Manifold Regularized Experimental Design for Active Learning

Lining Zhang, *Member, IEEE*, Hubert P. H. Shum, *Member, IEEE*, and Ling Shao, *Senior Member, IEEE*

***Abstract*— Various machine learning and data mining tasks in classification require abundant data samples to be labeled for training. Conventional active learning methods aim at labeling the most informative samples for alleviating the labor of the user. Many previous studies in active learning select one sample after another in a greedy manner. However, this is not very effective, because the classification models have to be retrained for each newly labeled sample. Moreover, many popular active learning approaches utilize the most uncertain samples by leveraging the classification hyperplane of the classifier, which is not appropriate, since the classification hyperplane is inaccurate when the training data are small-sized. The problem of insufficient training data in real-world systems limits the potential applications of these approaches. This paper presents a novel method of active learning called manifold regularized experimental design (MRED), which can label multiple informative samples at one time for training. In addition, MRED gives an explicit geometric explanation for the selected samples to be labeled by the user. Different from existing active learning methods, our method avoids the intrinsic problems caused by insufficiently labeled samples in real-world applications. Various experiments on synthetic data sets, such as the Yale face database and the Corel image database, have been carried out to show how MRED outperforms existing methods.**

***Index Terms*— Machine learning, active learning, manifold regularization, face recognition, content-based image retrieval.**

I. Introduction

IN MANY real-world systems [1]–[7], the effort of labeling samples is usually hard, even when a large number of unlabeled data samples are readily available and provide very useful information for the systems. Semi-supervised learning is widely designed to significantly enhance the general performance of conventional supervised learning by using abundant unlabeled samples [8]–[10]. Transfer learning borrows the knowledge from related domains to greatly improve the performance of the systems that have insufficient training samples [11]–[14]. Active learning alleviates the labor of the user in a different way by selecting the informative samples to label [15]–[18]. Thus, instead of passively receiving the label information, the system can actively decide which unlabeled samples are the most informative ones and then obtain label information from the user. In this way, the system achieves the high classification performance while using as few training samples as possible.

For active learning, the main challenge is finding an effective scheme to evaluate the informativeness and usefulness of the unlabeled samples in the database. A popular scheme for active learning methods is the uncertain criterion. The systems with uncertain criterion actively select those samples whose predicted label information is the most ambiguous based on the current trained model [17], [19]–[22]. Support Vector Machine based active learning (SVMactive) is one of the most effective active learning methods in this category, which is designed to find the uncertain samples with the help of the classification hyperplane of the corresponding Support Vector Machine (SVM) [17], [19]. During the past decade, numerous research works have been conducted to improve the performance of SVMactive for real-world applications [20]–[22]. However, the trained classification hyperplane of the classifier is not usually stable when training data are insufficient [23], [24]. In many real-world systems, the user does not label abundant data samples. Moreover, these data samples cannot be labeled very accurately [3], [25]. Therefore, the classification hyperplane of the classifier is not reliable for selecting the most informative samples with small-sized labeled training data. Another problem is that since these methods require a classification hyperplane to find the samples with the most information, SVMactive can not be utilized when labeled samples are not available.

To illustrate the principle of SVMactive, a simple synthetic dataset is shown in Fig.1. Here, we have two labeled data samples (i.e., the big solid circle for the sample with the positive label while the big hollow circle for the sample with the negative label) and a few unlabeled data samples (i.e., the small solid dots). The labeled data samples and unlabeled samples are used to illustrate the training and testing data, respectively. All six samples distribute along a line. Most of the previous studies in active learning (i.e., SVMactive) select the uncertain samples (i.e., A and B) one after another with a greedy strategy but cannot select a group of representative samples (i.e., B, C and D) in the database simultaneously. Moreover, these methods require a classification hyperplane to identify the uncertain data samples and thus cannot be used when there are no labeled data.

Manuscript received May 10, 2016; revised September 30, 2016; accepted November 22, 2016. Date of publication December 1, 2016; date of current version January 5, 2017. This work was supported by the Engineering and Physical Sciences Research Council under Grant EP/M002632/1. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Christopher Wyatt. *(Corresponding author: Ling Shao).*

L. Zhang and H. P. H. Shum are with the Department of Computer and Information Sciences, Northumbria University, Newcastle Upon Tyne NE1 8ST, U.K. (e-mail: liningzh@gmail.com; hubert.shum@northumbria.ac.uk).

L. Shao is with the School of Computing Sciences, University of East Anglia, Norwich NR4 7TJ, U.K. (e-mail: ling.shao@ieee.org).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2016.2635440

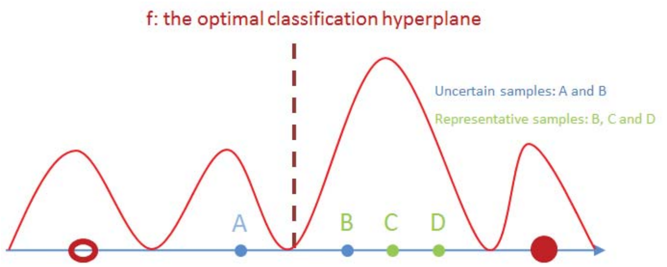

f: the optimal classification hyperplane
Uncertain samples: A and B
Representative samples: B, C and D
A
B
C
D

Fig. 1. An illustration of the most informative samples (i.e., uncertain samples and representative samples). Previous work in active learning studies find the uncertain data samples (i.e., A and B) one after another one greedily using the optimal classification hyperplane of the classifier f. Our method simultaneously selects a set of representative samples (i.e., B, C and D) in the database iteratively without using the classification hyperplane of f.

Active learning is often considered as experimental design in the machine learning community [26]. Optimum experimental design (OED) studies the selection of the most informative samples in the database to measure since conducting an experiment is usually expensive [26]. Conventional OED contains three different OED methods, which maximize the confidence of the predictive model when the measurement of estimated parameter covariance matrixes is minimized [26]. However, conventional OED methods do not show the informativeness of predictions on testing samples if the testing samples are presented firstly. Transductive experimental design (TED) [27], [28] was proposed to directly evaluate the predictions on testing samples, and to give an explicit geometric explanation to the selected samples for training. TED has obtained impressive performance compared with conventional OED approaches. Conventional OED methods only assess the labeled samples but ignore the unlabeled samples in the database, although these unlabeled samples provide useful information. A large number of semi-supervised learning approaches have been designed to improve the general performance of supervised learning models by using the manifold of unlabeled samples [9], [10], [29]–[34]. Moreover, most of the conventional OED methods select one data sample after another one [16], [35], which limit their potential applications to various real-world systems [2], [4], [36], [37].

To address the intrinsic drawbacks in OED, this paper presents an effective method for active learning called manifold regularized experimental design (MRED) by using the intrinsic manifold of the massive unlabeled samples. The new method allows us to simultaneously select a group of the most informative samples for training a classifier. Our method is largely inspired by the recent manifold assumption [10], [38], which plays an important role in semi-supervised learning models to significantly improve the generalization ability of conventional supervised learning in the machine learning community. Different from the previous methods based on the conventional manifold regularization [10] where the training samples are pre-given by the system, our method selects representative samples in the database for training. Moreover, this method learns a data-dependent deformed kernel function by using both a small number of labeled samples and abundant unlabeled data samples. These samples construct a data-dependent kernel function warped by a data-dependent norm to integrate the intrinsic manifold of unlabeled samples. A set of the most representative samples can be labeled by the user when the average prediction variance is minimized by using the deformed kernel function. Different from the TED methods [27], [28], MRED effectively utilizes the unlabeled samples in the new data-dependent deformed kernel space. Moreover, our method does not depend on any label information of training samples. At the same time, the sensitivity problem caused by insufficiently labeled samples is effectively alleviated. Various experiments on synthetic datasets, the Yale face database and the Corel image database have shown the general performance of the proposed MRED for real-world applications.

The main contributions of this work include the following:

- This paper has presented a novel method for active learning called MRED to simultaneously select a group of representative samples for training a classifier. Different from the conventional manifold regularization methods where the training samples are pre-given, our method can find the most representative samples to label.
- We intend to select the most informative samples with the global optimum, which are the most representative samples in the database.
- We use the deformed kernel function to identify multiple representative samples iteratively. Different from the previous SVMactive methods, our method does not require any label information and avoids the sensitivity problem caused by insufficiently labeled samples in SVMactive.

The rest of the paper is organized as follows: In Section 2, we provide a brief review of the related work, i.e., active learning, OED and TED. Then, we introduce the MRED method in Section 3. Section 4 presents the experimental results. Finally, we give the conclusions and suggestions for future work in Section 5.

II. Related Work

In this section, we give an overview of the conventional problem for active learning in the machine learning community and provide a review of OED and TED.

A. Active Learning

In the machine learning community, active learning is useful in labeling a small number of informative samples for obtaining sufficient information. In general, most of the active learning methods aim at selecting uncertain samples or representative samples for the user to label. Uncertain samples are defined as the most ambiguous unlabeled samples based on the current trained model. Representative samples effectively represent the intrinsic structure of unlabeled samples. SVMactive is a very effective technique to select the uncertain samples, which was very popular during the past few years [17], [19]–[22]. Hoi *et al.* [39] presented a method for active learning based on the batch model framework to find a set of the most informative samples simultaneously, which is fundamentally based on the kernel logistic regression model. To alleviate the problem of small sample size in relevance feedback (RF), Zhang *et al.* [25] proposed a general active learning framework by using the intrinsic manifold of the data to find the most informative samples in the database as the training samples. However, this method is specifically designed for a conventional binary classification problem, i.e., RF in collaborative image retrieval (CIR). Yang *et al.* [40] introduced a batch model multi-class active learning method for a visual concept recognition task, which can alleviate the problem of uncertainty sampling with a small seed set size to evaluate the uncertainty of data samples. Long and Gang [41] proposed a novel multi-annotator Gaussian process model to deal with multi-class visual recognition in the collaborative active learning framework with multiple annotators. Despite the vast research work in the past few years, conventional active learning approaches need an initial optimal classification hyperplane to find the useful samples. To incorporate the geometrical structure of the data space, Cai and He [42] proposed a Manifold Adaptive Experimental Design (MAED) method by introducing a data-dependent norm to integrate the unlabeled samples on reproducing kernel Hilbert space (RKHS) for text categorization. However, this method cannot show an explicit relationship between conventional active learning methods and semisupervised learning models, which is very important in handling the problems associated with small-sized training data. To alleviate the labor in defining multiple attributes in the large amount of data, You *et al.* [43] introduced a diverse expected gradient active learning method by combining an informativeness analysis and a diversity analysis for relative attributes.

B. OED

Active learning is formalized as follows. Suppose that we have a large number of unlabeled samples X in the high-dimensional space R^h, where h is the dimensionality of the high-dimensional space, the algorithm finds a subset of samples $Z \subseteq X$, which usually contains multiple informative samples for training. That is, if these samples $z_i (i = 1, \ldots, l)$ are labeled by the user and utilized as training samples, we can effectively obtain the label information of the unlabeled samples by using the auxiliary information.

We consider a linear regression problem as follows:

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

where y is the real-valued output, $w \in R^h$ is the weight parameters, $x \in R^h$ is the variable and ε is the measurement noise with zero mean and σ^2 variance. OED aims to find abundant samples with the most information $z_1, z_2, \ldots, z_l$ from X to learn a prediction function $f(x) = w^T x$ by minimizing the expected prediction. Given a large number of informative samples Z and the label information $Y = \{y_1, y_2, \ldots, y_l\}$, the prediction function f is estimated with the minimization of the objective function as follows: [44]:

$$J(w) = \sum_{i=1}^{l} (w^T z_i - y_i)^2. \tag{2}$$

The optimal solution to this problem is given by [44]:

$$\hat{w} = (Z^T Z)^{-1} Z^T y, \tag{3}$$

where $Z = [z_1, z_2, \ldots, z_l]$ is the feature matrix and $y = [y_1, y_2, \ldots, y_l]^T$ is the label information. It is verified that $\hat{w}$ is an unbiased estimation of w and the covariance matrix is shown as [27]:

$$Cov(\hat{w}) = \sigma^2 (Z^T Z)^{-1}. \tag{4}$$

OED only selects multiple samples with the most information in the database by minimizing various measurements of the estimated parameter covariance, i.e., Eq.(4). Three typical criteria are the trace of C_w, the determinant of C_w and the maximum eignevalue of C_w [26].

C. TED

Conventional OED approaches do not give a very clear geometric interpretation for these selected informative samples. TED tends to select multiple representative samples when the expected variance on the testing samples is minimized. Then, a set of representative samples in the database are selected as the most informative ones by directly minimizing the expected prediction variance on the test samples. By considering the regularized least squares formulation, TED is formulated as follows:

$$\hat{w} = \arg\min_{w} \sum_{i=1}^{k} (y_i - f(x_{s_i}))^2 + \gamma \, \|w\|^2, \tag{5}$$

where $\gamma \geq 0$ is the parameter to balance the loss function and the regularization term. It is verified that the solution to this problem is given as follows [44]:

$$\hat{w} = (Z^T Z + \gamma I)^{-1} Z^T y, \tag{6}$$

where I is an identity matrix to enhance the stability of the solution. The average prediction covariance matrix of the test samples X is given by

$$Cov(w) \approx \sigma^2 (Z^T Z + \gamma I)^{-1}. \tag{7}$$

TED tends to select the samples by minimizing the expected predictive variance on the given test samples. Let $X = [x_1, \ldots, x_n]^T$, the average predictive variance of TED for training samples is shown as follows:

$$\begin{aligned} &\frac{1}{n} \sum_{i=1}^{n} x_i^T Cov(\hat{w}) x_i \\ &\approx \frac{\sigma^2}{n} \sum_{i=1}^{n} x_i^T (Z^T Z + \gamma I)^{-1} x_i \\ &= \frac{\sigma^2}{n} Tr\left(X \left(Z^T Z + \gamma I \right)^{-1} X^T \right) \end{aligned} \tag{8}$$

Then, TED is formulated as the following optimization problem:

$$\min Tr\left(X \left(Z^T Z + \gamma I \right)^{-1} X^T \right). \tag{9}$$

It is verified that the optimization problem cannot be solved effectively. After some derivations, this problem can be

TABLE I
IMPORTANT NOTATIONS AND VARIABLES IN THIS PAPER

Notations and Variables	Descriptions
$X = \{x_1, \ldots, x_n\} \in R^h$	A set of unlabeled data samples in the high-dimensional space R^h
$Z = \{z_1, \ldots, z_l\} \subset X$	A subset of the most informative samples in X
$\alpha_i = (\alpha_{i,1}, \ldots, \alpha_{i,n})^T$	The reconstruction coefficient
$\beta = (\beta_1, \ldots, \beta_n)^T$	The most informative samples' selection coefficient
$y = f(x)$	A binary classifier used to predict the relationship from a sample x to its label y
y	The label information
ω	The weight vector
γ_1	The coefficient to trade off the loss function term and the regularization term $\|\omega\|^2$
γ_2	The coefficient to trade off the loss function term and the regularization term $\|w\|_I^2$
H_K	The original RKHS
$\tilde{H}_{\tilde{K}}$	The deformed RKHS
L	The graph Laplacian matrix for training samples
W	The data adjacency graph matrix for training samples
D	The diagonal matrix of the graph Laplacian
ϕ	The mapping function of the reproducing kernel Hilbert space

given as [27]:

$$\min_{\alpha_i \in R^l, Z} \sum_{i=1}^{n} \|x_i - Z\alpha_i\|^2 + \gamma_1 \|\alpha_i\|^2. \tag{10}$$

The term $\sum_{i=1}^{n} \|x_i - Z\alpha_i\|^2$ in Eq. (10) illustrates that the data samples selected by TED can reconstruct the abundant unlabeled samples in the database. In other words, the selected samples with the most information $z_i (i = 1, \ldots, l)$ reconstruct the data x_i precisely. The second term $\|\alpha_i\|^2$ shows that the TED penalizes the norm of the original reconstruction coefficients, and thus it effectively selects the samples with large norm.

III. MRED FOR ACTIVE LEARNING

In this section, we introduce a new method for active learning called MRED, which effectively finds multiple informative samples iteratively in the database for training. Compared with the popular SVMactive, our method avoids the problems caused by insufficiently labeled samples and generates more effective solutions for various real-world applications. Some important notations are summarized in Table I.

A. Active Learning Problem

Suppose that we have a binary classification problem, a classification model is usually learned to predict the relationship between the sample x and its label $y \in \{-1, 1\}$ via

$$\mathrm{y} = \mathrm{sign}\left(f(\mathrm{x})\right), \tag{11}$$

where the classifier is simply formulated as $f(x) = w^T x$. The bias term can be integrated into this formulation by replacing the weights and feature vector as in [28]. Given a set of labeled samples $z_1, \ldots, z_l$, the least-squares SVM (LSSVM) is equivalent to the least-squares ridge regression (LSRR) [45], which learns $f(x)$ by estimating w via

$$w^* = \arg\min_{w} \left(J(w) = \sum_{i=1}^{l} \left(w^T z_i - y_i\right)^2 + \gamma_1 \|w\|^2 \right), \tag{12}$$

where γ_1 is a trade-off parameter to balance the loss function and the regularization term, and $\gamma_1 > 0$. In general, the active learning problem is defined as the following. Given multiple unlabeled samples $X = \{x_1, \ldots, x_n\}$ in the high-dimensional space R^h, we want to find a subset of samples $Z = \{z_1, \ldots, z_l\}$ that contains a set of samples with the most information to be labeled. In general, these samples can significantly enhance the performance of the system if they are labeled by the user and adopted as the training samples.

B. MRED for Active Learning

In this subsection, we present a new method for active learning by using the intrinsic manifold of a large number of samples in the database to select the most informative samples to label. The proposed method is largely motivated by the recent research on manifold regularization [10], [38], which plays an important role in improving the generalization performance of supervised learning for semi-supervised learning models, i.e,

$$\omega^* = \arg\min_{\omega} \begin{pmatrix} J(w) = \sum_{i=1}^{l} (\omega^T z_i - y_i)^2 + \gamma_1 \|\omega\|^2 \\ +\gamma_2 \|\omega\|_I^2 \end{pmatrix} \tag{13}$$

where the $\|w\|_I^2$ is a smooth regularization penalty term to incorporate the intrinsic manifold of the abundant unlabeled samples. Parameters γ_1 and γ_2 are used to trade off the loss function $\sum_{i=1}^{l} \left(w^T z_i - y_i\right)^2$, $\|\omega\|^2$ and $\|\omega\|_I^2$. The term $\|w\|_I^2$

plays an important role in various semi-supervised learning studies. It is usually used to model the output smoothness of the classifier along the intrinsic manifold estimated from both a small number of labeled samples and a large number of unlabeled samples in the database [10], [38].

Our new method is similar with that of the Laplacian regularized LSRR in [10]. However, the informative samples in the database to be labeled can be effectively selected by MRED. Different from the popular active learning methods in the machine learning community, the new method can alleviate the labor of the user by using the intrinsic manifold structure of a large number of unlabeled samples [10], [29]–[31], [38], [46], [47]. In many real-world applications, the system effectively finds a set of the most informative samples to label, which is actually an active learning problem. After that, when these informative samples are labeled by the user, our system utilizes all of the data samples including both a small number of labeled samples and a large number of unlabeled samples to learn a classification model, which is actually a semi-supervised learning problem.

To integrate the intrinsic geometric structure of abundant unlabeled samples, many methods have been proposed in the literature [47]–[49]. In this work, we first design an effective RKHS deformed by a kernel Gram matrix K, and then find a solution to solve this problem by selecting a set of the most informative samples to label. In the following paragraphs, we first discuss how to integrate the intrinsic manifold of abundant unlabeled samples into the kernel space. Then, we also discuss how to find the informative samples $z_i (i = 1, \ldots, l)$ to label.

Kernel methods are useful techniques in discovering the intrinsic nonlinear manifold structure of the samples by embedding the original data into a higher dimensional kernel space [50]. Although the kernel methods can capture the intrinsic manifold of the database, the intrinsic nonlinear manifold captured by the kernel function may not be consistent with the intrinsic manifold structure of the data [10], [47]. In this paper, we employ a data-dependent deformed kernel function to incorporate the manifold structure of abundant unlabeled samples, which is constructed by a conventional kernel function from all the samples including a small number of labeled data samples and a large number of unlabeled samples with an effective kernel deformation principle [47].

We use H_K and $\tilde{H}_{\tilde{K}}$ to denote the original RKHS and the new kernel space, respectively. Reference [47] assumes the relationship between these two kinds of kernel spaces as follows:

$$\tilde{k}(x_i, x_j) = k(x_i, x_j) - \gamma k_{x_i}^T (I + MK)^{-1} M k_{x_j}, \tag{14}$$

where $k(\cdot, \cdot)$ is the conventional kernel function on both the labeled and unlabeled samples defined in H_K with its associated kernel Gram matrix $K = [k(x_i, x_j)]_{n \times n}$, k_{x_i} is defined as $k_{x_i} = [k(x_i, x_1), \ldots, (x_i, x_n)]^T$. It is important to note that all popular kernels (i.e., Gaussian kernel, polynomial kernel and linear kernel) can be transformed to the new kernel space. The second term in Eq. (14) is the deformed regularization term given by a data-dependent norm and is designed to incorporate the intrinsic manifold of the data. γ is a deformation parameter to balance the loss of the kernel function and the deformation term, and I is used to enhance the stability of the solution. The key problem here is how to choose M, which is designed to integrate the manifold of the samples in the database X.

As suggested by [10] and [47], we adopt the graph Laplacian L to capture the intrinsic manifold of unlabeled samples. In general, the graph Laplacian L is defined as $L = D - W$. The matrix $W \in R^n \times R^n$ is the data adjacency graph, and each element W_{ij} is an edge weight between two corresponding samples x_i and x_j. In the matrix $D \in R^n \times R^n$, the i^{th} entry $D_{ii} = \sum_{j=1}^n W_{ij}$. Different extensions of W were introduced in [51]. Here, we give a typical one as follows:

$$\mathrm{W_{ij}} = \begin{cases} 1, & \text{if } \mathrm{x_i} \in (\mathrm{x_j}) \text{ or } \mathrm{x_j} \in (\mathrm{x_i}) \\ 0, & \text{otherwise,} \end{cases} \tag{15}$$

where $N(x_i)$ is used to denote the k neighboring samples of the given sample x_i. The graph Laplacian term smooths the output of the classification model as follows:

$$f^T L f = \sum_{i=1}^{n} (f(x_i) - f(x_j))^2 W_{ij}. \tag{16}$$

As shown in [51], the definition in Eq. (16) corresponds to the approximation of a manifold on which the data samples X may reside. Motivated by [47], by setting $M = L$, $\tilde{K}$ can be used to design different algorithms for semi-supervised classification, and the new kernel can reinterpret them within the supervised learning models. In this paper, we formulate it as a new active learning method with a semi-supervised learning model for supervised learning in the new kernel space, that is,

$$\hat{w}^* = \arg\min_{\hat{w} \in \tilde{H}_{\tilde{K}}} \left\{ J(\hat{w}) = \sum_{i=1}^{l} \left(\hat{w}^T \tilde{\phi}(z_i) - y_i\right)^2 + \gamma_1 \left\| \hat{w} \right\|^2 \right\}, \tag{17}$$

where $\tilde{\phi}(z_i)$ indicates the data sample z_i in the high dimensional kernel space $\tilde{H}_K$, which shows the intrinsic manifold structure of a large number of the unlabeled samples in the database. Motivated by the theorem in representation learning [52], we notice that $\hat{w}^*$ is defined as a linear combination of $\tilde{\phi}(z_i), i = 1, \ldots, l$:

$$\hat{w} = \sum_{i=1}^{l} \upsilon_i \tilde{\phi}(z_i) = \tilde{\phi}(Z)\upsilon, \tag{18}$$

where $\upsilon = [\upsilon_1, \ldots, \upsilon_l]^T \in R^l$ is the expansion coefficient. By bringing Eq. (18) into Eq. (17), we have

$$\hat{w}^* = \arg\min_{\hat{w} \in \tilde{H}_K} \{ J(\upsilon) = \left\| \tilde{K}_z \upsilon - y \right\|^2 + \gamma_1 \upsilon^T \tilde{K}_z \upsilon \}, \tag{19}$$

where $y = [y_1, \ldots, y_l]^T$ is the label of the training samples, and $\tilde{K}_Z \in R^{l \times l}$ is constructed by the labeled set $\tilde{\phi}(Z) = [\tilde{\phi}(z_1), \ldots, \tilde{\phi}(z_l)]$ with the entries calculated as in the new kernel Gram matrix $\tilde{K}$. By setting $\frac{\partial J(\upsilon)}{\partial \upsilon} = 0$, we solve the problem of Eq. (19) as follows:

$$\upsilon^* = (\tilde{K}_z + \gamma_1 I)^{-1} y. \tag{20}$$

Generally, given an input data sample x, we obtain the label information of this sample in the following:

$$f(x) = \sum_{i=1}^{l} \tilde{k}(x, z_i)v^*, \tag{21}$$

where $\tilde{k}(\cdot,\cdot)$ is the new data-dependent kernel given in Eq. (14). Thus, Eq. (21) will be considered as the classification result for the sample x.

C. MRED Solution

To find multiple informative samples $\tilde{\phi}(z_i), i = 1, \ldots, l$ in the database for training, we first interpret the active learning method using the conventional supervised learning models in $\tilde{H}_{\tilde{K}}$, i.e., Eq. (19). Motivated by TED [27], we find the informative samples by minimizing the expected prediction variance on the test data. Similar to Eq. (10), the new optimization problem can be reformulated to find the optimal solution as follows:

$$\min_{\alpha_i \in R^l} \sum_{i=1}^{n} \left\| \tilde{\phi}(x_i) - \tilde{\phi}(Z)\alpha_i \right\|^2 + \gamma_1 \|\alpha_i\|^2. \tag{22}$$

Consequently, similar to TED, the data samples selected by MRED reconstruct the abundant unlabeled samples in the database. In other words, MRED tends to select a set of representative samples $\tilde{\phi}(Z)$ that can be used to span a linear space to retain most of the information of $\tilde{\phi}(X)$ in $\tilde{H}_{\tilde{K}}$. The new method gives an explicit geometric explanation to the selected samples $\tilde{\phi}(Z)$ as TED. MRED effectively integrates the geometric information of abundant unlabeled samples in the database by using the deformed kernel space [47]. Moreover, different from previous SVMactive methods, MRED does not require any label information $y_i (i = 1, \ldots, l)$, but only depends on the training samples $\tilde{\phi}(Z) = [\tilde{\phi}(z_1), \ldots, \tilde{\phi}(z_l)]$, which effectively alleviates the different potential problems led by insufficiently labeled samples in real-world applications.

Motivated by [28], by introducing the auxiliary variables $\beta = (\beta_1, \ldots, \beta_n)$ as the selection coefficients of the data samples, Eq. (22) is reformulated as:

$$\min_{\alpha_i, \beta \in R^n} \sum_{i=1}^{n} \left(\left\| \tilde{\phi}(x_i) - \tilde{\phi}(X)\alpha_i \right\|^2 + \sum_{j=1}^{n} \frac{\alpha_{i,j}^2}{\beta_j} \right) + \lambda \|\beta\|_1$$
$$s.t. \beta_j \geq 0, \quad j = 1, \ldots, n, \tag{23}$$

where $\alpha_i = (\alpha_{i,1}, \ldots, \alpha_{i,n})^T$ is the samples selection coefficient. As shown in [53], the $\|\beta\|_1$ results in a sparse coefficient β. When $\beta_j = 0$, all $\alpha_{i,j}, \ldots, \alpha_{n,j}$ should be 0. Otherwise, the objective function goes to infinity, which means the j^{th} sample is not identified as the most representative ones. In the objective function, both the first term (i.e., the square loss function) and the third term (i.e., the l_1 norm) are convex. A summation of convex functions is still convex. As demonstrated in [28], the Hessian of the second term is positive semidefinite and we can know that the second term is also convex. Finally, we get the global optimal solution since the objective function of Eq. (23) is convex. In the following parts, we discuss how to solve this problem step by step.

We define D_β as a diagonal matrix with entries $\beta_1, \ldots, \beta_n$, and thus,

$$\sum_{j=1}^{n} \frac{\alpha_{i,j}^2}{\beta_j} = \alpha_i^T D_\beta^{-1} \alpha_i. \tag{24}$$

To solve this problem, we take the derivative of Eq. (23) with α_{i}. By requiring this derivative to be zero, we get

$$-2\tilde{\phi}(X)^T \tilde{\phi}(x_i) + 2\tilde{\phi}(X)^T \tilde{\phi}(X)\alpha_i + 2D_\beta{}^{-1}\alpha_i = 0. \tag{25}$$

Finally, we have

$$\alpha_i = \left(D_\beta^{-1} + \tilde{\phi}(X)^T \tilde{\phi}(X) \right)^{-1} \tilde{\phi}(X)^T \tilde{\phi}(x_i). \tag{26}$$

In view of $\tilde{\phi}(X)^T \tilde{\phi}(X) = \tilde{K}$, Eq. (26) is reformulated as

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

Then, by taking the derivative of Eq. (23) with β_j and requiring this derivative be zero, we get

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

At last, we obtain the most informative samples selection coefficient as follows:

$$\beta_j = \sqrt{\sum\nolimits_{i=1}^{n} \alpha_{i,j}^2 / \lambda}, \tag{29}$$

where α_i and β_j are calculated iteratively according to Eq. (27) and Eq. (29). Because the objective function of Eq. (23) is convex, we get the global optimum iteratively.

The samples can be ranked by following the selection coefficient β. The top l samples are considered as the informative samples Z in the database. These selected samples are regarded as the most informative samples, which are utilized to train a classifier f according to Eq. (19) and Eq. (21.Finally the classifier are used to do the classification.

As shown in Algorithm 1, In Step 1, the computational complexity of constructing the k nearest neighbor graph is $O(kn^2)$, where n is the number of unlabeled samples. In Step 2, the computational complexity of computing the conventional kernel Gram matrix is $O(n^2)$. In Step 3, the computational complexity of computing the data-dependent kernel Gram matrix $\tilde{K}$ is $O(n^3)$, and it is $O(tn^3)$ in Step 4, where t is the iteration number. Since t is usually a small number, MRED converges very quickly. Therefore, the overall computational complexity of MRED is $O(n^3)$.

IV. Experimental Results

In this section, we compare the proposed method with state-of-the-art active learning methods. We evaluate the effectiveness of the new method based on synthetic datasets, the Yale face database and the Corel image database.

A. Synthetic Datasets

To show the performance of MRED in finding the most informative samples, we compare the proposed MRED with

```
Algorithm 1 MRED for Active Learning
Input: The $n$ unlabeled data samples $X$, the number of the selected most information data samples $l$, the number of the nearest neighbor data samples $k$
Step 1: Construct a nearest neighbor Laplacian graph with the weight matrix $W$ as calculated in Eq. (15) on the unlabeled samples $X$ and calculate
Step 2: Construct the kernel Gram matrix $K$ with an selected input kernel type and let $M = L$.
Step 3: Construct the data-dependent deformed kernel Gram matrix $\tilde{K}$ according to Eq. (14).
Step 4: Let $u_i$ be the $ith$ column vector of $K$ and initialize $\alpha_{i,j} = 1$.
    Step 4.1: Repeat
        Step 4.2: Compute $\beta_j$ according to Eq. (29), i.e., $\beta_j = \sqrt{\sum_{i=1}^{n} \alpha_{i,j}^2/\lambda}$.
        Step 4.3: Compute $\alpha_{\mathrm{i}}$ according to Eq. (27), i.e., $\alpha_{\mathrm{i}} = (D_\beta^{-1} + \tilde{K})^{-1}\tilde{K}_i$.
    Step 4.4: Until Convergence
Step 5: Rank the samples in $X$ by following $\beta_j (j = 1, \ldots, n)$ in a descending order and then return the top $l$ samples as the selected most informative ones $Z$.
Output The $l$ selected most informative samples can be labeled as the training samples.
```

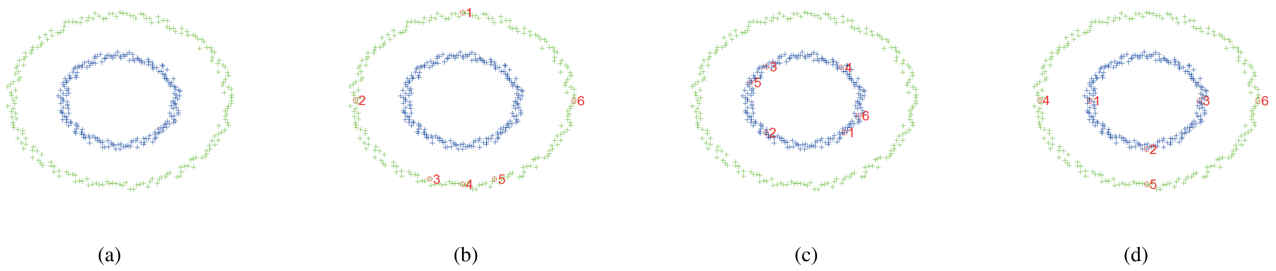


Fig. 2. Performance of several active learning methods in finding the most informative samples. The red circles represent the most informative samples found by A-OED, TED, and MRED in synthetic datasets. The numbers near the selected samples indicate the orders how they were selected. (a) the 2-circle synthetic dataset; (b) A-OED can select 6 informative samples on the large-sized green circle; (c) TED can find 6 informative samples on the small-sized blue circle; (d) MRED can find 3 informative samples on the small-sized blue circle and 3 informative samples on the large-sized green circle, respectively.

two related active learning approaches, i.e., A-OED and TED. It should be noted that SVMactive cannot be applied to this task since labeled samples are usually insufficient. The results are illustrated in Fig. 2. The most informative samples selected by each method are marked with red circles. The numbers near the selected informative samples denote the order of selection. As shown in Fig. 2, A-OED and TED select samples from the small-sized blue circle and the large-sized blue circle, respectively. Three data samples on the small-sized blue circle and three samples on the large-sized green circle are selected by MRED. Inspired by Eq. (22), we notice that the data samples selected by MRED reconstruct the unlabeled samples in the database with the minimum prediction variance, and thus these samples are the most representative ones. As shown in Fig. 2, MRED selects the informative samples, which show much better performance in representing the distribution of the original dataset (i.e., the small-sized blue circle and the large-sized green circle).

B. Real-World Databases

In this subsection, we conduct real-world experiments on two real-world databases to show the performance of different active learning methods.

1) Face Recognition: In this subsection, we first use the samples found by these methods as training samples to train a classifier. Then the unselected samples are adopted as the testing samples. In this experiment, we use the one-versus-all scheme to deal with the multi-class classification problem. If there are c classes in the training samples, we train c different two-class classifiers and each two-class classifier separates one class from all different classes. These c classifiers are used to classify this testing sample, and its label is given based on the largest output value from the c classifiers. SVM [45] and Laplacian regularized LSRR [10] are used as the classifiers to evaluate the effectiveness of different active learning methods.

The Yale face database [54] is used to evaluate the effectiveness of compared methods for face recognition. This database includes 165 grayscale images of 15 different individuals, with 11 images per person. In our implementation, each face image is normalized by fixing the position of 2 eyes and scaled to the size of 32 × 32 pixels. Thus, each face image is represented in a 1,024-dimensional feature space. Fig. 3 illustrates some face images from the Yale face database.

To evaluate the effectiveness of different active learning methods, 20 subsets are randomly generated from the original database. For each subset, 10 images are randomly chosen from each class to form the subset. Therefore, 150 images exist in each subset, and each method is applied to select a given number $k = 5, 10, \ldots, 50$ of training faces. In the experiment, average precision (AP) and standard deviation (SD) are adopted to evaluate the effectiveness of the compared methods.

Figs. 4 (a) and (b) show the APs of different active learning methods versus the number of training data by using the SVM and Laplacian regularized LSRR, respectively. As we can

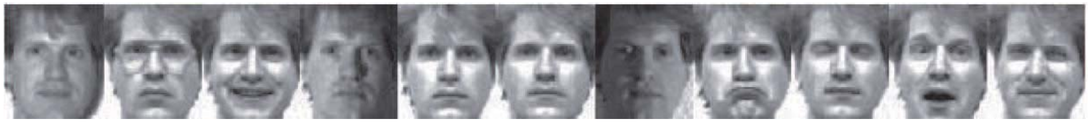
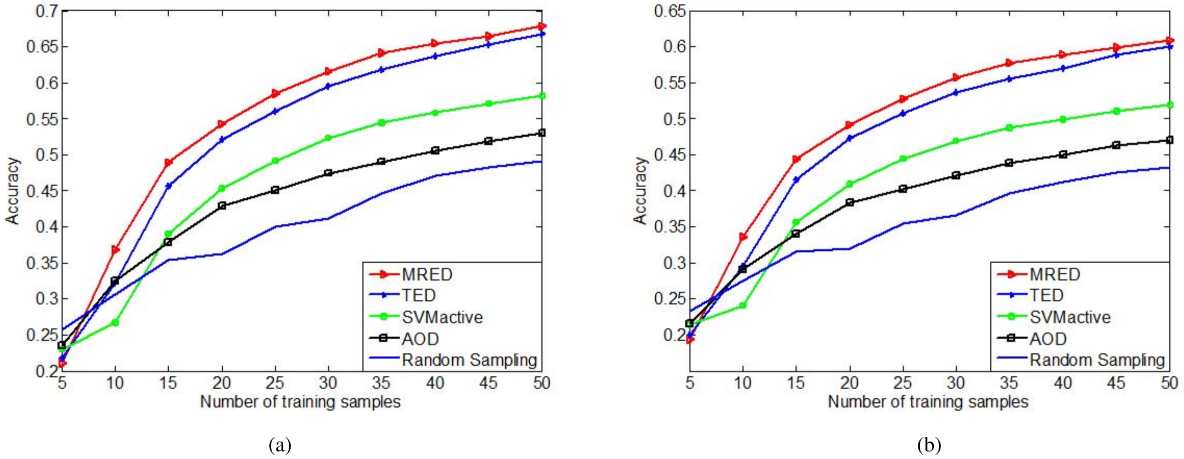


Fig. 3. Eleven images of one person in the Yale face database and the images are aligned well by fixing the positions of the two eyes.



(a)

(b)

Fig. 4. Performance comparison of different active learning approaches (i.e., MRED, TED, SVMactive, A-OED and Random Sampling) on the Yale face database. The face images found by the active learning algorithms are adopted as the training samples and these unselected images are used as the testing samples.

TABLE II

PERFORMANCE COMPARISON OF DIFFERENT ACTIVE LEARNING ALGORITHMS (i.e., MRED, TED, SVMactive, A-OED AND RANDOM SAMPLING) ON YALE FACE DATABASE (APs ± SDs(PERCENT))

k	The classification accuracy by SVM					The classification accuracy by Laplacian regularized LSRR				
	Random	AOD	SVMactive	TED	MRED	Random	AOD	SVMactive	TED	MRED
5	25.65 ± 2.2	23.52 ± 2.0	23.01 ± 3.2	21.77 ± 3.1	21.15 ± 2.1	23.31 ± 2.3	21.67 ± 2.1	21.41 ± 2.3	20.05 ± 2.4	19.45 ± 2.1
10	30.58 ± 3.1	32.42 ± 3.2	26.72 ± 4.3	32.27 ± 1.3	36.84 ± 2.3	27.50 ± 3.2	29.03 ± 2.4	24.09 ± 4.3	29.54 ± 2.3	33.63 ± 2.4
15	35.48 ± 2.8	37.90 ± 3.1	39.04 ± 2.3	45.65 ± 3.5	48.87 ± 4.5	31.57 ± 2.8	33.93 ± 3.4	35.63 ± 3.2	41.49 ± 3.5	44.36 ± 3.6
20	36.14 ± 3.2	42.83 ± 4.5	45.37 ± 1.4	52.11 ± 4.3	54.23 ± 3.3	31.98 ± 1.3	38.29 ± 2.8	40.96 ± 1.5	47.37 ± 2.8	49.12 ± 2.5
25	40.02 ± 2.1	45.12 ± 2.7	49.15 ± 4.3	56.05 ± 2.3	58.53 ± 1.6	35.46 ± 3.2	40.23 ± 3.7	44.42 ± 5.3	50.68 ± 2.4	52.79 ± 3.4
30	41.23 ± 1.5	47.47 ± 4.2	52.36 ± 2.8	59.95 ± 3.8	61.60 ± 3.4	36.60 ± 2.5	42.11 ± 2.4	46.93 ± 4.7	53.64 ± 3.7	55.75 ± 1.3
35	44.62 ± 4.3	44.98 ± 1.3	54.44 ± 3.4	61.77 ± 4.3	64.12 ± 3.7	39.53 ± 4.8	43.84 ± 2.7	48.74 ± 3.2	55.51 ± 2.9	57.72 ± 4.1
40	47.03 ± 2.6	50.52 ± 4.3	55.82 ± 3.4	61.77 ± 4.3	64.12 ± 3.7	39.53 ± 4.8	43.84 ± 2.7	48.74 ± 3.2	55.51 ± 2.9	57.72 ± 4.1
45	48.28 ± 3.4	51.94 ± 3.8	57.12 ± 2.1	65.23 ± 4.5	66.46 ± 3.4	42.53 ± 4.3	46.23 ± 4.2	51.02 ± 3.2	58.82 ± 3.8	59.88 ± 3.2
50	49.18 ± 2.8	47.92 ± 2.3	58.12 ± 3.4	66.77 ± 2.3	67.82 ± 2.1	43.33 ± 2.8	47.05 ± 4.3	51.90 ± 4.1	59.93 ± 4.3	60.81 ± 3.1

see, MRED significantly outperforms the other related active learning methods in most cases. Compared with TED, MRED consistently shows better performance with the increase of the number of training samples.

The performance difference becomes larger when the number of training samples increases. However, when only 5 most informative samples are selected by the active learning methods, some classes should not have any labeled samples. Therefore, on this occasion, all of these active learning methods cannot obtain good performance. When the amount of selected informative samples increases, the performance of all compared methods increase. Therefore, the performance of the system can consistently improve by using the most informative samples selected by the active learning methods.

Table II shows the detailed APs and SDs for each active learning method. As we can notice, when the initially labeled set is small-sized, the random sampling method outperforms other related methods. This is mainly because the initially trained model is not very accurate given a small number of labeled samples. However, when there are only 20-35 most informative samples, our MRED method outperforms the other active learning methods which require more than 50 selected informative samples. With a large number of labeled samples, the initial model can be more accurate. Thus, the most representative samples selected by our method can provide the largest amount of new information. Therefore, we conclude that the labeling efforts of the user are alleviated by our MRED method.

2) Content-Based Image Retrieval (CBIR): In this subsection, we show how to use the proposed MRED for a CBIR task. We first give a brief description of the Corel database and the low-level feature representations.

The original Corel gallery is collected as a real-world image database and widely used as a benchmark database to demonstrate the effectiveness of the CBIR system in the past decade [24], [55]–[57]. We group the images into 80 different categories according to the ground truth of the images. Some example images are illustrated in Fig. 5.

Fig. 5. Some images in the Corel database.

We utilize three different kinds of visual features, i.e., color [58], local descriptors [59] and shape [60] to represent the images. The color moment feature vector is firstly adopted to represent the color information. 240-D Webber Local Descriptors [59] are used to describe the local visual descriptors of images. The edge directional histogram from the Y component is employed as the shape information. These visual features can characterize the contents of the images from different aspects. These three different low-level visual features are combined into a 510-dimensional vector to represent the images in the database.

The original Corel image database is divided into five subsets to evaluate the compared methods. In each round of RF, we select one subset as the query subset, and use the other four subsets as the evaluation database. We randomly select 500 images as the query subset and do the image retrieval task. The system can retrieve and rank the images in the database.

To evaluate the performance of our MRED, we compare the new MRED with MAED, TED, Locally Linear Reconstruction (LLR) [35], LSRR and SVMactive. Out of these six methods, MRED, MAED, LLR, LapROD, TED and SVMactive are considered as the conventional active learning based methods, whereas LSRR is a standard classification method. We label the first three relevant images in top twenty images as the positive samples, and label all other irrelevant images as the negative samples. For conventional active learning-based RF methods, the system selects the informative samples automatically. In experiments, we use AP and SD to evaluate the performance of the compared methods. AP is considered as the percentage of relevant images in top images presented to the user and is calculated as the averaged value of all query images. SD describes the stability of different methods. In the following, we show the performance of the compared methods using the APs and SDs from top 10 to top 60. All results are computed by averaging the results of 5-fold cross validation.

Fig. 6 and Fig. 7 show the compared performance of different methods. As shown in Fig. 6, our MRED consistently outperforms all other compared methods. Three different methods, MRED, MAED and TED are developed by following the conventional LSRR; however, these three methods can select the informative samples for training an effective classifier, and thus can significantly outperform the original LSRR. Because our MRED uses the informative samples as training samples by leveraging the manifold of the database, our method can show much better performance than the original TED. Because the classification hyperplane of SVMactive is not as good when the training data are small-sized, SVMactive does not outperform both MRED and TED. MRED and TED can label the representative samples in the database, which do not depend on the label information and is more appropriate for real-world applications. SVMactive can not be applied in the first round of RF since it requires an initial hyperplane. In the experiment, for SVMactive, we first use the standard SVM to build an initial hyperplane. Since MRED can find the informative samples for training the classifier, it shows much better performance for most of the results in experiments.

TABLE III

APs (PERCENT) IN TOP N RESULTS OF SIX ALGORITHMS (i.e., MRED, MAED, TED, LLR, LSRR, SVMactive) AFTER THE NINTH ROUND OF RF

Algorithms	MRED	MAED	TED	LLR	LSRR	SVMactive
Top 10	88.09	86.31	84.53	87.87	70.80	78.05
Top 20	78.49	76.69	75.32	76.60	62.66	74.11
Top 30	72.05	70.59	69.13	69.59	52.14	67.71
Top 40	66.83	65.48	64.13	64.16	45.13	59.53
Top 50	62.43	61.16	59.90	59.70	40.15	52.84
Top 60	54.95	53.84	52.73	52.14	32.87	43.47
Top 70	51.58	50.53	49.49	48.91	30.18	39.89
Top 80	48.58	47.60	46.62	45.98	27.88	36.95
Top 90	45.97	45.04	44.11	43.37	26.01	34.48

We can see that MRED outperforms other methods among the top 10 to top 40 results as shown in Fig. 7. For other results, MRED is similar to other related approaches. We can notice that MRED shows its effectiveness in finding the most informative samples of the database.

In [17], the system requires the user to label a large number of unlabeled samples for training a classifier. Then the uncertain samples are labeled by the user. Basically, the negative samples outnumbers than the positive ones. The user also would not like to label a large number of samples in each round of RF. Therefore, the system selects 3 relevant images and all irrelevant images in top 20 images as positive and negative samples, respectively, which can simulate the real-world CBIR systems.

The detailed results of the compared algorithms after nine rounds of RF are shown in Table III. As given in Table III, MRED achieves better performance compared with other approaches for all top results. MAED can also obtain satisfactory performance, as compared with TED, LSRR and SVMactive. Therefore, we can conclude that the proposed MRED shows the better performance in labeling the most representative samples in the database for the user to label.

We also show some qualitative results of CBIR. In this experiment, some query images are randomly selected, e.g., tiger, lion, wolf, and castle. The RF is automatically conducted based on the ground truth of images. In CBIR, four rounds of RF are performed automatically. The positive and negative samples are selected from the relevant and irrelevant images in the first screen, which contains 20 images in total. All of these positive and negative samples contain 20 images in total. In general, we select about five positive and five negative feedback samples based on the ground truth of the images. The experimental results are shown in Fig. 8. The query images are given as the first images of corresponding rows. We give the results of the initial results from top 1 to 9 without RF, SVMactive and MRED by using four rounds of RF, respectively. We also highlighted the incorrect results by red boxes. The new MRED shows much better performance for CBIR compared with related methods. For the

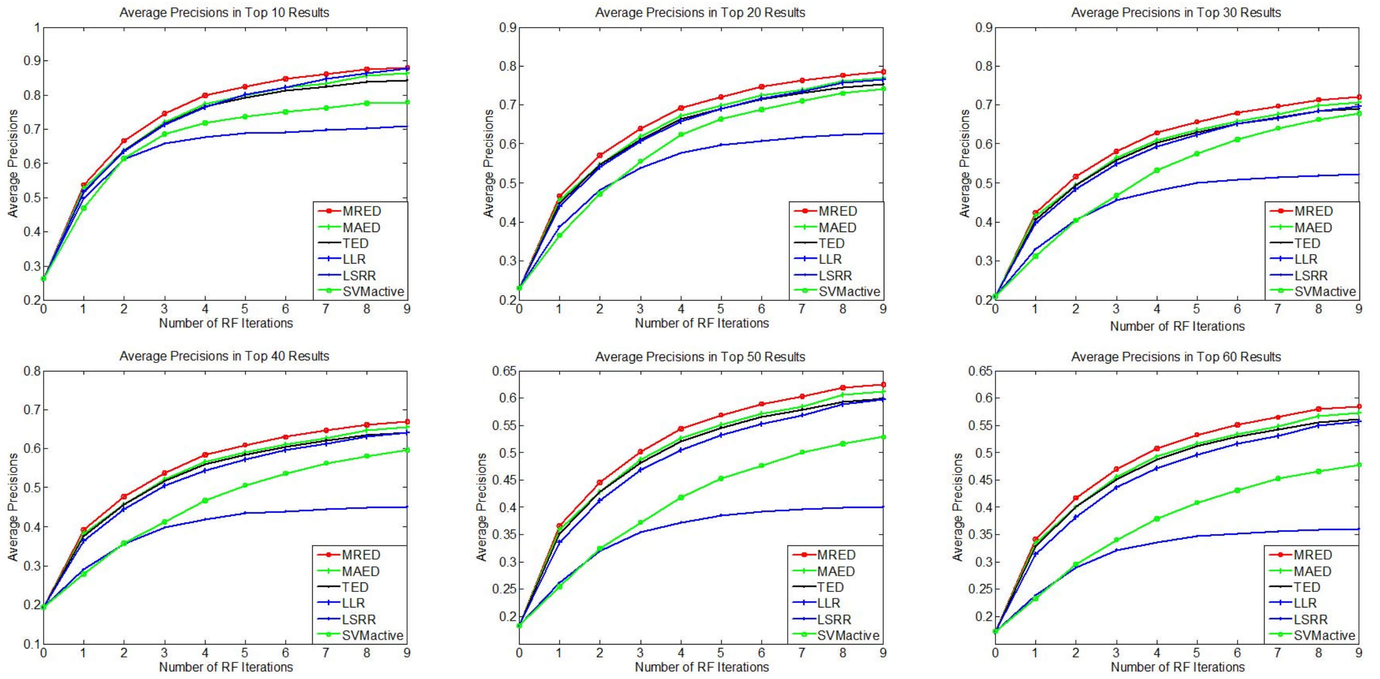


Fig. 6. APs in top 10 to top 60 results of 6 different RF approaches (i.e., MRED, MAED, TED, LLR, LSRR and SVMactive). The compared methods are based on 9 rounds of RF for CBIR.

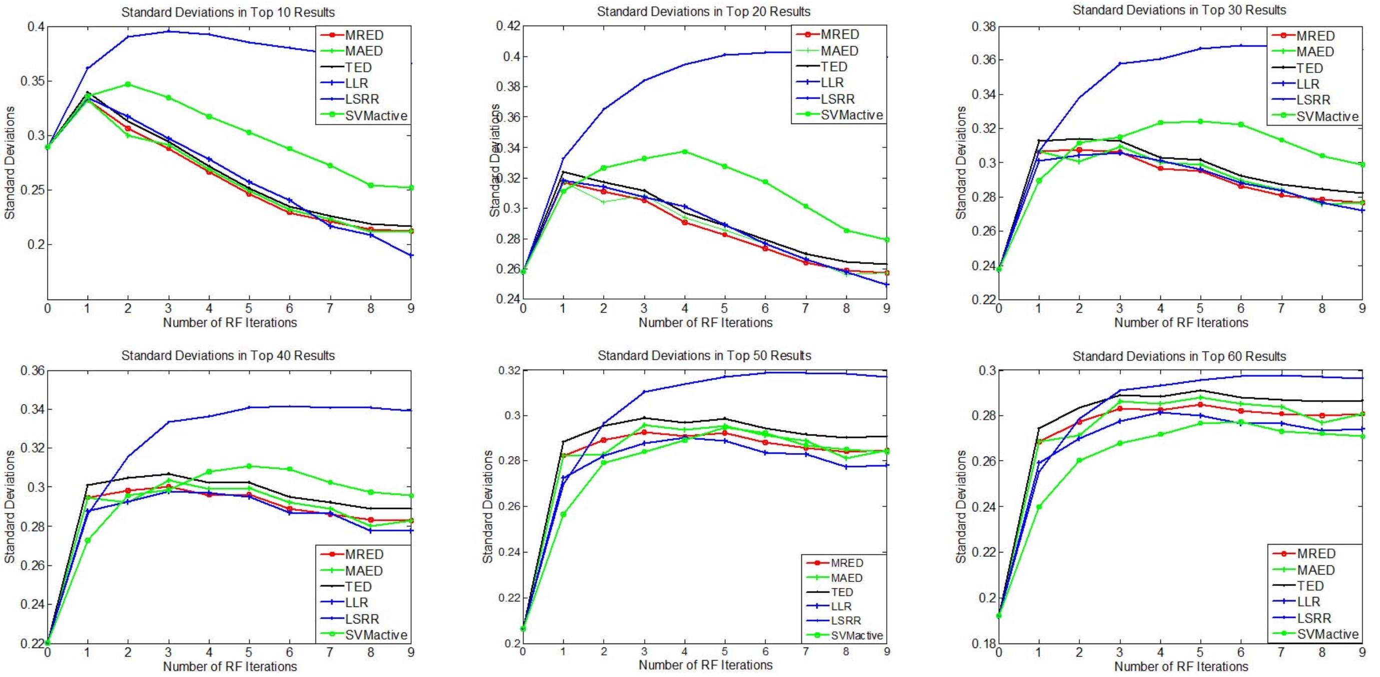


Fig. 7. SDs in top 10 to top 60 results of 5 different RF approaches (i.e., MRED, MAED, TED, LLR, LSRR and SVMactive). The compared methods are based on 9 rounds of RF for CBIR.

1^{st}, 2^{nd} and 3^{rd} query images, 9 relevant images are produced out of top 9 retrieved images. For the 4^{th} query image, the system produces 8 relevant images out of top 9 images. SVMactive also achieves comparable performance compared with the initial results. Therefore, the performance of CBIR can be significantly improved by labeling the most informative samples by our MRED.

C. Discussions

In the machine learning community, there are usually two research directions for active learning [17], [19]–[22], [26]–[28]. Conventional SVM-based active learning methods can only select uncertain samples to label by using the optimal hyperplane [17], [19]. Different from the SVM based active learning methods, MRED explores the whole database

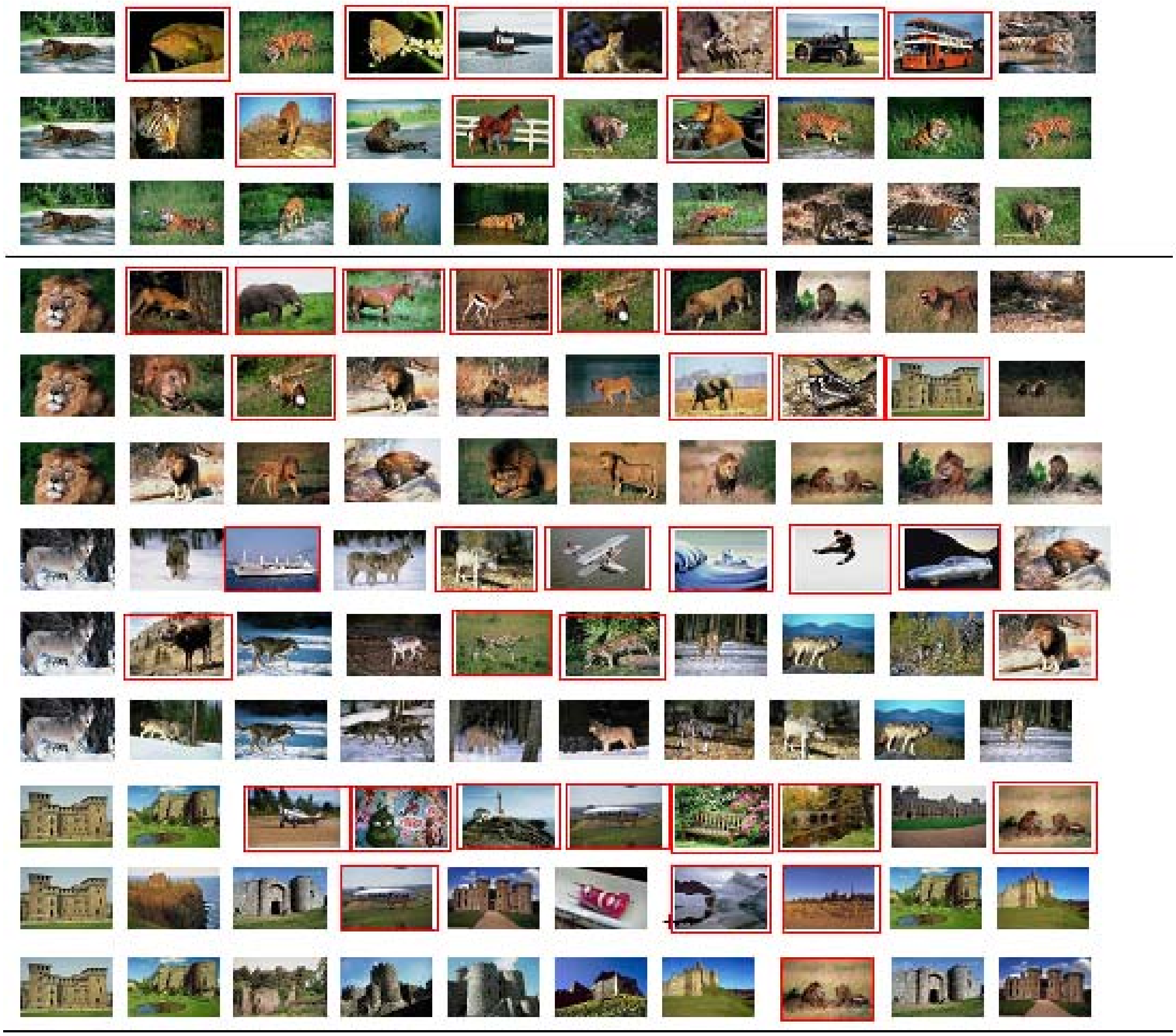


Fig. 8. Top 9 results for 4 different query images based on different active learning methods after 4 rounds of RF. The first row for each query image is the initial image retrieval result without RF. The second row for each query image is the image retrieval result based on SVMactive. The third row for each query image is the image retrieval result based on MRED.

and show much better performance when dealing with a small number of training samples by selecting the most representative samples. Similar to MRED, TED directly evaluates the predictions on testing samples and also gives a very clear geometric explanation to the selected samples [27], [28]. However, conventional TED [27], [28] only assesses the labeled samples but ignores the unlabeled samples in the database. LLR reconstructs each sample by the linear combination of its neighbors [35]. The representative samples are defined as those whose coordinates can be used to best reconstruct every other sample. However, this method is still not very appropriate since the classification model is inaccurate when training data are small-sized. This definitely affects the applications to real-world applications. Different from the conventional manifold regularization framework in [61], our method effectively selects the most informative samples in the database for the user to label. Then, the system utilizes all of the data samples including both a small number of labeled samples and a larger number of unlabeled samples to learn a classification model, which can be considered as a semisupervised learning problem. Similar to semisupervised learning, the conventional kernel deformed method [47] effectively utilizes the auxiliary information of unlabeled samples, which significantly improves the performance of conventional active learning when the size of training samples is small.

V. Conclusions

This paper presents an effective method for active learning called manifold regularized experimental design (MRED) to alleviate the labor of the user by using the most informative samples for training a classifier. Compared with other popular active learning methods, which only focus on selecting one sample in each iteration, our method allows multiple informative samples to be selected iteratively. The new method is largely inspired by the popular manifold assumption in the machine learning community, which plays an important role in semi-supervised models to significantly enhance the generalization of conventional supervised learning. Different from the previous SVMactive methods, our method does not depend on any label information of training samples and can avoid different problems caused by insufficiently labeled training samples. The new method is more appropriate and useful for different real-world applications. Various experiments on both synthetic datasets and real-world databases have demonstrated the performance of our proposed MRED. In future, we will extend our method to image classification and image annotation tasks.

References

[1] Y. Huang, K. Huang, D. Tao, T. Tan, and X. Li, "Enhanced biologically inspired model for object recognition," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 41, no. 6, pp. 1668–1680, Dec. 2011.

[2] K. Huang, D. Tao, Y. Yuan, X. Li, and T. Tan, "Biologically inspired features for scene classification in video surveillance," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 41, no. 1, pp. 307–313, Feb. 2011.

[3] L. Zhang, L. Wang, and W. Lin, "Generalized biased discriminant analysis for content-based image retrieval," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 1, pp. 282–290, Feb. 2012.

[4] W. Bian, D. Tao, and Y. Rui, "Cross-domain human action recognition," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 2, pp. 298–307, Apr. 2012.

[5] M. Yu, L. Liu, and L. Shao, "Structure-preserving binary representations for RGB-D action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 8, pp. 1651–1664, Aug. 2016.

[6] M. Yu, L. Shao, X. Zhen, and X. He, "Local feature discriminant projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 9, pp. 1908–1914, Sep. 2016.

[7] L. Zhang, H. P. H. Shum, and L. Shao, "Discriminative semantic subspace analysis for relevance feedback," *IEEE Trans. Image Process.*, vol. 25, no. 3, pp. 1275–1287, Mar. 2016.

[8] X. Zhu, "Semi-supervised learning literature survey," Dept. Comput. Sci., Univ. Wisconsin-Madison, Wisconsin, WI, USA, Tech. Rep. 1530, 2005.

[9] D. Zhou, O. Bousquet, T. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 16. 2004, pp. 321–328.

[10] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *J. Mach. Learn. Res.*, vol. 7, pp. 2399–2434, Nov. 2006.

[11] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[12] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, "Self-taught learning: Transfer learning from unlabeled data," in *Proc. 24th Int. Conf. Mach. Learn.*, 2007, pp. 759–766.

[13] L. Shao, F. Zhu, and X. Li, "Transfer learning for visual categorization: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 5, pp. 1019–1034, May 2015.

[14] L. Liu, Z. Lin, L. Shao, F. Shen, G. Ding, and J. Han, "Sequential discrete hashing for scalable cross-modality similarity retrieval," *IEEE Trans. Image Process.*, vol. 26, no. 1, pp. 107–118, Jan. 2017.

[15] B. Settles, "Active learning literature survey," Ph.D. thesis, Univ., Dept. Comput. Sci., Wisconsin–Madison, Madison, WI, USA, 2010.

[16] X. He, "Laplacian regularized D-optimal design for active learning and its application to image retrieval," *IEEE Trans. Image Process.*, vol. 19, no. 1, pp. 254–263, Jan. 2010.

[17] S. Tong and E. Chang, "Support vector machine active learning for image retrieval," in *Proc. 9th ACM Int. Conf. Multimedia*, 2001, pp. 107–118.

[18] A. Bordes, S. Ertekin, J. Weston, and L. Bottou, "Fast kernel classifiers with online and active learning," *J. Mach. Learn. Res.*, vol. 6, pp. 1579–1619, Sep. 2005.

[19] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *J. Mach. Learn. Res.*, vol. 2, pp. 45–66, Mar. 2001.

[20] L. Wang, K. Chan, and Z. Zhang, "Bootstrapping SVM active learning by incorporating unlabelled images for image retrieval," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jul. 2003, pp. 629–634.

[21] C. H. Hoi and M. R. Lyu, "A semi-supervised active learning framework for image retrieval," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, pp. 302–309.

[22] C. K. Dagli, S. Rajaram, and T. S. Huang. "Leveraging active learning for relevance feedback using an information theoretic diversity measure," in *Proc. Int. Conf. Image Video Retr.*, Berlin, Germany, 2006.

[23] 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 Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 7, pp. 1088–1099, Jul. 2006.

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

[25] L. Zhang, L. Wang, W. Lin, and S. Yan, "Geometric optimum experimental design for collaborative image retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 2, pp. 346–359, Feb. 2014.

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

[27] K. Yu, J. Bi, and V. Tresp, "Active learning via transductive experimental design," in *Proc. 23rd Int. Conf. Mach. Learn.*, vol. 23. 2006, pp. 1081–1088.

[28] K. Yu, S. Zhu, W. Xu, and Y. Gong, "Non-greedy active learning for text categorization using convex ansductive experimental design," in *Proc. 31st Int. Conf. Res. Develop. Inf. Retr.*, 2008, pp. 635–642.

[29] O. Chapelle *et al.*, *Semi-Supervised Learning*. Cambridge, MA, USA: MIT Press, 2006.

[30] X. Zhu, Z. Ghahramani, and J. Lafferty, "Semi-supervised learning using Gaussian fields and harmonic functions," in *Proc. 20th Int. Conf. Mach. Learn.*, 2003, vol. 20. no. 2, pp. 912–920.

[31] D. Cai, X. He, and J. Han, "Semi-supervised discriminant analysis," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–7.

[32] Y. Yang, F. Nie, D. Xu, J. Luo, Y. Zhuang, and Y. Pan, "A multimedia retrieval framework based on semi-supervised ranking and relevance feedback," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 723–742, Apr. 2012.

[33] J. Yu, Y. Rui, and B. Chen, "Exploiting click constraints and multi-view features for image re-ranking," *IEEE Trans. Multimedia*, vol. 16, no. 1, pp. 159–168, Jan. 2014.

[34] Z. Pan, X. You, H. Chen, D. Tao, and B. Pang, "Generalization performance of magnitude-preserving semi-supervised ranking with graph-based regularization," *Inf. Sci.*, vol. 221, pp. 284–296, Feb. 2013.

[35] L. Zhang, C. Chen, J. Bu, D. Cai, X. He, and T. S. Huang, "Active learning based on locally linear reconstruction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 10, pp. 2026–2038, Oct. 2011.

[36] J. Yu, Y. Rui, Y. Y. Tang, and D. Tao, "High-order distance-based multiview stochastic learning in image classification," *IEEE Trans. Cybern.*, vol. 44, no. 12, pp. 2431–2442, Dec. 2014.

[37] J. Yu, D. Liu, D. Tao, and H. Seah, "On combining multiple features for cartoon character retrieval and clip synthesis," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 5, pp. 1413–1427, Oct. 2012.

[38] B. Geng, D. Tao, C. Xu, L. Yang, and X.-S. Hua, "Ensemble manifold regularization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 6, pp. 1227–1233, Jun. 2012.

[39] S. C. H. Hoi, R. Jin, and M. R. Lyu, "Batch mode active learning with applications to text categorization and image retrieval," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1233–1248, Sep. 2009.

[40] Y. Yang, Z. Ma, F. Nie, X. Chang, and A. G. Hauptmann, "Multi-class active learning by uncertainty sampling with diversity maximization," *Int. J. Comput. Vis.*, vol. 113, no. 2, pp. 113–127, 2015.

[41] C. Long and G. Hua, "Multi-class multi-annotator active learning with robust Gaussian process for visual recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 2839–2847.

[42] D. Cai and X. He, "Manifold adaptive experimental design for text categorization," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 4, pp. 707–719, Apr. 2012.

[43] X. You, R. Wang, and D. Tao, "Diverse expected gradient active learning for relative attributes," *IEEE Trans. Image Process.*, vol. 23, no. 7, pp. 3203–3217, Jul. 2014.

[44] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer-Verlag, 2000.

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

[46] M. Belkin and P. Niyogi, "Using manifold structure for partially labelled classification," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 15. 2002, pp. 953–960.

[47] V. Sindhwani, P. Niyogi, and M. Belkin, "Beyond the point cloud: From transductive to semi-supervised learning," in *Proc. 22nd Int. Conf. Mach. Learn.*, 2005, pp. 824–831.

[48] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proc. 7th Annu. Conf. Comput. Learn. Theory*, Jul. 1998, pp. 92–100.

[49] X. Zhu, "Semi-supervised learning literature survey," Dept. Comput. Sci., Univ. Wisconsin-Madison, Madison, WI, USA, Tech. Rep. TR1530, 2005.

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

[51] P. Niyogi and X. He, "Locality preserving projections," in *Proc. Adv. Neural Inf. Process. Syst.*, 2004.

[52] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[53] M. Yuan and Y. Lin, "Model selection and estimation in the Gaussian graphical model," *Biometrika*, vol. 94, no. 1, pp. 19–35, 2007.

[54] Yale Univ. (2002). *Face Database*. [Online]. Available: http://cvc.yale.edu/projects/yalefaces/yalefaces.html

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

[56] C. H. Hoi, W. Liu, and S. F. Chang, "Semi-supervised distance metric learning for collaborative image retrieval and clustering," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 6, no. 3, pp. 1–26, 2010.

[57] L. Zhang, L. Wang, and W. Lin, "Conjunctive patches subspace learning with side information for collaborative image retrieval," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3707–3720, Aug. 2012.

[58] M. J. Swain and D. H. Ballard, "Color indexing," *Int. J. Comput. Vis.*, vol. 7, no. 1, pp. 11–32, 1991.

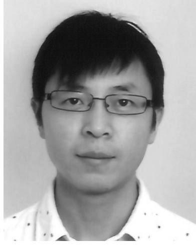
[59] J. Chen *et al.*, "WLD: A robust local image descriptor," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1705–1720, Sep. 2010.

[60] A. K. Jain and A. Vailaya, "Image retrieval using color and shape," *Pattern Recognit.*, vol. 29, no. 8, pp. 1233–1244, 1996.

[61] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2002, pp. 585–591.

Lining Zhang (S'11–M'14) received the B.Eng. and M.Eng. degrees from Xidian University, Xi'an, China, and the Ph.D. degree from Nanyang Technological University, Singapore. He was a Research Scientist with the Ocular Imaging Program, Institute for Infocomm Research, and a Research Engineer with the Learning and Vision Research Group, National University of Singapore. He is currently with Northumbria University, Newcastle upon Tyne, U.K. He has authored extensively in top venues, such as the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and the IEEE TRANSACTIONS ON CYBERNETIC. His research interests include computer vision, video/image processing, medical image analysis, machine learning, and computational intelligence.

Hubert P. H. Shum received the M.Sc. and B.Eng. degrees from the City University of Hong Kong and the Ph.D. degree from the School of Informatics, The University of Edinburgh. He was a Lecturer with the University of Worcester, a Post-Doctoral Researcher with RIKEN Japan, and a Research Assistant with the City University of Hong Kong. He is currently an Associate Professor (Reader) with Northumbria University. His research interests include character animation, machine learning, human motion analysis, and computer vision.



Ling Shao (M'09–SM'10) is a professor with the School of Computing Sciences at the University of East Anglia, Norwich, UK. Previously, he was a professor (2014-2016) with Northumbria University, a senior lecturer (2009-2014) with the University of Sheffield and a senior scientist (2005-2009) with Philips Research, The Netherlands. His research interests include computer vision, image/video processing and machine learning. He is an associate editor of IEEE Transactions on Image Processing, IEEE Transactions on Neural Networks and Learning Systems and several other journals. He is a Fellow of the British Computer Society and the Institution of Engineering and Technology. He is a senior member of the IEEE.