Discriminative Semantic Subspace Analysis for Relevance Feedback

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

***Abstract*— Content-based image retrieval (CBIR) has attracted much attention during the past decades for its potential practical applications to image database management. A variety of relevance feedback (RF) schemes have been designed to bridge the gap between low-level visual features and high-level semantic concepts for an image retrieval task. In the process of RF, it would be impractical or too expensive to provide explicit class label information for each image. Instead, similar or dissimilar pairwise constraints between two images can be acquired more easily. However, most of the conventional RF approaches can only deal with training images with explicit class label information. In this paper, we propose a novel discriminative semantic subspace analysis (DSSA) method, which can directly learn a semantic subspace from similar and dissimilar pairwise constraints without using any explicit class label information. In particular, DSSA can effectively integrate the local geometry of labeled similar images, the discriminative information between labeled similar and dissimilar images, and the local geometry of labeled and unlabeled images together to learn a reliable subspace. Compared with the popular distance metric analysis approaches, our method can also learn a distance metric but perform more effectively when dealing with high-dimensional images. Extensive experiments on both the synthetic data sets and a real-world image database demonstrate the effectiveness of the proposed scheme in improving the performance of the CBIR.**

***Index Terms*— Content-based image retrieval, relevance feedback, pairwise constraints, distance metric analysis.**

I. Introduction

CONTENT-BASED image retrieval (CBIR) has attracted much attention during the past decades [1]–[5]. Conventional CBIR systems usually adopt the Euclidean distance metric in a high-dimensional low-level visual feature space to measure the similarity between the query image and the images in the database [1]–[4], [6]. However, the Euclidean distance metric in a high-dimensional space is usually not very effective due to the gap between the low-level visual features and the high-level semantic concepts.

Relevance feedback (RF) is one of the most powerful tools to narrow down this semantic gap and thus to improve the performance of a CBIR system [7], [8]. In general, RF focuses on the interactions between a user and a search engine by requiring the user to label semantically similar and dissimilar images with the query image [7], [8], which are positive and negative feedback samples, respectively. During the past decades, various RF approaches have been designed based on different assumptions for the positive and negative feedback samples [8]. One-class support vector machine (SVM) estimates the density of positive feedback samples but ignores the negative feedback samples [9]. Two-class SVM can identify both positive and negative feedback samples but treats these two different groups equally [10]. In [3], Tao et al. included positive feedback samples in a single set and split negative feedback samples into a small number of subsets, and a series of kernel marginal convex machines were developed between one positive group and several negative subgroups. The results indicate that clustering the negative feedback samples into several subgroups can indeed improve the overall retrieval performance.

Beyond conventional RF approaches, several new schemes have emerged to attack this semantic gap in CBIR [11]–[16]. For instance, image annotation techniques intend to directly acquire the semantic concepts from the low-level visual features of an image [11]. However, major challenges still remain in image annotation. Recently, collaborative image retrieval (CIR) was introduced to alleviate the labeling efforts of conventional RF approaches by leveraging various auxiliary information [12]–[16]. We can roughly 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 [12], [13], [15]. In [12], 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 the low-level visual features and the high-level semantic concepts. In [14], Liu et al. proposed a novel RF method for personal image retrieval via a cross-domain learning scheme, and it can effectively alleviate the labeling efforts of conventional RF 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 image database [16]–[21], which could be labeled by the user in RF and used as the training data to define an effective similarity metric for image retrieval.

However, for a conventional CBIR task, the need for online RF stems from the fact that different semantic concepts may occur in different subspaces and the selection of such

Manuscript received June 13, 2015; revised October 6, 2015; accepted December 28, 2015. Date of publication January 18, 2016; date of current version February 2, 2016. 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 Prof. Wen Gao. *(Corresponding author: Ling Shao.)*

The authors are with the Department of Computer Science and Digital Technologies, Northumbria University, Newcastle Upon Tyne NE1 8ST, U.K. (e-mail: liningzh@gmail.com; hubert.shum@northumbria.ac.uk; 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.2016.2516947

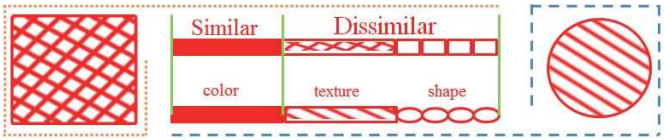


Fig. 1. Two synthetic images and the associated low-level visual features in a high-dimensional space for concept illustration. It is not very effective to measure the similarity between two images in the original high-dimensional visual feature space due to the semantic gap. Two images can only be similar in a low-dimensional semantic subspace.

subspaces cannot be done offline [22], and it is the goal of RF schemes to figure out which one [22]. However, it will be a burden for conventional RF schemes to tune the internal parameters to adapt to the changes of such semantic subspaces. Such difficulties have severally degraded the effectiveness of conventional RF for an image retrieval task.

Subspace analysis approaches play an important role in various tasks in computer vision, such as, face recognition [23], [24], gait recognition [25]–[27], image classification [28] and web image annotation [11], [29]. Let us first use a toy example to show the importance of subspace analysis in measuring the similarity between a pair of images, which is usually the key issue in image retrieval. For a conventional CBIR task, the images are usually represented by a set of low-level visual features with various semantic concepts (e.g., color, shape and texture) in a high-dimensional visual feature space. With an assumption that different semantic concepts occupy different subspaces and each image can reside many different subspaces, Fig.1 shows two toy images, each of which is associated with a number of semantic concepts, i.e., color, shape, texture and size, in a high-dimensional visual feature space. However, in RF, it is not appropriate to directly measure the similarity between two images based on the Euclidean distance metric in a high-dimensional multiple semantic concept space (e.g., color, texture and shape) due to the semantic gap. This is mainly because there are many different semantic concept subspaces in the original high-dimensional visual feature space and the two images can only be similar in one low-dimensional semantic concept subspace, e.g., color, but dissimilar with each other in the other semantic concept subspaces, e.g., texture and shape. Therefore, it is more reasonable to measure the similarity between two images in the low-dimensional semantic subspace than in the original multiple high-dimensional semantic concept space. By selecting a 1-D semantic subspace, measuring the similarity between a pair of images will be easy and obvious.

Subspace analysis approaches project the original high-dimensional feature space to a low-dimensional subspace, where specific statistical properties can be well preserved. For example, Fisher's Linear Discriminant Analysis (LDA) [30], the most traditional supervised subspace analysis method, minimizes the trace ratio between the within-class scatter and the between-class scatter so that Gaussian distributed samples can be well separated in the selected subspace; locality preserving projections (LPP) preserve the local geometry of samples by processing an undirected weighted graph that represents the neighborhood relations of pairwise samples [31]. The aforementioned subspace analysis methods function impressively on both artificial datasets and practical applications, such as face recognition. However, most of these traditional subspace analysis approaches (e.g., LDA [30], [32]) normally need to acquire explicit class label information. In RF, explicit class label information for each image might be too expensive to obtain [8]. Compared with explicit class label information of each image, the similar or dissimilar pairwise constraints between two images can be acquired more easily when the user-labeled information is available. Therefore, it is more attractive to learn a semantic concept subspace directly from the similar and dissimilar pairwise constraints without using the explicit image class label information. Recently, distance metric analysis with similar and dissimilar pairwise constraints have been actively studied in the machine learning community [33]–[38]. Despite the active research efforts during the past few years, most of these approaches in this area have involved a high-computational burden when dealing with high-dimensional images and also cannot give explicit image representations in the low-dimensional semantic concept subspace, which are thus not appropriate and will significantly limit their potential applications to the CBIR research [33]–[36].

In this paper, we propose a novel discriminative semantic subspace analysis (DSSA) method to bridge the gap between low-level visual features and high-level semantic concepts by exploiting the training images with pairwise constraints in RF. The proposed DSSA method can effectively learn a reliable subspace both from labeled and unlabeled images with similar and dissimilar pairwise constraints without using any explicit class label information. Specifically, DSSA can effectively integrate the local geometry of labeled similar images, the discriminative information between labeled similar and dissimilar images, and the local geometry of unlabeled images together to learn a reliable subspace. Compared with the popular distance metric analysis methods with pairwise constraints, our method can also learn a distance metric but perform more effectively when dealing with high-dimensional images, which is more appropriate for a CBIR task.

The rest of this paper is organized as follows. Section II reviews the related work. DSSA with similar and dissimilar pairwise constraints are detailed in Section III. A CBIR system based on the proposed method 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 some analysis to the important parameters in DSSA. 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: 1) RF and 2) distance metric analysis with pairwise constraints.

A. Review of RF

During the past few years, various RF methods have been developed based on different assumptions for the positive and negative feedback samples. One-class support vector machine (SVM) estimates the density of positive feedback

samples but ignores the negative feedback samples [9]. Two-class SVM makes use of both the positive and negative feedback samples but treat the two groups equally [10]. Biased discriminant analysis techniques define a (1+x) class problem and find a subspace to separate one positive class from the unknown number of negative classes [22], [39]–[41].

CIR attempts to alleviate the labeling efforts of conventional RF schemes by resorting to the user historical feedback log data or the large-scale web data. In [12], Hoi et al. proposed a log-based RF scheme with the SVM by engaging the user feedback log data in a regular RF task. In [14], a textual query-based personal image retrieval system was proposed, which can significantly alleviate the labeling efforts of conventional RF by leveraging millions of loosely labeled web images via a cross-domain learning scheme. Most of the conventional RF approaches can only deal with training images with explicit class label information. However, in RF, explicit class label information for each image might be too expensive to obtain.

B. Review of Distance Metric Analysis With Pairwise Constraints

Suppose we have a database X consisting of n images $x_i (1 \leq i \leq n)$ in a high-dimensional visual feature space R^h, i.e., $X = [x_1, \ldots, x_n] \in R^{h \times n}$. Given prior information that certain pairs of images are similar: $S : (x_i, x_j) \in S$ if x_i and x_j are judged as a similar pair, and dissimilar: $D : (x_i, x_j) \in D$ if x_i and x_j are judged as a dissimilar pair. Distance metric analysis methods aim to learn a distance metric $d_M(x_i, x_j)$ between images x_i and x_j, such that dissimilar images are far from each other and similar images are close to each other. The distance metric between two images x_i and x_j is defined as:

$$d_M(x_i, x_j) = ||x_i - x_j||_M = \sqrt{(x_i - x_j)^T M (x_i - x_j)}, \quad (1)$$

where $M \in R^{h \times h}$ is a positive semi-definite matrix. Setting $M = I$ means using the Euclidean distance metric. More generally, M represents a family of Mahalanobis distance metrics. By adopting the eigenvalue decomposition, M can be rewritten as $M = WW^T, W \in R^{h \times l}, l \leq h$, so Eq.(1) can be rewritten as:

$$\sqrt{(x_i - x_j)^T (WW^T)(x_i - x_j)} = ||W^T x_i - W^T x_j||, \quad (2)$$

Let $y = W^T x$, then:

$$\begin{aligned} d(y_i, y_j) &= ||W^T x_i - W^T x_j|| \\ &= \sqrt{(x_i - x_j)^T M (x_i - x_j)}, \end{aligned} \quad (3)$$

Therefore, learning a Mahalanobis distance metric M in the high-dimensional visual feature space is equivalent to learning an efficient mapping matrix W that replaces each image x with $W^T x$ and applying the standard Euclidean distance metric to the images in the low-dimensional space.

Distance metric analysis methods are usually accomplished based on a set of labeled data with pairwise constraints. For example, neighborhood component analysis (NCA) was proposed to learn a Mahalanobis distance metric by directly maximizing the leave-one-out cross validation accuracy of k-nearest neighbors. The large margin nearest neighbor (LMNN) method was proposed to take the margin into account and separate the samples of different classes in a large margin manner [42]. In [34], a relevant component analysis (RCA) technique was proposed to exploit only similar pairwise constraints for distance metric analysis. In detail, given the pairwise constraints, RCA first forms a set of chunklets, each of which is defined as a group of samples linked together by similar pairwise constraints. The optimal distance metric learned by RCA can be computed as the inverse of the average covariance matrix of the chunklets. In [33], Xing et al. proposed a distance metric analysis approach (called Xing hereafter) and formulated the task into a convex optimization problem, which can be solved by an iterative projection algorithm. RCA is simple to calculate but ignores the dissimilar pairwise constraints. Discriminative component analysis (DCA) was proposed to incorporate the dissimilar pairwise constraints [36], which can show slightly better discriminative performance compared with RCA for some datasets. Lately, an information-theoretic metric learning approach was proposed to express the weakly supervised learning problem as a Bregman optimization problem [43]. In [44], Guillaumin et al. offered a probabilistic view on learning a Mahalanobis distance metric and posteriori class probabilities were treated as similar and dissimilar measures. A simple but effective strategy to learn a distance metric from equivalence constraints was introduced based on a statistical inference perspective [45]. Different from the previous metric learning methods, a Probabilistic Relative Distance Comparison (PRDC) model was proposed to maximize the probability of a pair of true match having a smaller distance than that of a wrong match pair [46]. Although encouraging performance has been shown, most of these approaches in this area have involved a high-computational burden when dealing with high-dimensional images and also cannot give explicit image semantic representations in the low-dimensional semantic concept subspace, which are thus not appropriate and will significantly limit their potential applications to the CBIR research [33]–[37].

III. DSSA FOR RF

In RF, the images returned for a certain query are usually represented by low-level visual features, i.e., $X = [x_1, \ldots, x_n] \in R^{h \times n}$ in a high-dimensional space with $x_i \in R^h$ for an image. The performance of CBIR using the Euclidean distance metric in a high-dimensional space is usually poor because of the gap between low-level visual features and high-level semantic concepts.

With the RF information, this semantic gap can be reduced significantly. By mining the user-labeled information, we can learn a submanifold to encode the user intention. This submanifold is embedded in the ambient space, i.e., the high-dimensional low-level visual feature space R^h. In this paper, a linear subspace W is used to approximate this submanifold such that the images can be represented as $Y = W^T X = [y_1, \ldots, y_n] \in R^{l \times n} (l < h)$ with $y_i \in R^l$ for an image x_i. Therefore, in the low-dimensional subspace, an improved image retrieval performance is expected.

To measure the similarity between two images in the low-dimensional subspace, we adopt the widely used

Euclidean distance metric. Learning a mapping matrix W is actually equivalent to learning an efficient Mahalanobis distance metric M in the original high-dimensional space. In recent years, a variety of techniques have been proposed to learn such an optimal Mahalanobis distance metric M from training data that are given in the form of pairwise constraints [33]–[36], [47], [48]. However, most of these methods are inappropriate for CBIR, since they either require solving a convex optimization problem with gradient decent and iterative projections or involve solving a semidefinite programming problem that often suffers from high computational costs, which limits their potential applications for high-dimensional data [33], [34], [48]. Moreover, most of these methods, which can learn distance metrics from the training data, are unable to explicitly give the new representations of data in the new metric space.

Therefore, in this paper, we present a DSSA method to learn such a mapping matrix W. Particularly, the DSSA can effectively integrate the local geometry and the discriminative information of labeled images, and the local geometry of labeled and unlabeled images together. This process is conducted by building different kinds of local patches for each image, and then aligning these different kinds of patches together to learn a consistent coordinate [49], [50]. One patch is a local area, which is formed by one image and its associated neighboring images. Particularly, in DSSA, we build three different kinds of patches: 1) local geometric patches for labeled similar images to represent the local geometry of labeled similar images; 2) local discriminative patches for labeled similar and dissimilar images to represent the discriminative information between labeled similar and dissimilar images; 3) local similar patches for labeled and unlabeled images to represent the similar information of labeled and unlabeled images.

A. DSSA for RF

1) Local Geometric Patches for Labeled Similar Images: With an observation that all labeled similar images are alike, while each labeled dissimilar image is dissimilar in its own way, BDA was introduced as a principled way to select a subset of image visual features and define a suitable similarity metric [22]. Thus, all labeled similar images are required to be close to each other in the learned subspace. However, this assumption is usually not reliable in conventional RF.

Labeled similar images may vary in appearance and the corresponding visual features. For instance, for query "train", labeled similar images are usually different from each other, as shown in Fig. 2. For this reason, instead of requiring all labeled similar images to be close to each other in the projected subspace, it is more appropriate to only retain the local geometry of labeled similar images in RF.

Specifically, for each image x_i associated with a local geometric patch $X_{g(i)} = [x_i, x_{i_1} \ldots, x_{i_{k_1}}]$, wherein $x_{i_1}, x_{i_2}, \ldots, x_{i_{k_1}}$, i.e., the k_1 nearest images with similar constraints. This paper assumes that the new representation y_i of x_i can be linearly reconstructed by its k_1 nearest images with similar constraints, i.e.,

$$x_i = c_{i_1} x_{i_1} + c_{i_2} x_{i_2} + \ldots + c_{i_{k_1}} x_{i_{k_1}} + \varepsilon_i, \tag{4}$$

Fig. 2. For query "train", labeled similar images are different from each other in appearance. Therefore, it is not reasonable to require all labeled similar images to be close to each other in the projected subspace.

where c_i is a k_1 dimensional vector encoding the reconstruction coefficients and ε_i is the reconstruction error. Minimizing the error yields

$$\arg\min_{c_i} ||\varepsilon_i||^2 = \arg\min_{c_i} ||x_i - \sum_{j=1}^{k_1} c_{i_j} x_{i_j}||^2, \tag{5}$$

With the sum-to-one constraint: $\sum_{j=1}^{k_1} (c_i)_j = 1, c_i$ can be computed in a closed form:

$$c_{i_j} = \frac{\sum_{t=1}^{k_1} G_{jt}^{-1}}{\sum_{p=1}^{k_1} \sum_{q=1}^{k_1} G_{pq}^{-1}}, \tag{6}$$

where $G_{jt} = \left(x_i - x_{i_j}\right)^T (x_i - x_{i_t})$ is called the local Gram matrix [51].

We assume that c_i reconstructs both x_i from $x_{i_1}, \ldots, x_{i_{k_1}}$ in the high-dimensional space and y_i from $y_{i_1}, \ldots, y_{i_{k_1}}$ in the low-dimensional subspace. Based on this point, the cost function can be reformulated as

$$\begin{aligned} \arg\min_{Y_{g(i)}} ||\sigma_i||^2 &= \arg\min_{Y_{g(i)}} ||y_i - \sum_{j=1}^{k_1} (c_i)_j y_{i_j}||^2 \\ &= \arg\min_{Y_{g(i)}} tr\left(Y_{g(i)} \begin{bmatrix} -1 \\ c_i \end{bmatrix} \begin{bmatrix} -1 & c_i^T \end{bmatrix} Y_{g(i)}^T\right) \\ &= \arg\min_{Y_{g(i)}} tr\left(Y_{g(i)} L_{g(i)} Y_{g(i)}^T\right) \end{aligned} \tag{7}$$

where

$$\begin{aligned} Y_{g(i)} &= [y_i, y_{i_1}, \ldots, y_{i_{k_1}}]; \\ L_{g(i)} &= \begin{bmatrix} -1 \\ c_i \end{bmatrix} \begin{bmatrix} -1 & c_i^T \end{bmatrix} = \begin{bmatrix} 1 & -c_i^T \\ -c_i & c_i c_i^T \end{bmatrix} \end{aligned}$$

with

$$\bar{c}_i = [c_i^T, \underbrace{0, \ldots, 0}_{k_1}]^T; g(i)$$

is used to encode the local geometry of labeled similar images in RF.

2) Local Discriminative Patches for Labeled Similar and Dissimilar Images: In RF, given an image x_i, according to the user-labeled information, we can divide the other images into two categories: images with similar pairwise constraints and images with dissimilar pairwise constraints. We select k_1 images with respect to x_i from similar images and term them neighbor images with similar pairwise constraints denoted by $x_{i_1}, \ldots, x_{i_{k_1}}$. We select k_2 nearest neighbors with respect to x_i from dissimilar images and term them

neighbor images with dissimilar pairwise constraints denoted by $x_{i_1}, \ldots, x_{i_{k_2}}$. By putting $x_i, x_{i_1}, \ldots, x_{i_{k_1}}$, and $x_{i_1}, \ldots x_{i_{k_2}}$ together, we can build the local discriminative patch for an image x_i as $X_d(i) = [x_i, x_{i_1}, \ldots, x_{i_{k_1}}, x_{i_1}, \ldots, x_{i_{k_2}}]$.

Especially, for the new representations of each local discriminative patch, i.e., $Y_{d(i)} = [y_i, y_{i_1}, \ldots, y_{i_{k_1}}, y_{i_{k_1+1}}, \ldots, y_{i_{k_1+k_2}}]$, we expect that distances between the given image and the neighbor similar images are as small as possible, while distances between the given measurement and the neighbor dissimilar images are as large as possible.

For each patch in the low-dimensional subspace, we expect that distances between y_i and the neighbor images with similar pairwise constraints are as small as possible, so we have

$$\arg\min_{y_i} \sum_{j=1}^{k_1} ||y_i - y_{i_j}||^2, \tag{8}$$

Meanwhile, we expect that distances between y_i and the neighbor images with dissimilar pairwise constraints are as large as possible, so we have

$$\arg\max_{y_i} \sum_{m=1}^{k_2} ||y_i - y_{i_m}||^2, \tag{9}$$

Since the patch formed by the local neighborhood can be approximately regarded as linear [50], [52], we formulate the part discriminator by using the linear manipulation as follows:

$$\arg\min_{y_i} \left(\sum_{j=1}^{k_1} ||y_i - y_{i_j}||^2 - \beta \sum_{m=1}^{k_2} ||y_i - y_{i_m}||^2 \right), \tag{10}$$

where β is a scaling factor in [0, 1] to unify different measures of the within-class distances and the between-class distances. We define the coefficients vector as

$$\omega_i = \left[\overbrace{1, \ldots, 1}^{k_1}, \overbrace{-\beta, \ldots, -\beta}^{k_2} \right]^T, \tag{11}$$

To rewrite Eq.(10) into a compact form, we have

$$\begin{aligned}
&\arg\min_{y_i} \left(\sum_{j=1}^{k_1} ||y_i - y_{i_j}||^2 (\omega_i)_j + \sum_{p=1}^{k_2} ||y_i - y_{i_p}||^2 (\omega_i)_{p+k_1} \right) \\
&= \arg\min_{y_i} \left(\sum_{j=1}^{k_1+k_2} ||y_{F_i(1)} - y_{F_i(i+j)}||^2 (\omega_i)_j \right) \\
&= \arg\min_{Y_{d(i)}} tr \left(\begin{matrix} Y_{d(i)} \begin{bmatrix} -e_{k_1+k_2}^T \\ I_{k_1+k_2} \end{bmatrix} diag(\omega_i) \\ \times \begin{bmatrix} -e_{k_1+k_2} & I_{k\cdot+k_2} \end{bmatrix} Y_{d(i)}^T \end{matrix} \right) \\
&= \arg\min_{Y_{d(i)}} tr \left(Y_{d(i)} L_{d_{(i)}} Y_{d(i)}^T \right),
\end{aligned} \tag{12}$$

where

$$L_d(i) = \begin{bmatrix} \sum_{j=1}^{k_1+k_2} (\omega_i)_j & -\omega_i^T \\ -\omega_i & diag(\omega_i) \end{bmatrix};$$

$F_i = \{i, i_1, \ldots, i_{k_1}, i_{k_1+1}, \ldots, i_{k_1+k_2}\}$ is the set of indices for images on the patch; $e_{k_1+k_2} = [1, \ldots, 1]^T \in R^{k_1+k_2}$; and $I_{k_1+k_2}$ is a $(k_1 + k_2) \times (k_1 + k_2)$ identity matrix.

3) Local Similar Patches for Labeled and Unlabeled Images: Conventional RF approaches are developed based on supervised learning (i.e., BDA RF or SVM RF) models. However, in RF, the efforts of requiring the user to label a large number of images is generally laborious, although vast amounts of unlabeled images are readily available in the database and can also provide useful information to enhance the performance of CBIR. Semi-supervised learning under such a scenario is often designed to significantly improve the generalization ability of supervised learning by leveraging abundant unlabeled images in the database [53], [54].

Unlabeled images are valuable in improving the local geometry of supervised learning models [53], [54]. Unlabeled images are attached to the original data set: $X_u = [x_1, \ldots, x_n, x_{n+1}, \ldots, x_{n+n_u}]$, where the first n images are labeled, and the remaining n_u images are unlabeled. For each image $x_i, = 1, \ldots, n + n_u$, we search its k_3 nearest neighbors $x_{i_1}, \ldots, x_{i_{k_3}}$ in all training data including both labeled and unlabeled images. Let $X_{u(i)} = [x_i, x_{i_1}, \ldots, x_{i_{k_3}}]$ denote the i^{th} patch.

To preserve the local geometry of labeled and unlabeled images, the nearby images should stay nearby in the low-dimensional space, or $y_i \in R^l$ is close to $y_{i_1}, \ldots, y_{i_{k_3}}$, i.e.,

$$\arg\min_{y_i} \sum_{j=1}^{k_3} ||y_i - y_{i_j}||^2 (\omega_i)_j, \tag{13}$$

where $y_{i_j}, j = 1, \ldots, k_3$ are k_3 connected images of a given image y_i and ω_i is the k_3-dimensional vector weighted by $(\omega_i)_j = \exp(-||x_i - x_{i_j}||^2/t)$, where t is set as a suitable constant according to [31]. Therefore, Eq. (13) can be reformulated as

$$\begin{aligned}
&\arg\min \sum_{j=1}^{k_3} tr \left(\begin{matrix} \begin{bmatrix} (y_i - y_{i_1})^T \\ \vdots \\ (y_i - y_{i_1})^T \end{bmatrix} \begin{bmatrix} y_i - y_{i_1}, \ \ldots \ y_i - y_{i_1} \end{bmatrix} \\ \times diag(\omega_i) \end{matrix} \right) \\
&= \arg\min_{Y_{u(i)}} tr \left(\begin{matrix} Y_{u(i)} \begin{bmatrix} -e_{k_3}^T \\ I_{k_3} \end{bmatrix} diag(\omega_i) \\ \times \begin{bmatrix} -e_{k_3} & I_{k_3} \end{bmatrix} Y_i^T \end{matrix} \right) \\
&= \arg\min_{Y_{u(i)}} tr \left(Y_{u(i)} L_{u(i)} Y_{u(i)}^T \right),
\end{aligned} \tag{14}$$

where

$$Y_{u(i)} = [y_i, y_{i_1}, \ldots, y_{i_1}];$$

$$L_i = \begin{bmatrix} \sum_{j=1}^{k_3} (\omega_i)_j & -\omega_i^T \\ -\omega_i & diag(\omega_i) \end{bmatrix}; e_{k_3} = [1, \ldots, 1]^T \in R^{k_3}$$

is a $k_3 \times k_3$ identity matrix; $L_{u(i)}$ is used to encode the local geometry of labeled and unlabeled images.

4) DSSA: Each patch has its own coordinate system. With the calculated local patches, we can align them together into a consistent coordinate [49], [50]. For each image I_i, $Y = [y_i, y_{i_1}, \ldots, y_{i_k}]$ can be rewritten as $Y = YS_i$, where $Y = [y_1, \ldots, y_N]$ and $S_i \in R^{N \times (k+1)}$ is the selection matrix. The S_i is defined according to [50] and [51] as

$$(S_i)_{pq} = \begin{cases} 1, & if \ \ p = F_i(q) \\ 0, & else \end{cases} \tag{15}$$

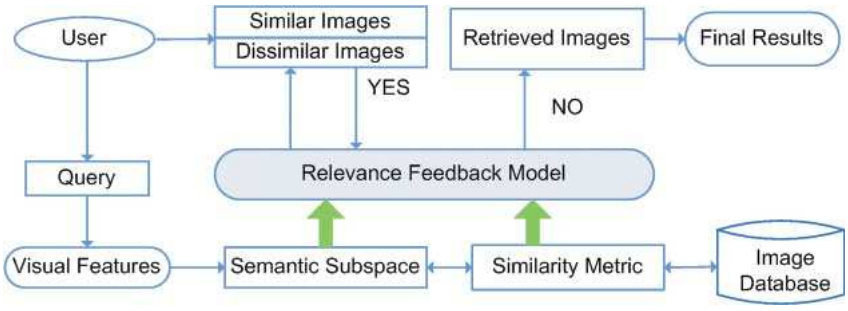


Fig. 3. Framework of our CBIR system. Our system can return the most semantically similar and dissimilar images for the user to label in RF.

where $F_i = [i, i_1, \ldots, i_k]$ is the index vector for samples in Y_i. Then, we can combine all the patches defined in Eq. (7), Eq.(12) and Eq.(14), together as follows:

$$\begin{aligned}
&\sum_{i=1}^{n} \min tr\left(Y_{g(i)} L_{g(i)} Y_{g(i)}^T\right) + \sum_{i=1}^{n} \max tr\left(Y_{d(i)} L_{d(i)} Y_{d(i)}^T\right) \\
&\quad + \sum_{i=1}^{n+n_u} \min tr\left(Y_{u(i)} L_{u(i)} Y_{u(i)}^T\right) \\
&= \sum_{i=1}^{n} \min tr\left(Y_{g(i)} L_{g(i)} Y_{g(i)}^T\right) - \gamma \sum_{i=1}^{n} \min tr\left(Y_{d(i)} L_{d(i)} Y_{d(i)}^T\right) \\
&\quad + \lambda \sum_{i=1}^{n+n_u} \min tr\left(Y_{u(i)} L_{u(i)} Y_{u(i)}^T\right) \\
&= \min tr\left(Y\left(\sum_{i=1}^{n} S_{g(i)} L_{g(i)} (S_{g(i)})^T - \gamma \sum_{i=1}^{n} S_{d(i)} L_{d(i)} (S_{d(i)})^T \right.\right. \\
&\qquad \left.\left. + \lambda \sum_{i=1}^{n+n_u} S_{u(i)} L_{u(i)} (S_{u(i)})^T\right) Y^T\right) \\
&= \min tr\left(W^T X\left(G - \gamma D + \lambda U\right) X^T W\right), \qquad (16)
\end{aligned}$$

where G encodes the local geometric information of labeled similar images and

$$G = \sum_{i=1}^{n} S_{g(i)} L_{g(i)} (S_{g(i)})^T;$$

D encodes the local discriminative information and

$$D = \sum_{i=1}^{n} S_{d(i)} L_{d(i)} (S_{d(i)})^T,$$

U encodes the local information of unlabeled images and

$$U = \sum_{i=1}^{n+n_u} S_{u(i)} L_{u(i)} (S_{u(i)})^T,$$

and $\gamma, \lambda > 0$ are tuning parameters that are used to tradeoff the contributions of three different terms.

By imposing $W^T W = I$, the mapping matrix $W = [w_1, \ldots, w_l]$ can be obtained by solving the standard eigendecomposition problem

$$X L X^T w = \lambda w, \qquad (17)$$

where W consists of the eigenvectors corresponding to the l largest eigenvalues.

Fig. 4. Some example images in the Corel image database.

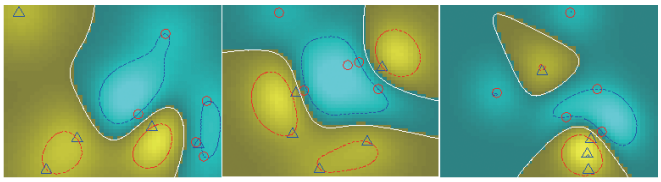
IV. Content-Based Image Retrieval System

A. Overview of Our CBIR Framework

In this section, we first give an overview of our CBIR system. As shown in Fig. 3, when a query image is provided, the low-level visual features are first extracted. Then, all image 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, in most situations, RF is actually required because of the poor performance of the system. The CBIR requires the user to label some semantically similar and dissimilar images as the positive and negative feedback samples, respectively. Using these labeled similar and dissimilar 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 sorted 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 the most semantically similar images with the query image. Otherwise, RF is performed iteratively. Some example images in the Corel photo gallery are shown in Fig. 4.

B. Corel Image Database and Image Representations

To perform an empirical evaluation of the proposed method, we first require 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 few years [3], [15], [16], [39]. To validate the effectiveness of the proposed algorithm, we group the images into a number of classes based on the provided 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 same category while many images with the same concept are in different categories. Therefore, existing categories of the original Corel



(a) (b) (c)

Fig. 5. RF models (i.e., the SVM hyperplane) are diverse with different semantic subspaces of feedback samples. The open circles and triangles denote the positive feedback samples and the negative feedback samples, respectively. The white solid line indicates the optimal hyperplane of SVM, which separates the positive and negative feedback samples.

photo gallery are ignored and reorganized into 80 conceptual classes based on the ground truth, such as lion, castle, 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. As a result, the Corel image database comprises totally 10,763 real-world images. This way of using the images with semantic categories can help evaluate the retrieval performance automatically, which significantly reduces subjective errors compared to manual evaluations.

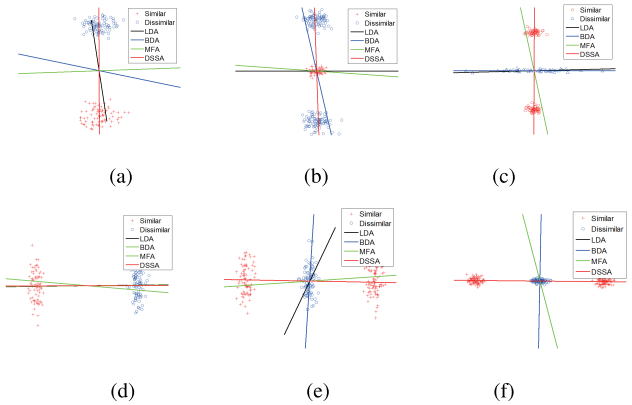
To represent images in the database, we use three different sets of low-level visual features in a 503-D space, i.e., 128-D RGB color histogram, 75-D edge distribution histogram and 300-D Bag-of-words (BOW) [55]. For the generation of visual words, we briefly apply the difference of Gaussians filter on the gray scale image to detect a set of salient points; then we compute the Scale-Invariant-Feature-Transform (SIFT) feature over the local areas defined by the detected salient points [56]; 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) The RF Models Are Diverse With Different Semantic Subspaces of Low-Level Visual Features: In RF, an image is usually represented by a high-dimensional low-level visual feature vector in the CBIR research. However, one key issue is about which subset of visual features can reflect the semantic properties of different groups of feedback samples and benefit the construction of RF models. This problem can be illustrated from some real-world samples in RF. There are five positive feedback samples and five negative feedback samples. We randomly select two features to construct the optimal RF model (i.e., SVM RF) for three times. As shown in Fig. 5, we can see that the resultant RF models are diverse with different semantic subspaces of visual features. And thus, selecting an effective semantic subspace and defining an effective similarity metric for the feedback samples are important steps in RF.

2) The DSSA Is Effective in Dealing With the Feedback Samples With Similar and Dissimilar Pairwise Constraints in RF: To visualize the effectiveness of DSSA in seeking the discriminative semantic subspace with similar and dissimilar



(a) (b) (c)

(d) (e) (f)

Fig. 6. Performance comparison of four different subspace analysis approaches (i.e., DSSA, LDA, BDA and MFA) for feedback samples with different distributions. (a)-(f) show the experimental results of four subspace analysis approaches when dealing with feedback samples with various distributions, respectively.

pairwise constraints in RF, this experiment is conducted on six synthetic datasets. In each round of RF, the user judges a set of images with similar and dissimilar pairwise constraints, which are positive and negative feedback samples, respectively. The positive and negative feedback samples are generated with different distributions, since the distributions of feedback samples are usually complicated in real-world applications. Regarding the set of positive feedback samples and the set of negative feedback samples as two different classes, LDA treats the two different sets of feedback samples equally. BDA was proposed to formulate the RF as a (1+x) class subspace analysis problem. However, it is still not very reasonable to conclude that all positive feedback samples come from one single class. Actually, each positive feedback sample is similar to each of the remaining positive feedback samples, and each negative feedback sample is dissimilar to each of the positive feedback sample. Consequently, different from conventional supervised subspace analysis methods (e.g., LDA and BDA), RF is intrinsically a weakly supervised learning problem and can only involve similar and dissimilar pairwise constraints for feedback samples. Any unreasonable assumption for the class label information of feedback samples will result in performance degradation.

From Fig. 6, we can clearly notice that LDA can find the best discriminative direction only when the set of positive feedback samples and the set of the negative feedback samples are distributed as Gaussians with similar covariance matrices, as shown in Fig.6 (a), but may be confused when the distribution of the feedback samples is more complicated, as given in Figs. 6 (b), (c), (d), (e), (f). Regarding the RF as a (1+x) class problem, BDA can only find the direction in which the positive feedback samples are well separated with the negative feedback samples when the positive feedback samples have a Gaussian distribution, e.g., Fig. 6 (b). However, BDA may also be confused when the distribution of positive feedback samples is more complicated, as shown in Figs. 6 (c), (e), (f). Marginal Fisher Analysis (MFA) defines the separation of positive and negative feedback samples with the marginal samples of different classes [23]. However, MFA treats the two different classes equally. The DSSA method only

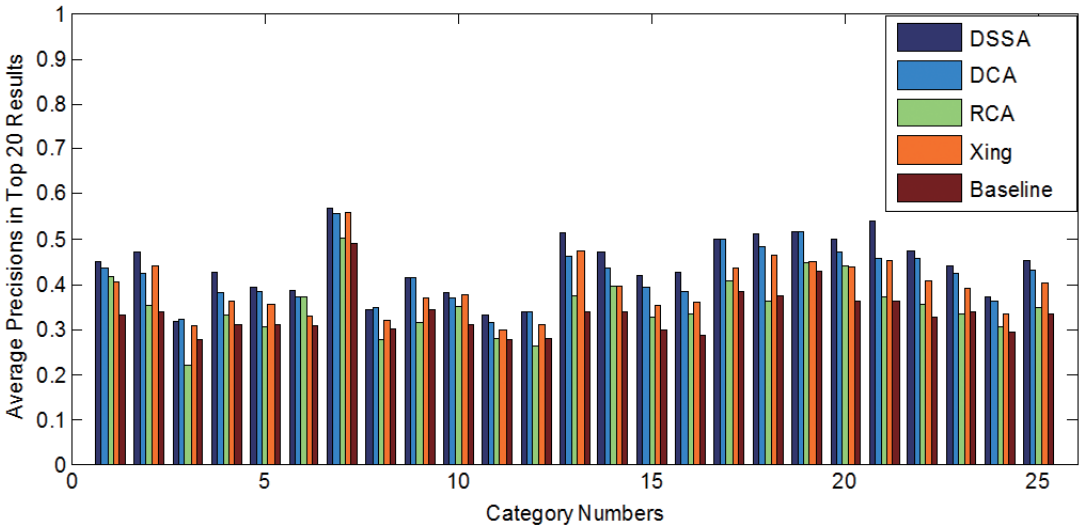


Fig. 7. APs with 25 categories in top 20 results of the compared algorithms (i.e., DSSA, DCA, RCA, Xing, and Baseline) based on the small-scale image database.

involves the local similar and dissimilar pairwise constraints of feedback samples and does not impose any label constraints on feedback samples, which is more appropriate for RF. Consequently, the DSSA can effectively find the most discriminative subspace compared with classical supervised subspace analysis methods (e.g., LDA, BDA, and MFA) with explicit class label information.

B. Experiments on the CBIR System

In this subsection, we will evaluate the effectiveness of the proposed DSSA in RF based on two experiments: first, we investigate the DSSA method for a CIR task [15], [35] by comparing it with a number of representative distance metric analysis methods; then, we show the performance of our CBIR system by comparing it with some popular RF approaches for an image retrieval task based on a real-world image database.

We use the widely used average precision (AP) and standard deviation (SD) to evaluate the performance of the compared algorithms. AP refers to the percentage of similar images in top ranked images presented to the user and is calculated as the averaged values of all the queries. SD is used to measure the amount of variation of APs. AP is the major evaluation criterion, which evaluates the effectiveness of the compared algorithms.

1) Performance Evaluation on a Small-Scale Image Database: In this part, we intend to examine whether the proposed method is comparable to or better than previous distance metric analysis techniques with similar and dissimilar constraints. We compare the proposed DSSA method with the Euclidean distance metric and three representative distance metric analysis methods (i.e., RCA [34], DCA [36] and Xing [33]). In our experiments, we do not compare the proposed method with supervised learning techniques since they require explicit class labels, which are not suitable for this task. Moreover, in this subsection, the DSSA method does not involve any unlabeled samples for fair comparison with RCA, DCA, and Xing. Parameters in each method were determined empirically to achieve its best performance in this paper. The parameter sensitivity of the DSSA method will be analyzed carefully in the next subsection.

In our experiments, to conduct objective evaluation and effectively investigate the performance of the proposed method, we have to provide a reliable image database with similar and dissimilar constraints to run these algorithms. It is not difficult to build a user historical feedback log database based on an existing real-world database, e.g., Corel image database. Here, we randomly select 25 classes according to the ground truth of images from the Corel image database and form a user historical feedback log database, which contains 2,497 real-world images. We randomly select 20 images uniformly from each class, and therefore, we can gather a user historical feedback database with 500 log images. Similar constraints are imposed on the images within the same class, while dissimilar constraints are imposed on the images with different classes. All 2,497 images in the 25 categories are used as the query images to evaluate the compared algorithms.

Fig. 7 shows the experimental results of the compared algorithms on the database with 500 log images. From the results, we can draw several observations. First, we notice that directly using the Euclidean distance metric in a high-dimensional visual feature space is not proper because of the semantic gap. All of the distance metric analysis methods (i.e., RCA, DCA, Xing and DSSA) can perform better than the baseline (i.e., Euclidean distance metric) by exploiting the user historical feedback log data. In the experiments, the optimal metric learned by RCA is computed as the inverse of the average covariance matrix of the chunklets. RCA will encounter the singular covariance matrix when dealing with high-dimensional images. The RCA is preceded by constraints-based LDA, which reduces the dimension to that of the DSSA method. By doing this, we notice that the RCA can show much better performance than the Euclidean distance metric by exploiting similar pairwise constraints. The DCA incorporates the dissimilar constraints into the RCA and was formulated into a trace ratio problem. In [36], the authors proposed to attack this problem by using a direct

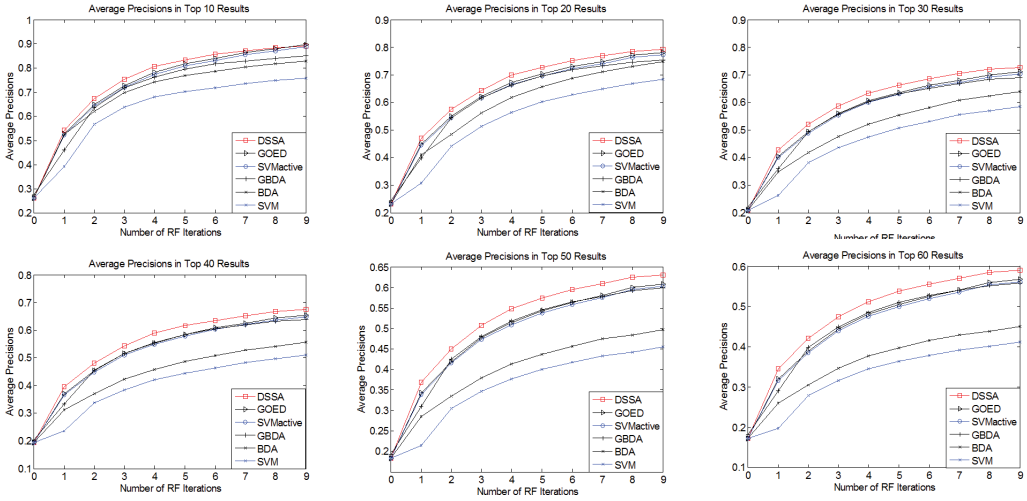


Fig. 8. APs of DSSA compared with conventional RF approaches, i.e., GOED, SVMactive, GBDA, SVM and BDA.

method as in Fisher's LDA. However, much discriminative information in the null space of the dissimilar scatter has been discarded in solving this problem. The DSSA can learn a distance metric by resorting to the mapping matrix and solve this function with a standard Eigen value decomposition method, which is very effective and efficient when handling high-dimensional images and never meets the problem of numerical computation. From the results, we can see that the proposed DSSA can significantly outperform the Euclidean distance metric and three compared metric learning approaches for overall evaluation.

2) Performance Evaluation on a Large-Scale Image Database: In this part, we design a slightly different scheme to model the real-world image retrieval process. In the real-world CBIR 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 an equal size. Therefore, there are 20 percent of the categories 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, nine rounds of RF are automatically conducted by the system.

To show the effectiveness of the proposed DSSA, we compare it with the popular conventional RF methods, i.e., geometric optimum experimental design (GOED) [16], SVMactive [17], SVM [9], BDA [22], and generalized BDA [39]. Out of these four algorithms, GOED and SVMactive are active learning methods, whereas SVM is a standard classification-based scheme, 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 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 sequentially to mark as the positive or negative feedback samples. In general, in a real-world image retrieval system, the dissimilar images usually largely outnumber the similar ones. To simulate such a case in the system, the first three similar images are labeled as positive feedback samples, and all other dissimilar images in the top 20 images are automatically labeled as the negative feedback samples. The images that have been selected in previous RF iterations are excluded from later sections. It should be noted that, for active learning-based RF methods, the 20 images are selected from the algorithms themselves, whereas for conventional classification-based RF methods (i.e., SVM) and discriminant analysis-based RF methods (i.e., GDBA 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 top 10 to top 60 to obtain the APs and the SDs 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, DSSA consistently outperforms all the other compared algorithms on the entire scope. SVMactive cannot show better performance, since the optimal hyperplane of SVM is usually not very stable and accurate with small-sized training data in a high-dimensional space. 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. 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 the experiments, we use the standard SVM to build an initial classifier. When considering more rounds of RF, SVMactive can get some improvement over the standard SVM.

Regarding the stability of the compared algorithms, we can also notice that DSSA performs best among the top 10,

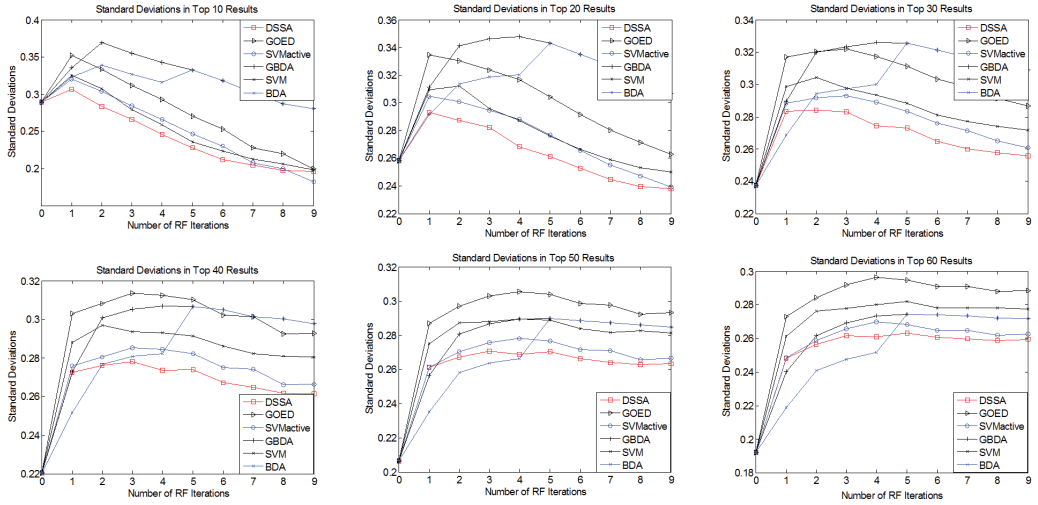


Fig. 9. SDs of DSSA compared with conventional RF approaches, i.e., GOED, SVMactive, SVM, BDA and GBDA.

TABLE I

AVERAGE PRECISIONS IN TOP N RESULTS OF THE COMPARED METHODS (i.e., DSSA, GOED, SVMACTIVE, GBDA SVM AND BDA)

Methods	DSSA	GOED	SVMactive	GBDA	SVM	BDA
Top 10	0.88±0.20	**0.89±0.21**	0.87±0.20	0.82±0.29	0.80±0.21	0.75±0.28
Top 20	**0.78±0.24**	0.77±0.20	0.77±0.25	0.73±0.32	0.71±0.26	0.67±0.32
Top 30	**0.72±0.26**	0.70±0.23	0.69±0.27	0.62±0.31	0.65±0.28	0.57±0.31
Top 40	**0.67±0.26**	0.65±0.20	0.64±0.27	0.54±0.30	0.60±0.28	0.50±0.30
Top 50	**0.62±0.27**	0.61±0.24	0.59±0.27	0.48±0.29	0.57±0.28	0.44±0.29
Top 60	**0.58±0.26**	0.57±0.20	0.56±0.26	0.44±0.27	0.53±0.28	0.40±0.27
Top 70	**0.55±0.25**	0.54±0.24	0.52±0.26	0.40±0.26	0.49±0.27	0.37±0.26
Top 80	**0.52±0.24**	0.50±0.21	0.49±0.24	0.37±0.25	0.46±0.26	0.34±0.24
Top 90	**0.48±0.24**	0.46±0.23	0.46±0.24	0.35±0.23	0.44±0.25	0.32±0.23

20, 30, and 40 results as shown in Fig. 9. Then, for other top results, the performance of DSSA is similar to the other compared algorithms. The detailed results of the compared algorithms after nine rounds of RF are shown in Table I. As given in Table I, DSSA achieves much better performance compared with other approaches for all top results. Therefore, we can conclude that the proposed DSSA has shown its effectiveness in learning an effective discriminative semantic subspace in RF.

In our experiments, the mapping matrix W can be obtained by using the Eigen value decomposition. The time cost to calculate W is $O\left((n+n_u)\right)^3$. Afterwards, we project all images to this semantic subspace and then apply the new similarity metric with respect to the query to sort all images in the database. The time cost for calculating the Euclidean distance in the semantic subspace L between the query and all images in the database is $O\left((NL)\right)$, where N is the cardinality of the database. Therefore, for a query image, the time cost for the DSSA based CBIR system is $O\left((n+n_u)^3\right)+O\left(NL\right)$. And the time cost for a conventional CBIR system in the high dimensional visual feature space H is $O\left(NH\right)$. Usually, for a CBIR system, the cardinality of the database N is very large and $H \gg L$; therefore, the proposed method is very efficient for an image retrieval task.

C. Parameter Sensitivity

In this subsection, we study the parameter sensitivity of the DSSA method for an image retrieval task. The analyses are performed based on the experiments conducted on the small-scale image database (i.e., 2,497 real-world image database). We analyze the trade-off parameter γ in Eq. (16), and the dimension of the projected features for DSSA. First, 500 query images are randomly selected from the database, and then the image retrieval process is automatically done by a computer. The APs in top 50 results are used for the overall performance evaluation.

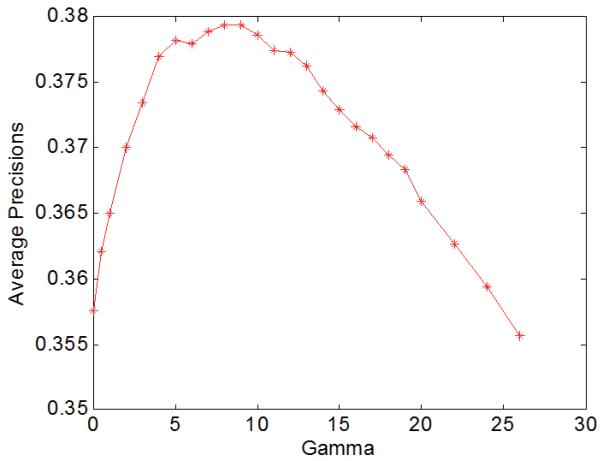


Fig. 10. Performance of DSSA with different γ values for the small-scale image database.

1) Evaluation of the Tradeoff Parameter γ: Empirically, the local geometry is useful for finding the semantic subspace. In this part, we intend to investigate the influence of the tradeoff parameter γ in Eq. (16) for DSSA when building the local discriminative patches and the local geometric patches for labeled log images. A small γ reflects the importance of separating dissimilar samples from similar ones, i.e., the DSSA focuses on the local geometric information and ignores the local discriminative information. Fig. 10 shows the performance of DSSA by varying γ, from which we can have the following observations.

When γ is small, e.g., $\gamma = 0$, the performance is unsatisfactory. This is because in this situation the local information is mainly preserved while important local discriminative information within labeled images with similar and dissimilar

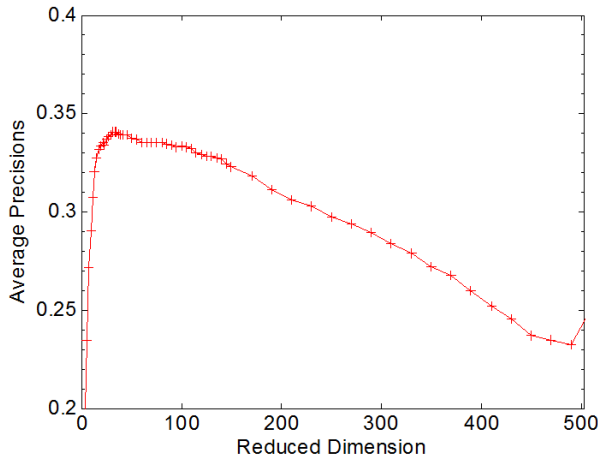


Fig. 11. Performance of DSSA with features projected onto the subspaces with different dimensions on the small-scale image database.

pairwise constraints is less considered. The performance of the DSSA increases when γ increases its value and reaches the optimal value at $\gamma = 5$. Then, the APs decrease when γ is larger than this best setup, in which case the local discriminative information dominates the local patches and the local geometric information is ignored.

Therefore, both the local geometry and the discriminative information can reflect the important information contained in the local patches from different aspects and are complementary. A suitable combination of them is essential to achieve good performance of DSSA.

2) Evaluation on the Projected Subspace: Different from the distance metric analysis methods, the proposed DSSA method aims to learn a mapping matrix that can find a low-dimensional subspace from the original high-dimensional space. To find an appropriate dimension of the projected semantic subspace, we investigate the influence of the dimension in this experiment. Fig. 11 shows the performance of DSSA with features projected onto the subspaces with different dimensions. From Fig. 11, we can notice that, when the projected dimension is too low (e.g., less than 25), the reduced subspace is insufficient to encode the semantic concepts of images, which makes the retrieval performance poor. When the dimension equals or is close to that of the original high-dimensional space (i.e., 503 in this paper), no or less benefit can be obtained from this subspace analysis method. From the experimental results, we can notice that the DSSA method achieves its best performance with the dimension of 25 for the small-scale image database. Moreover, low-dimensional data can lead to lower computation costs than higher-dimensional data for an image retrieval task.

D. Discussions and Future Work

In the proposed CBIR 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 Flikr images) has attracted much attention. In these systems, the images are first selected from social web sites (e.g., Flickr), most of which are accompanied by rich surrounding textual description (e.g., tags). And then, these images are grouped into plenty of semantic groups according to the associated textual descriptions. However, different users have different opinions on the same image, and thus will annotate significantly different textual information. 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 future studies.

To enhance the retrieval performance, the indexing of database is very important for a CBIR system. Generally, there are two types of image indexing methods [1], [2]. A classification based indexing technique aims to improve the retrieval precision of the system [57]. In this method, each image in the database is assigned one or more distinct labels. Then, based on these labels, indexing the database can be constructed through their associated semantic labels. Therefore, the search results will be more satisfactory for most of the users. The other indexing method is the low level visual feature based indexing [58], which can be used to speed up the retrieval procedure. There are many low level visual feature based indexing techniques, e.g., various tree-based indexing structures for high dimensional data. The two indexing methods have their respective advantages from different aspects. As a consequence, it is promising to combine the classification and visual feature information in the indexing structures to improve both the retrieval precision and speed.

VI. Conclusion

In this paper, we have proposed a novel discriminative semantic subspace analysis (DSSA) method to bridge the gap between low-level visual features and high-level semantic concepts by exploiting the training images with pairwise constraints in RF. The proposed DSSA method can effectively learn a reliable subspace both from labeled and unlabeled images with similar and dissimilar pairwise constraints without using any explicit class label information. Especially, DSSA can effectively integrate the local geometry of labeled similar images, the discriminative information of labeled similar and dissimilar images and the local geometry of labeled and unlabeled images together to learn a reliable subspace. Compared with the popular distance metric analysis methods with pairwise constraints, our method can also learn a distance metric but perform more effectively when dealing with high-dimensional images. Extensive experiments on both synthetic datasets and the real-world Corel image database have shown the effectiveness of the proposed scheme in exploiting the training images with pairwise constraints in RF.

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 Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1349–1380, Dec. 2000.

[2] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Comput. Surv.*, vol. 40, no. 2, May 2008, Art. ID 5.

[3] 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.

[4] 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.

[5] L. Liu, M. Yu, and L. Shao, "Multiview alignment hashing for efficient image search," *IEEE Trans. Image Process.*, vol. 24, no. 3, pp. 956–966, Mar. 2015.

[6] 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.

[7] Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra, "Relevance feedback: A power tool for interactive content-based image retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, no. 5, pp. 644–655, Sep. 1998.

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

[9] Y. Chen, X. S. Zhou, and T. S. Huang, "One-class SVM for learning in image retrieval," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2001, pp. 34–37.

[10] P. Hong, Q. Tian, and T. S. Huang, "Incorporate support vector machines to content-based image retrieval with relevance feedback," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2000, pp. 750–753.

[11] D. Zhang, M. M. Islam, and G. Lu, "A review on automatic image annotation techniques," *Pattern Recognit.*, vol. 45, no. 1, pp. 346–362, 2012.

[12] S. C. H. Hoi, M. R. Lyu, and R. Jin, "A unified log-based relevance feedback scheme for image retrieval," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 4, pp. 509–524, Apr. 2006.

[13] L. Si, R. Jin, S. C. H. Hoi, and M. R. Lyu, "Collaborative image retrieval via regularized metric learning," *Multimedia Syst.*, vol. 12, no. 1, pp. 34–44, 2006.

[14] Y. Liu, D. Xu, I. W. Tsang, and J. Luo, "Textual query of personal photos facilitated by large-scale Web data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 1022–1036, May 2011.

[15] 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.

[16] 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.

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

[19] S. 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.

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

[21] S. C. H. Hoi, R. Jin, J. Zhu, and M. R. Lyu, "Semi-supervised SVM batch mode active learning for image retrieval," *ACM Trans. Inf. Syst.*, vol. 27, no. 3, May 2009, Art. ID 16.

[22] X. S. Zhou and T. S. Huang, "Small sample learning during multimedia retrieval using BiasMap," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2001, pp. I-11–I-17.

[23] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 40–51, Jan. 2007.

[24] S. Si, D. Tao, and B. Geng, "Bregman divergence-based regularization for transfer subspace learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 7, pp. 929–942, Jul. 2010.

[25] D. Tao, X. Li, X. Wu, and S. J. Maybank, "General tensor discriminant analysis and Gabor features for gait recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 10, pp. 1700–1715, Oct. 2007.

[26] M. Yu, L. Liu, and L. Shao, "Structure-preserving binary representations for RGB-D action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published. DOI: 10.1109/TPAMI.2015.2491925

[27] L. Shao, L. Liu, and M. Yu, "Kernelized multiview projection for robust action recognition," *Int. J. Comput. Vis.*, to be published. DOI: 2015.10.1007/s11263-015-0861-6

[28] M. Yu, L. Shao, X. Zhen, and X. He, "Local feature discriminant projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published. DOI: 10.1109/TPAMI.2015.2497686

[29] C. Xu, D. Tao, and C. Xu, "Multi-view intact space learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 12, pp. 2531–2544, Dec. 2015.

[30] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[31] X. He and P. Niyogi, "Locality preserving projections," in *Proc. Adv. Neural Inf. Process. Syst.*, 2004, pp. 153–160.

[32] D. Tao, X. Li, X. Wu, and S. J. Maybank, "Geometric mean for subspace selection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 260–274, Feb. 2009.

[33] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. J. Russell, "Distance metric learning with application to clustering with side-information," in *Proc. Adv. Neural Inf. Process. Syst.*, 2003, pp. 521–528.

[34] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall, "Learning a Mahalanobis metric from equivalence constraints," *J. Mach. Learn. Res.*, vol. 6, no. 1, pp. 937–965, 2006.

[35] S. 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, 2010, Art. ID 18.

[36] S. C. H. Hoi, W. Liu, M. R. Lyu, and W.-Y. Ma, "Learning distance metrics with contextual constraints for image retrieval," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2006, pp. 2072–2078.

[37] Y. Luo, T. Liu, D. Tao, and C. Xu, "Decomposition-based transfer distance metric learning for image classification," *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 3789–3801, Sep. 2014.

[38] F. Zhu and L. Shao, "Weakly-supervised cross-domain dictionary learning for visual recognition," *Int. J. Comput. Vis.*, vol. 109, nos. 1–2, pp. 42–59, 2014.

[39] 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.

[40] 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.

[41] D. Tao and X. Tang, "Kernel full-space biased discriminant analysis," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jun. 2004, pp. 1287–1290.

[42] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *J. Mach. Learn. Res.*, vol. 10, pp. 207–244, Feb. 2009.

[43] J. V. Davis and I. S. Dhillon, "Structured metric learning for high dimensional problems," in *Proc. 14th ACM Int. Conf. Knowl. Discovery Data Mining*, 2008, pp. 195–203.

[44] M. Guillaumin, J. Verbeek, and C. Schmid, "Is that you? Metric learning approaches for face identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 498–505.

[45] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2288–2295.

[46] W.-S. Zheng, S. Gong, and T. Xiang, "Person re-identification by probabilistic relative distance comparison," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 649–656.

[47] L. Wu, R. Jin, C. H. Hoi, J. Zhu, and N. Yu, "Learning bregman distance functions and its application for semi-supervised clustering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 2089–2097.

[48] A. Ghodsi, D. Wilkinson, and F. Southey, "Improving embeddings by flexible exploitation of side information," in *Proc. 20th Int. Joint Conf. Artif. Intell.*, 2007, pp. 810–816.

[49] Z. Zhang and H. Zha, "Principal manifolds and nonlinear dimension reduction via local tangent space alignment," *SIAM J. Sci. Comput.*, vol. 26, no. 1, pp. 313–338, 2002.

[50] T. Zhang, D. Tao, X. Li, and J. Yang, "Patch alignment for dimensionality reduction," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1299–1313, Sep. 2009.

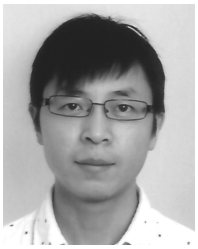
[51] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[52] D. L. Donoho and C. Grimes, "Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data," *Proc. Nat. Acad. Sci. USA*, vol. 100, no. 10, pp. 5591–5596, 2003.

[53] 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.
[54] 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.
[55] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: A comprehensive study," *Int. J. Comput. Vis.*, vol. 73, no. 2, pp. 213–238, Jun. 2007.
[56] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
[57] N. Vasconcelos, "Image indexing with mixture hierarchies," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Dec. 2001, p. 1.
[58] A. Natsev, R. Rastogi, and K. Shim, "WALRUS: A similarity retrieval algorithm for image databases," *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 3, pp. 301–316, Mar. 2004.

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

Hubert P. H. Shum received the B.Eng. and M.Sc. degrees from the City University of Hong Kong, and the Ph.D. degree from the School of Informatics, 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 a Senior Lecturer (Associate Professor) 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 Department of Computer Science and Digital Technologies at Northumbria University, Newcastle Upon Tyne, U.K. Previously, he was a Senior Lecturer (2009- 2014) with the Department of Electronic and Electrical Engineering at 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 CYBERNETICS and several other journals. He is a Fellow of the British Computer Society and the Institution of Engineering and Technology.