algorithms on the entire scope. GOED and TED are mainly designed based on the LSRR model in RKHS; however, both of them can significantly improve the performance of LSRR by selecting the most informative samples for the user to label. Compared with TED, GOED performs much better for all top results, since GOED tries to find the most informative samples for the user to label by leveraging the manifold structure of the data on which the classifier is as smooth as possible. SVMactive cannot show better performance than GOED and TED, since the optimal hyperplane of SVM is usually not very stable and accurate with small-sized training data in a high dimensional space (i.e., about tens of samples in a 631 dimensional space in this paper). Therefore, it is not appropriate to directly use the optimal hyperplane of SVM to identify the most informative samples when the number of the training data is small. Different from SVMactive, GOED and TED can select representative samples in the database as the most informative ones for the user to label, which are actually label-independent and therefore more appropriate for image retrieval. Moreover, we should indicate that SVMactive can only be applied when there is an initial classifier. Therefore, it cannot be applied in the first round of RF. In experiments, we use the standard SVM to build an initial classifier. When considering more rounds of RF, SVMactive can get little improvement over the standard SVM. Compared with LSRR, SVM can show slightly better performance than LSRR for all top results especially after a few rounds of RF since the loss function in SVM (i.e., hinge-loss) has much better generalization ability than the loss function in LSRR (i.e., squared-loss). GOED can outperform other compared methods for most of the results since it can select the most representative samples in the database for the user to label, which is more practical and useful for image retrieval.

Regarding the stability of the compared algorithms, we can also notice that GOED performs best among the top 10, 20, 30 and 40 results as shown in Fig.9. Then, for other top results, the performance of GOED is similar to the other compared algorithms. The detailed results of the compared algorithms after 9 rounds of RF are shown in

TABLE II
APS AND SDS IN TOP N RESULTS OF FIVE ALGORITHMS (I.E., GOED, TED, LSRR, SVMACTIVE, SVM, GBDA AND BDA) AFTER THE NINTH ROUND OF RF (APS ± SDS).

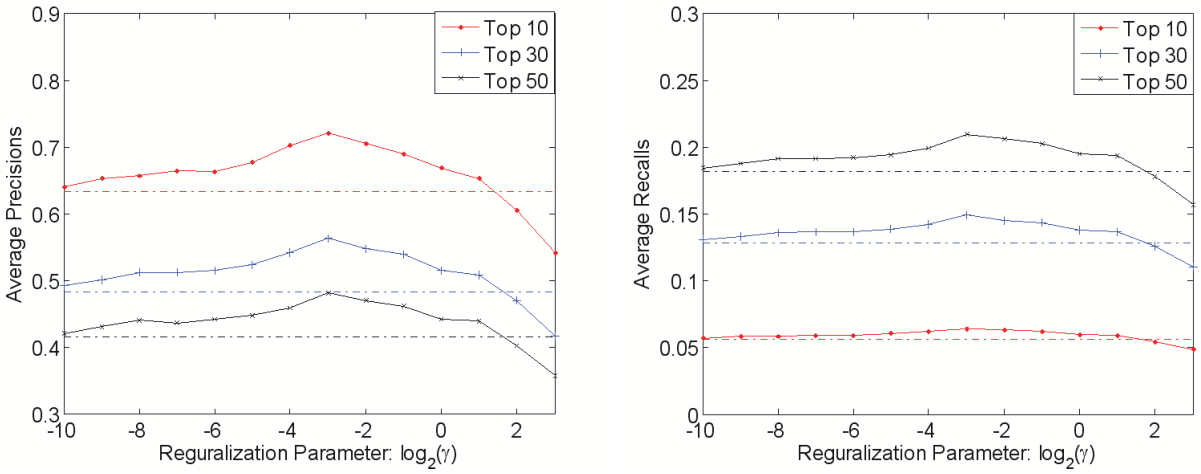
Methods	GOED	TED	LSRR	SVMactive	SVM	GBDA	BDA
Top 10	0.93±0.19	0.93±0.20	0.78±0.37	0.84±0.25	0.82±0.28	0.86±0.20	0.85±0.28
Top 20	0.83±0.24	0.81±0.25	0.69+0.40	0.80±0.28	0.74±0.31	0.76±0.25	0.76±0.30
Top 30	0.76±0.25	0.74±0.27	0.57±0.37	0.73±0.30	0.63±0.31	0.69±0.27	0.65±0.31
Top 40	0.71±0.26	0.68±0.27	0.49±0.34	0.64±0.30	0.55±0.30	0.64±0.28	0.57±0.29
Top 50	0.66±0.26	0.63±0.27	0.43±0.32	0.57±0.28	0.49±0.28	0.60±0.28	0.51±0.28
Top 60	0.62±0.26	0.59±0.27	0.40±0.30	0.52±0.27	0.45±0.27	0.56±0.28	0.46±0.27
Top 70	0.58±0.25	0.55±0.26	0.36±0.28	0.47±0.26	0.41±0.26	0.53±0.27	0.42±0.25
Top 80	0.55±0.24	0.52±0.25	0.33±0.26	0.43±0.24	0.37±0.24	0.50±0.26	0.39±0.24
Top 90	0.52±0.25	0.49±0.24	0.31±0.25	0.40±0.23	0.35±0.23	0.47±0.25	0.36±0.23

Table II. As given in Table II, GOED achieves much better performance compared with other approaches for all top results. TED still obtains satisfactory performance, as compared with SVM, LSRR and SVMactive. Therefore, we can conclude that the proposed GOED has shown its effectiveness in selecting representative samples in the database for the user to label in image retrieval.

It should be noted that the performance difference between this work and the previous research is mainly due to different experimental setup and platforms. As indicated in [17], to obtain a stable classifier, the user is usually required to randomly select a large number of samples to label in the first round of RF and a large number pool samples closest to optimal hyperplane of SVM (e.g., more than 20 samples in a 144 dimensional space), which can then be used to train a classifier and identify the ambiguous samples for the user to label. However, in practice, the positive feedback samples are usually much less than the negative feedback ones [9] and it is also not appropriate to require the user to label a large amount of samples in RF. In this work, we select at most 3 relevant images and all other irrelevant images in 20 returned samples in a 631 dimensional space as the feedback samples, which is similar to a real-world image retrieval process. Moreover, we use a fivefold cross validation database to evaluate the effectiveness of the compared algorithms, and this is different from previous studies [17].

C. Parameter sensitivity

In this subsection, we study the sensitivity of GOED with regard to the kernel deformation parameter γ in constructing the data-dependent kernel function, which is the key in GOED to significantly improve the performance of image retrieval by leveraging the geometric structure of unlabeled samples in RKHS. With the observation that the difference between GOED and TED is whether the geometric structure of unlabeled samples is fully exploited, we show the performance difference between the two methods. The analysis is performed based on experiments conducted on the small-scale image database given in subsection V-B-1. In experiments, we use all samples in the database to construct the data-dependent kernel function and 500 images are randomly selected from the database and used as the query images. For each query image, the system retrieves and ranks the images in the database



(a) (b)

Fig. 10. The performance sensitivity of GOED with regard to the data-dependent kernel deformation parameter γ. (a) APs (b) ARs

and 4 rounds of RF are automatically carried out by the system. In each round of RF, 20 most informative samples selected by the algorithms are examined from the top. The first 3 relevant and first 3 irrelevant images in the 20 selected informative samples are automatically marked as the positive and negative feedback samples by the system, respectively.

In experiments, we empirically set the kernel deformation parameter γ as a value in a sequence, i.e., $\{2^i, i = -10, -9, \ldots, 2, 3\}$. Fig.10 shows how APs and ARs of GOED vary with γ in the top 10, 30 and 50 results after 4 rounds of RF. The dashed lines indicate the corresponding results of TED. We can notice that the parameter γ can significantly affect the performance of GOED with regard to APs and ARs. As shown in Fig.10, when γ is small enough, the APs and ARs of GOED will approximate the corresponding results of TED. This is mainly because GOED will result in TED when the parameter γ is set as 0, i.e., GOED will be a standard supervised learning method and cannot exploit the geometric structure of unlabeled samples as TED. When γ becomes larger, GOED shows much better performance regarding the APs and ARs. However, if γ is too large, the performance may be degenerated, and this is mainly due to the overdeformation in the new RKHS. As we can see, GOED achieves consistently good performance with γ varying from $2^{-4} \sim 2^{-2}$. The parameter can be further tuned to achieve better performance. The analysis above also indicates that the GOED can effectively alleviate the small-sized training data problem by leveraging the geometric structure of unlabeled samples in RKHS.

D. Discussions

In the proposed CIR system, several aspects can be improved. For instance, a much larger image database will be utilized in the current platform. Recently, CBIR based on a large scale social web database (e.g., 1 million Flickr images) has attracted much attention [3], [56]. In these systems, large scale social web images are first

selected from social web sites (e.g., Flickr) and then manually grouped into semantic classes according to the associated textual information. However, different users have different opinions on a same web image, and thus will categorize the same image into different semantic groups. The CBIR results from such image databases will be subjective and are difficult to objectively evaluate or compare. Moreover, due to the noise textual information, it is still a problematic issue to categorize the images into semantic groups according to their rich associated tags. Consequently, it is interesting to objectively evaluate the performance of a CBIR system based on a large scale noisy social web database in our future study; newly proposed features may outperform the conventional ones, e.g., sparse coding representations [57].

The last two decades have witnessed the significance of RF provided by the user in improving the performance of image retrieval. Various RF schemes have been widely developed based on different assumptions for the positive and negative feedback samples [4], [58], [59], [11], [12], [60], [13], [6], [7], [8], [9]. However, it is more appropriate for image retrieval to achieve satisfactory results within as few rounds of RF as possible since an online learning task is usually tedious and boring for the user.

Active learning is one of the most powerful tools to reduce the labeling efforts of the user by labeling the most informative samples. Discriminative models, e.g., SVMactive, aim to select the most ambiguous samples and generative models attempt to identify the most representative samples. In general, when the number of the training data is large, the methods which select the most ambiguous samples (e.g., SVMactive) may outperform the methods which select the most representative samples (e.g., GOED). This is mainly because with a large amount of training data, the classifier will be more accurate and thus the ambiguous samples selected by the algorithm can provide the most amount of information. However, for image retrieval, it is not appropriate to require the user to label a large number of samples in RF. Consequently, when the number of the training data is small, GOED is more appropriate in selecting representative samples in the database as the most informative ones for the user to label compared with the popular SVMactive for CIR.

VI. Conclusion

In this paper, we have introduced a novel active learning method, i.e., GOED, to select multiple representative samples in the database as the most informative ones for the user to label in image retrieval. Especially, GOED alleviates the small-sized training data problem by leveraging the geometric structure of unlabeled samples in RKHS, and thus further enhances the performance of image retrieval. Different from the conventional manifold regularization framework, the new method can effectively select the most informative samples for the user to label in image retrieval. By minimizing the expected average prediction variance on the test data, GOED has a clear geometric interpretation to select the most representative samples in the database iteratively with the global optimum, and it is more effective and efficient for the user to label. Compared with the popular SVMactive, our method is label-independent and thus can avoid various potential problems caused by insufficient and inexactly labeled feedback samples, and is more appropriate and useful for image retrieval. Extensive experiments on both synthetic datasets and a real-world image database have shown the advantages of the proposed GOED for CIR.

Despite the promising results, several questions remain to be investigated in our further work. In general, ambiguity and representativeness are two important aspects in considering an active learning problem, both of which can provide useful and necessary information for alleviating the labeling efforts of conventional RF. Therefore, an effective technique is required to integrate the two aspects of active learning for selecting the most informative samples in the database. In addition, theoretic questions need to be investigated regarding how the proposed method affects the generalization error of the classification model. More specifically, we expect that the active learning method can improve the generalization ability of the classifier and further enhance the performance of the system.

ACKNOWLEDGMENT

The authors would like to acknowledge the Ph.D. grant from the Institute for Media Innovation, Nanyang Technological University, Singapore. This work was partially supported by the SINGAPORE MINISTRY OF EDUCATION Academic Research Fund (AcRF) Tier 2, Grant Number: T208B1218.

REFERENCES

[1] A.W.M. 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, Dec. 2000.

[2] Y. Rui, T.S. Huang, and S.F. Chang, "Image retrieval: current techniques, promising directions, and open issues," *Journal of Visual Communication and Image Representation*, vol. 10, no. 1, pp. 39–62, 1999.

[3] R. Datta, D. Joshi, J. Li, and J.Z. Wang, "Image retrieval: ideas, influences, and trends of the new age," *ACM Computing Surveys*, vol. 40, no. 2, pp. 1–60, May 2008.

[4] X.S. Zhou and T.S. Huang, "Relevance feedback in image retrieval: A comprehensive review," *Multimedia Systems*, vol. 8, no. 6, pp. 536–544, Apr. 2003.

[5] Y. Chen, X. S. Zhou, and T.S. Huang, "One-class svm for learning in image retrieval," in *Proceedings of IEEE International Conference on Image Processing*, 2001, pp. 34 –37.

[6] G. Guo, A.K. Jain, W. Ma, and H. Zhang, "Learning similarity measure for natural image retrieval with relevance feedback," *IEEE Transactions on Neural Networks,*, vol. 13, no. 4, pp. 811 – 820, July 2002.

[7] P. Hong, Q. Tian, and T.S. Huang, "Incorporate support vector machines to content-based image retrieval with relevance feedback," in *Proceedings of IEEE International Conference on Image Processing*, 2000, pp. 750 –753.

[8] J. Li, N. Allinson, D. Tao, and X. Li, "Multitraining support vector machine for image retrieval," *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp. 3597 –3601, 2006.

[9] D. Tao, X. Tang, X. Li, and X. Wu, "Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 7, pp. 1088 –1099, 2006.

[10] L. Zhang, L. Wang, and W. Lin, "Semisupervised biased maximum margin analysis for interactive image retrieval," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 2294–2308, 2012.

[11] X. S. Zhou and T.S. Huang, "Small sample learning during multimedia retrieval using biasmap," in *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2001, pp. 11 –17.

[12] D. Tao, X. Tang, X. Li, and Y. Rui, "Direct kernel biased discriminant analysis: a new content-based image retrieval relevance feedback algorithm," *IEEE Transactions on Multimedia*, vol. 8, no. 4, pp. 716 –727, 2006.

[13] D. Xu, S. Yan, D. Tao, S. Lin, and H. Zhang, "Marginal fisher analysis and its variants for human gait recognition and content- based image retrieval," *IEEE Transactions on Image Processing*, vol. 16, no. 11, pp. 2811 –2821, 2007.

[14] L. Zhang, L. Wang, and W. Lin, "Generalized biased discriminant analysis for content-based image retrieval," *IEEE Transactions on Systems, Man, Cybernetics-Part B: Cybernetics*, vol. 42, no. 1, pp. 282–290, Feb. 2012.

[15] C.H. Hoi, M.R. Lyu, and 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.

[16] Y. Liu, D. Xu, I.W. Tsang, and J. Luo, "Textual query of personal photos facilitated by large-scale web data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, pp. 1022–1036, 2011.

[17] S. Tong and E. Chang, "Support vector machine active learning for image retrieval," in *Proceedings of the ninth ACM international conference on Multimedia*, 2001, Multimedia '01, pp. 107–118.

[18] L.P. Wang, *Support Vector Machines: Theory and Applications*, Springer Berlin, 2005.

[19] L.P. Wang and X.J. Fu, *Data Mining with Computational Intelligence*, Springer Berlin, 2005.

[20] L. Wang, K. Chan, and Z. Zhang, "Bootstrapping SVM active learning by incorporating unlabelled images for image retrieval," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2003*.

[21] C. H. Hoi and M. R. Lyu, "A semi-supervised active learning framework for image retrieval," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2005*.

[22] C. Dagli, S. Rajaram, and T. Huang, "Leveraging active learning for relevance feedback using an information theoretic diversity measure," *Image and Video Retrieval*, pp. 123–132, 2006.

[23] C. H. Hoi, R. Jin, J. Zhu, and M. R. Lyu, "Semisupervised svm batch mode active learning with applications to image retrieval," *ACM Transactions on Information System*, vol. 27, no. 3, pp. 16:1–16:29, May 2009.

[24] B. Settles, "Active learning literature survey," Computer Sciences Technical Report 1648, University of Wisconsin Madison, 2009.

[25] D.A. Cohn, Z. Ghahramani, and M.I. Jordan, "Active learning with statistical models," *Journal of Artificial Intelligence Research*, vol. 4, pp. 129–145, 1996.

[26] R. Liere and P. Tadepalli, "Active learning with committees for text categorization," in *Proceedings of the National Conference on Artificial Intelligence*. John Wiley & Sons Ltd, 1997, pp. 591–597.

[27] A. McCallum and K. Nigam, "Employing em and pool-based active learning for text classification," in *Proceedings of the 5th International Conference on Machine Learning*, 1998, ICML '98, pp. 350–358.

[28] G. Schohn and D. Cohn, "Less is more: Active learning with support vector machines," in *Proceedings of the 7th International Conference on Machine Learning*, 2000, ICML '00, pp. 839–846.

[29] A.C.Atkinson and A.N.Donev, *Optimum Experimental Designs*, Oxford,U.K.,Oxford Univ.Press, 2007.

[30] K. Yu, J. Bi, and V. Tresp, "Active learning via transductive experimental design," in *Proceedings of the 23rd International Conference on Machine Learning*, 2006, vol. 23, pp. 1081–1088.

[31] K. Yu, S. Zhu, W. Xu, and Y. Gong, "Non-greedy active learning for text categorization using convex ansductive experimental design," in *Proceedings the 31st International Conference on Research and Development in Information Retrieval*, 2008, pp. 635–642.

[32] O. Chapelle, B. Schölkopf, A. Zien, et al., *Semi-supervised learning*, vol. 2, MIT press Cambridge, MA:, 2006.

[33] X. Zhu, Z. Ghahramani, and J. Lafferty, "Semi-supervised learning using gaussian fields and harmonic functions," in *Proceedings of International Conference on Machine Learning*, 2003, vol. 20, pp. 912–920.

[34] D. Cai, X. He, and J. Han, "Semi-supervised discriminant analysis," in *Proceedings of IEEE International Conference on Computer Vision*, 2007.

[35] K. Bennett and A. Demiriz, "Semi-supervised support vector machines," *Advances in Neural Information processing systems*, pp. 368–374, 1999.

[36] D. Zhou, O. Bousquet, T.N. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," *Advances in neural information processing systems*, vol. 16, pp. 321–328, 2004.

[37] V. Sindhwani, P. Niyogi, and M. Belkin, "Beyond the point cloud: from transductive to semi-supervised learning," in *Proceedings of the 22nd International Conference on Machine Learning*. ACM, 2005, pp. 824–831.

[38] M. Belkin, Partha N., and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *Journal of Machine Learning Research*, vol. 7, pp. 2399–2434, Dec. 2006.

[39] X. He, "Laplacian regularized d-optimal design for active learning and its application to image retrieval," *IEEE Transactions on Image Processing*, vol. 19, no. 1, pp. 254–263, 2010.

[40] L. Zhang, C. Chen, W. Chen, J. Bu, D. Cai, and X. He, "Convex experimental design using manifold structure for image retrieval," in *Proceedings of the 17th ACM international conference on Multimedia*. ACM, 2009, pp. 45–54.

[41] B. Geng, D. Tao, Chao Xu, L. Yang, and X. Hua, "Ensemble manifold regularization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, pp. 1227–1233, 2012.

[42] L. Si, R. Jin, C.H. Hoi, and M.R. Lyu, "Collaborative image retrieval via regularized metric learning," *Multimedia Systems*, vol. 12, pp. 34–44, 2006.

[43] C.H. Hoi, W. Liu, and S.F. Chang, "Semi-supervised distance metric learning for collaborative image retrieval and clustering," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 6, no. 3, pp. 1–26, 2010.

[44] L. Zhang, L. Wang, and W. Lin, "Conjunctive patches subspace learning with side information for colaborative image retrieval," *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3707–3720, 2012.

[45] C.H. Hoi, R. Jin, and M.R. Lyu, "Batch mode active learning with applications to text categorization and image retrieval," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1233–1248, 2009.

[46] X. Tian, D. Tao, X. Hua, and X. Wu, "Active reranking for web image search," *IEEE Transactions on Image Processing*, vol. 19, no. 3, pp. 805–820, Mar. 2010.

[47] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[48] T. Evgeniou, M. Pontil, and T. Poggio, "Regularization networks and support vector machines," *Advances in Computational Mathematics*, vol. 13, no. 1, pp. 1–50, 2000.

[49] V.N. Vapnik, *The nature of statistical learning theory*, Springer Verlag, 2000.

[50] M. Belkin and P. Niyogi, "Using manifold structure for partially labeled classification," *Advances in Neural Information Processing Systems*, vol. 15, pp. 929–936, 2002.

[51] B. Scholkopf and Alexander J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, MIT Press, Cambridge, MA, USA, 2001.

[52] X. He and P. Niyogi, "Locality Preserving Projections," in *Advances in Neural Information Processing Systems*, Cambridge, MA, 2004, MIT Press.

[53] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: A comprehensive study," *International Journal of Computer Vision*, vol. 73, no. 2, pp. 213–238, 2007.

[54] D. Marr and A. Vision, "A computational investigation into the human representation and processing of visual information," *WH San Francisco: Freeman and Company*, 1982.

[55] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[56] D. Jia, A.C. Berg, and F. Li, "Hierarchical semantic indexing for large scale image retrieval," in *Proceedings IEEE International Conference on Computer Vision and Pattern Recognition*, 2011, pp. 785 –792.

[57] J. Yang, K. Yu, and T. S. Huang, "Supervised translation-invariant sparse coding," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2010, pp. 3517–3524.

[58] Y. Rui, T.S. Huang, M. Ortega, and S. Mehrotra, "Relevance feedback: a power tool for interactive content-based image retrieval," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 8, no. 5, pp. 644 –655, Sept. 1998.

[59] Y. Rui and T.S. Huang, "Optimizing learning in image retrieval," in *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2000, pp. 236 –243.

[60] L. Wang, K. L. Chan, and P. Xue, "A criterion for optimizing kernel parameters in kbda for image retrieval," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 35, no. 3, pp. 556 –562, 2005.