Semi-supervised Discriminant Analysis*

Deng Cai
UIUC
dengcai2@cs.uiuc.edu

Xiaofei He
Yahoo!
hex@yahoo-inc.com

Jiawei Han
UIUC
hanj@cs.uiuc.edu

Abstract

Linear Discriminant Analysis (LDA) has been a popular method for extracting features which preserve class separability. The projection vectors are commonly obtained by maximizing the between class covariance and simultaneously minimizing the within class covariance. In practice, when there is no sufficient training samples, the covariance matrix of each class may not be accurately estimated. In this paper, we propose a novel method, called Semi-supervised Discriminant Analysis (SDA), which makes use of both labeled and unlabeled samples. The labeled data points are used to maximize the separability between different classes and the unlabeled data points are used to estimate the intrinsic geometric structure of the data. Specifically, we aim to learn a discriminant function which is as smooth as possible on the data manifold. Experimental results on single training image face recognition and relevance feedback image retrieval demonstrate the effectiveness of our algorithm.

1. Introduction

In many visual analysis applications, such as image retrieval, face recognition, etc., one is often confronted with high-dimensional data. However, there might be reason to suspect that the naturally generated high-dimensional data probably resides on a lower dimensional manifold. This leads one to consider methods of dimensionality reduction that allow one to represent the data in a lower dimensional space. Two of the most popular techniques for this purpose are Principal Component Analysis (PCA) [16] and Linear Discriminant Analysis (LDA) [9].

PCA is an unsupervised method. It performs dimensionality reduction by projecting the original n-dimensional data onto the $d(\ll n)$-dimensional linear subspace spanned by the leading eigenvectors of the data's covariance matrix. Its goal is to find a set of mutually orthogonal basis functions that capture the directions of maximum variance in the data so that the pairwise *Euclidean* distances can be best preserved. If the data is embedded in a linear subspace, PCA is guaranteed to discover the dimensionality of the subspace and produces a compact representation.

LDA is a supervised method. It searches for the project axes on which the data points of different classes are far from each other while requiring data points of the same class to be close to each other. When label information available, *e.g.*, for classification task, LDA can achieve significant better performance than PCA [1]. However, when there is no sufficient training samples relative to the number of dimensions, the covariance matrix of each class may not be accurately estimated. In this case, the generalization capability on testing samples can not be guaranteed. A possible solution to deal with insufficient training (labeled) samples could be learning on both labeled and unlabeled data (*semi-supervised* and *transductive* learning). It is natural and reasonable since in reality we usually have only part of input data labeled, along with a large number of unlabeled data.

In the last decades, semi-supervised learning (or transductive learning) has attracted an increasing amount of attention. Two well known algorithms are Transductive SVM (TSVM) [23] and Co-Training . Recently, there are considerable interest and succuss on graph based semi-supervised learning algorithms [3, 20, 26, 27], which consider the graph over all the samples as a prior to guide the decision making. All these algorithms considered the problem of classification, either transductive or inductive.

In this paper, we aim at dimensionality reduction in the semi-supervised case. We proposed a semi-supervised dimensionality reduction algorithm, called **Semi-supervised Discriminant Analysis** (SDA). SDA aims to find a projection which respects the discriminant structure inferred from the labeled data points, as well as the intrinsic geometrical structure inferred from both labeled and unlabeled data points. Specifically, the labeled data points, combined with

*The work was supported in part by the U.S. National Science Foundation NSF IIS-05-13678, NSF BDI-05-15813 and MIAS (a DHS Institute of Discrete Science Center for Multimodal Information Access and Synthesis). Any opinions, findings, and conclusions or recommendations expressed here are those of the authors and do not necessarily reflect the views of the funding agencies.

the unlabeled data points, are used to build a graph incorporating neighborhood information of the data set. The graph provides a discrete approximation to the local geometry of the data manifold. Using the notion of graph Laplacian, a smoothness penalty on the graph can be incorporated into the objective function. In this way, our SDA algorithm can optimally preserves the manifold structure.

The rest of this paper is organized as follows. In Section 2, we provide a brief review of LDA. We introduce our Semi-supervised Discriminant Analysis (SDA) algorithm for dimensionality reduction in Section 3. The experimental results are presented in Section 4. Finally, we conclude the paper and provide suggestions for future work in Section 5.

2. Graph Perspective of LDA

Linear Discriminant Analysis (LDA) seeks directions on which the data points of different classes are far from each other while requiring data points of the same class to be close to each other [9]. Suppose we have a set of l samples $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_l \in \mathbb{R}^n$, belonging to c classes. The objective function of LDA is as follows:

$$\mathbf{a}_{opt} = \arg\max_{\mathbf{a}} \frac{\mathbf{a}^T S_b \mathbf{a}}{\mathbf{a}^T S_w \mathbf{a}}, \tag{1}$$

$$S_b = \sum_{k=1}^{c} l_k (\boldsymbol{\mu}^{(k)} - \boldsymbol{\mu})(\boldsymbol{\mu}^{(k)} - \boldsymbol{\mu})^T, \tag{2}$$

$$S_w = \sum_{k=1}^{c} \left(\sum_{i=1}^{l_k} (\mathbf{x}_i^{(k)} - \boldsymbol{\mu}^{(k)})(\mathbf{x}_i^{(k)} - \boldsymbol{\mu}^{(k)})^T \right), \tag{3}$$

where $\boldsymbol{\mu}$ is the total sample mean vector, l_k is the number of samples in the k-th class, $\boldsymbol{\mu}^{(k)}$ is the average vector of the k-th class, and $\mathbf{x}_i^{(k)}$ is the i-th sample in the k-th class. We call S_w the within-class scatter matrix and S_b the between-class scatter matrix.

Define the total scatter matrix $S_t = \sum_{i=1}^{l} (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T$, we have $S_t = S_b + S_w$ [9]. The objective function of LDA in Eqn. (1) is equivalent to

$$\mathbf{a}_{opt} = \arg\max_{\mathbf{a}} \frac{\mathbf{a}^T S_b \mathbf{a}}{\mathbf{a}^T S_t \mathbf{a}}. \tag{4}$$

The optimal $\mathbf{a}$'s are the eigenvectors corresponding to the non-zero eigenvalue of eigen-problem:

$$S_b \mathbf{a} = \lambda S_t \mathbf{a}. \tag{5}$$

Since the rank of S_b is bounded by $c - 1$, there are at most $c - 1$ eigenvectors corresponding to non-zero eigenvalues [9].

Without loss of generality, we assume $\boldsymbol{\mu} = \mathbf{0}$[1]. We have

$$\begin{aligned} S_b &= \sum_{k=1}^{c} l_k (\boldsymbol{\mu}^{(k)})(\boldsymbol{\mu}^{(k)})^T \\ &= \sum_{k=1}^{c} l_k \left(\frac{1}{l_k} \sum_{i=1}^{l_k} \mathbf{x}_i^{(k)} \right) \left(\frac{1}{l_k} \sum_{i=1}^{l_k} \mathbf{x}_i^{(k)} \right)^T \\ &= \sum_{k=1}^{c} X^{(k)} W^{(k)} (X^{(k)})^T \end{aligned}$$

where $W^{(k)}$ is a $l_k \times l_k$ matrix with all the elements equal to $1/l_k$ and $X^{(k)} = [\mathbf{x}_1^{(k)}, \cdots, \mathbf{x}_{l_k}^{(k)}]$ denote the data matrix of k-th class.

Let the data matrix $X = [X^{(1)}, \cdots, X^{(c)}]$ and define a $l \times l$ matrix $W_{l \times l}$ as:

$$W_{l \times l} = \begin{bmatrix} W^{(1)} & 0 & \cdots & 0 \\ 0 & W^{(2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & W^{(c)} \end{bmatrix} \tag{6}$$

We have

$$S_b = \sum_{k=1}^{c} X^{(k)} W^{(k)} (X^{(k)})^T = X W_{l \times l} X^T. \tag{7}$$

Thus, the objective function of LDA in Eqn. (4) can be rewritten as

$$\mathbf{a}_{opt} = \arg\max_{\mathbf{a}} \frac{\mathbf{a}^T S_b \mathbf{a}}{\mathbf{a}^T S_t \mathbf{a}} = \arg\max_{\mathbf{a}} \frac{\mathbf{a}^T X W_{l \times l} X^T \mathbf{a}}{\mathbf{a}^T X X^T \mathbf{a}}. \tag{8}$$

This formulation of LDA objective function will be very helpful in developing our algorithm. It is first introduced in [14].

3. Semi-supervised Discriminant Analysis

LDA considers seeking the optimal projections purely on the training (labeled) set. In reality, it is possible to acquire a large set of unlabeled data. In this section, we are trying to extend LDA model to incorporate the manifold structure illustrated by unlabeled data.

3.1. The Objective Function

LDA aims to find a projection vector $\mathbf{a}$ such that the ratio between $\mathbf{a}^T S_b \mathbf{a}$ and $\mathbf{a}^T S_t \mathbf{a}$ is maximized. When there is no sufficient training sample, overfitting may happen. A typical way to prevent overfitting is to impose a regularizer

[1] This can be achieved by centering the data, *i.e.*, subtract the mean vector from all the sample vectors.

[11]. The optimization problem of the regularized version of LDA can be written as follows:

$$\max_{\mathbf{a}} \frac{\mathbf{a}^T S_b \mathbf{a}}{\mathbf{a}^T S_t \mathbf{a} + \alpha J(\mathbf{a})} \tag{9}$$

where $J(\mathbf{a})$ controls the learning complexity of the hypothesis family, and the coefficient α controls balance between the model complexity and the empirical loss. One of the most popular regularizers is the Tikhonov regularizer [21]:

$$J(\mathbf{a}) = \|\mathbf{a}\|^2.$$

LDA model with Tikhonov regularizer is usually referred as Regularized Discriminant Analysis (RDA) [8].

The regularizer term $J(\mathbf{a})$ provides us the flexibility to incorporate our prior knowledge on some particular applications. When a set of unlabeled examples available, we aim to construct a $J(\mathbf{a})$ incorporating the manifold structure. The key to semi-supervised learning algorithm is the prior assumption of consistency. For classification, it means nearby points are likely to have the same label [26]. For dimensionality reduction, it can be interpreted as nearby points will have similar embeddings (low-dimensional representations). Given a set of examples $\{\mathbf{x}_i\}_{i=1}^m$, we can use a p-nearest neighbor graph G to model the relationship between nearby data points. Specifically, we put an edge between nodes i and j if $\mathbf{x}_i$ and $\mathbf{x}_j$ are "close", *i.e.*, $\mathbf{x}_i$ and $\mathbf{x}_j$ are among p nearest neighbors of each other. Let the corresponding weight matrix be S, defined by

$$S_{ij} = \begin{cases} 1, & \text{if } \mathbf{x}_i \in N_p(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in N_p(\mathbf{x}_i) \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

where $N_p(\mathbf{x}_i)$ denotes the set of p nearest neighbors of $\mathbf{x}_i$. In general, the mapping function should be as smooth as possible on the graph. Specifically, if two data points are linked by an edge, they are likely to be in the same class. Moreover, the data points lying on a densely linked subgraph are likely to have the same label. Thus, a natural regularizer can be defined as follows:

$$J(\mathbf{a}) = \sum_{ij} \left(\mathbf{a}^T \mathbf{x}_i - \mathbf{a}^T \mathbf{x}_j\right)^2 S_{ij} \tag{11}$$

This formulation is motivated from spectral dimensionality reduction [2, 13], which also plays a key role in spectral clustering [17] and various kinds of graph based semi-supervised learning algorithms [3, 6, 20].

Let $X = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_m]$. We have

$$\begin{aligned} J(\mathbf{a}) &= \sum_{ij} (\mathbf{a}^T \mathbf{x}_i - \mathbf{a}^T \mathbf{x}_j)^2 S_{ij} \\ &= 2 \sum_i \mathbf{a}^T \mathbf{x}_i D_{ii} \mathbf{x}_i^T \mathbf{a} - 2 \sum_{ij} \mathbf{a}^T \mathbf{x}_i S_{ij} \mathbf{x}_j^T \mathbf{a} \\ &= 2\mathbf{a}^T X(D - S)X^T \mathbf{a} \\ &= 2\mathbf{a}^T X L X^T \mathbf{a} \end{aligned}$$

where D is a diagonal matrix; its entries are column (or row, since S is symmetric) sum of S, $D_{ii} = \Sigma_j S_{ij}$. $L = D - S$ is the Laplacian matrix [7].

With this data dependent regularizer, we get the objective function of our semi-supervised discriminant analysis:

$$\max_{\mathbf{a}} \frac{\mathbf{a}^T S_b \mathbf{a}}{\mathbf{a}^T \left(S_t + \alpha X L X^T\right) \mathbf{a}}. \tag{12}$$

The projective vector $\mathbf{a}$ that maximizes the objective function is given by the maximum eigenvalue solution to the generalized eigenvalue problem:

$$S_b \mathbf{a} = \lambda (S_t + \alpha X L X^T) \mathbf{a} \tag{13}$$

3.2. The Algorithm

Given a labeled set $\{(\mathbf{x}_i, y_i)\}_{i=1}^l$ belonging to c classes and an unlabeled set $\{\mathbf{x}_i\}_{i=l+1}^m$. The k-th class have l_k samples, $\sum_{k=1}^c l_k = l$. Without loss of generality, we assume that the data points in $\{\mathbf{x}_1, \cdots, \mathbf{x}_l\}$ are ordered according to their labels. The algorithmic procedure of semi-supervised discriminant analysis is stated below:

1. **Construct the adjacency graph**: Construct the p-nearest neighbors graph matrix S as in Eqn. (10) and calculate the graph Laplacian $L = D - S$.

2. **Construct the labeled graph**: Construct the weight matrix $W \in \mathbb{R}^{m \times m}$ for labeled graph as

   $$W = \begin{bmatrix} W_{l \times l} & 0 \\ 0 & 0 \end{bmatrix}$$

   where $W_{l \times l} \in \mathbb{R}^{l \times l}$ is defined in Eqn. (6). Define

   $$\widetilde{I} = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}$$

   where I is an identity matrix of size $l \times l$.

3. **Eigen-problem**: Compute the eigenvectors with respect to the non-zero eigenvalues for the generalized eigenvector problem:

   $$XWX^T \mathbf{a} = \lambda X(\widetilde{I} + \alpha L)X^T \mathbf{a}, \tag{14}$$

   where $X = [\mathbf{x}_1, \cdots, \mathbf{x}_l, \mathbf{x}_{l+1}, \cdots, \mathbf{x}_m]$. It is easy to check that W is of rank c and we will have c eigenvectors with respect to non-zero eigenvalues [2] [10]. We denote them as $\mathbf{a}_1, \cdots, \mathbf{a}_c$.

4. **SDA Embedding**: Let $A = [\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_c]$, A is a $n \times c$ transformation matrix. The samples can be embedded into c dimensional subspace by

   $$\mathbf{x} \to \mathbf{z} = A^T \mathbf{x}$$

[2] We consider the case that the number of features $n > c$.

Let $X_l = [\mathbf{x}_1, \cdots, \mathbf{x}_l]$ be the labeled data matrix. It is easy to check that

$$XWX^T = X_l W_{l\times l} X_l^T = S_b$$

and

$$X\widetilde{I}X^T = X_l X_l^T = S_t.$$

Thus, the eigen-problem in Eqn. (14) is same as the eigen-problem in Eqn. (13).

To get a stable solution of the eigen-problem in Eqn. (14), the matrix $X(\widetilde{I} + \alpha L)X^T$ is required to be non-singular [10] which is not true when the number of features is larger than the number of samples. In this case, we can apply the Tikhonov regularization idea as the way in regularized discriminant analysis [8]. Thus, our generalized eigen-problem becomes:

$$XWX^T\mathbf{a} = \lambda\Big(X(\widetilde{I} + \alpha L)X^T + \beta I\Big)\mathbf{a} \qquad (15)$$

For $\beta > 0$, the matrix $X(\widetilde{I}+\alpha L)X^T + \beta I$ is certainly non-singular. We can also use the spectral regression technique to solve this singularity problem, please see [5] for details.

3.3. Kernel SDA

The algorithm described above is a linear method. It may fail to discover the intrinsic geometry when the data manifold is highly nonlinear. In this subsection, we discuss how to perform SDA in Reproducing Kernel Hilbert Space (RKHS), which gives rise to kernel SDA. The approach used here is essentially similar to [13].

We consider the problem in a feature space $\mathcal{F}$ induced by some nonlinear mapping

$$\phi : \mathbb{R}^n \to \mathcal{F}$$

For a proper chosen ϕ, an inner product $\langle , \rangle$ can be defined on $\mathcal{F}$ which makes for a so-called reproducing kernel Hilbert space (RKHS). More specifically,

$$\langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle = \mathcal{K}(\mathbf{x}, \mathbf{y})$$

holds where $\mathcal{K}(.,.)$ is a positive semi-definite kernel function. Several popular kernel functions are: Gaussian kernel $\mathcal{K}(\mathbf{x}, \mathbf{y}) = \exp(-\|\mathbf{x} - \mathbf{y}\|^2/\sigma^2)$; polynomial kernel $\mathcal{K}(\mathbf{x}, \mathbf{y}) = (1 + \langle \mathbf{x}, \mathbf{y} \rangle)^d$; Sigmoid kernel $\mathcal{K}(\mathbf{x}, \mathbf{y}) = tanh(\langle \mathbf{x}, \mathbf{y} \rangle + \alpha)$.

Given a set of vectors $\{\mathbf{v}_i \in \mathcal{F} | i = 1, 2, \cdots, d\}$ which are orthonormal ($\langle \mathbf{v}_i, \mathbf{v}_j \rangle = \delta_{i,j}$), the projection of $\phi(\mathbf{x}_i) \in \mathcal{F}$ to these $\mathbf{v}_1, \cdots, \mathbf{v}_d$ leads to a mapping from $\mathbb{R}^n$ to Euclidean space $\mathbb{R}^d$ through

$$\mathbf{y}_i = \big(\langle \mathbf{v}_1, \phi(\mathbf{x}_i) \rangle, \langle \mathbf{v}_2, \phi(\mathbf{x}_i) \rangle, \cdots, \langle \mathbf{v}_d, \phi(\mathbf{x}_i) \rangle\big)^T$$

We look for such $\{\mathbf{v}_i \in \mathcal{F} | i = 1, 2, \cdots, d\}$ that helps $\{\mathbf{y}_i | i = 1, \cdots, m\}$ preserve local geometrical and discriminant structure of the data manifold.

Let Φ denote the data matrix in RKHS:

$$\Phi = [\phi(\mathbf{x}_1), \phi(\mathbf{x}_2), \cdots, \phi(\mathbf{x}_m)]$$

Now, the eigenvector problem of Eqn. (14) in RKHS can be written as follows:

$$\Phi W \Phi^T \mathbf{v} = \lambda \Phi \big(\widetilde{I} + \alpha L\big)\Phi^T \mathbf{v} \qquad (16)$$

Because the eigenvector of (16) are linear combinations of $\phi(\mathbf{x}_1), \phi(\mathbf{x}_2), \cdots, \phi(\mathbf{x}_m)$, there exist coefficients $\alpha_i, i = 1, 2, \cdots, m$ such that

$$\mathbf{v} = \sum_{i=1}^{m} \alpha_i \phi(\mathbf{x}_i) = \Phi\boldsymbol{\alpha}$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \cdots, \alpha_m)^T \in \mathbb{R}^m$.

Following some algebraic formulations, we get:

$$\begin{aligned} & \Phi W \Phi^T \mathbf{v} = \lambda \Phi \big(\widetilde{I} + \alpha L\big)\Phi^T \mathbf{v} \\ \Rightarrow \quad & \Phi W \Phi^T \Phi\boldsymbol{\alpha} = \lambda \Phi \big(\widetilde{I} + \alpha L\big)\Phi^T \Phi\boldsymbol{\alpha} \\ \Rightarrow \quad & \Phi^T \Phi W \Phi^T \Phi\boldsymbol{\alpha} = \lambda \Phi^T \Phi \big(\widetilde{I} + \alpha L\big)\Phi^T \Phi\boldsymbol{\alpha} \\ \Rightarrow \quad & KWK\boldsymbol{\alpha} = \lambda K \big(\widetilde{I} + \alpha L\big)K\boldsymbol{\alpha} \end{aligned} \qquad (17)$$

where K is the kernel matrix, $K_{ij} = \mathcal{K}(\mathbf{x}_i, \mathbf{x}_j)$. Let the column vectors $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \cdots, \boldsymbol{\alpha}_c$ be the eigenvectors with respect to the non-zero eigenvalues of eigen-problem in Eqn. (17) and the $m \times c$ transformation matrix $\Theta = [\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_c]$. A data point can be embedded into c dimensional subspace by

$$\mathbf{x} \to \mathbf{z} = \Theta^T K(:, \mathbf{x}) \qquad (18)$$

where $K(:, \mathbf{x}) \doteq [\mathcal{K}(\mathbf{x}_1, \mathbf{x}), \cdots, \mathcal{K}(\mathbf{x}_m, \mathbf{x})]^T$

4. Experimental Results

In this section, several experiments are performed to test our algorithm. We choose two scenarios in which semi-supervised learning is natural and necessary. They are single training image face recognition [4] and relevance feedback image retrieval [18].

Many of proposed graph based semi-supervised learning algorithms [26, 27] can only work on *transductive* setting. That is, both the training and test set (without label information) are available during the learning process. In reality (*e.g.*, face recognition), a more natural setting for semi-supervised learning is as follows. The available training set contains both labeled and unlabeled examples, and the testing set is not available during the training phrase, which we refer here as *semi-supervised setting*. To this end, manifold regularization [3, 20] is one of the most successful approaches that address both two settings. Manifold regularization extends many of the existing inductive algorithms (*e.g.*, SVM, Regression) to semi-supervised learning by adding a geometrically based regularization term.

Figure 1. Sample face images from the CMU PIE face database. For each subject, there are 43 face images under different illumination with fixed pose and expression.

The SDA algorithm essentially shares the similar idea while focuses on dimensionality reduction. In the SDA subspace, any ordinary classifier can then be used. In our experiments, we simply choose the nearest centroid method.

4.1. Single Training Image Face Recognition

One of the most successful and well-studied techniques to face recognition is the appearance-based method [22]. Previous works have demonstrated that the face recognition performance can be improved significantly in lower dimensional linear subspaces [1, 14, 22]. Two of the most popular appearance-based methods include *Eigenface* [22] (based on PCA) and *Fisherface* [1] (based on LDA). In general, face appearance does not depend solely on identity. It is also influenced by illumination and viewpoint. Changes in pose and illumination will cause large changes in the appearance of a face. Thus, appearance-based methods need a number of training images for each subject, in order to cope with pose and illumination variabilities.

One of the classical challenges in face recognition is recognition from a single training image [4]. In this setting, the ordinary appearance-based methods (*e.g.*, Eigenface and Fisherface) tend to fail. Actually, with a single training sample per class, it is easy to check that the between-class scatter matrix will be same as the total scatter matrix. Thus, LDA can not be applied. Recent studies show that the face images are sampled from a nonlinear low-dimensional manifold which is embedded in the high-dimensional ambient space [14]. If we have a large set of unlabeled face images (which is possible due to the fast growth of digital photography industry), the intrinsic image manifold can still be estimated even with a single labeled face image per subject. In this experiment, we test our SDA algorithm in this single training image face recognition setting.

The CMU PIE face database [19] is used in this experiment. It contains 68 subjects with 41,368 face images as a whole. The face images were captured under varying pose, illumination and expression. In this experiment, we choose the frontal pose (C27) with varying lighting and illumination which leaves us 43 images per subject. The size of each cropped face image is 32×32 pixels, with 256 grey levels per pixel. Figure 1 shows some sample images for a certain subject. For each subject, 30 images are randomly selected as the training set. Among these 30 images, 1 image is randomly selected and labeled which leaves other 29 images unlabeled. We average the results over 20 random split.

Table 1. Recognition error rates on PIE (mean±std-dev%)

	Unlabeled error	Test error
Baseline	74.7±1.7	74.4±1.6
Eigenface (PCA) [22]	74.7±1.7	74.4±1.6
Laplacianface (LPP) [14]	43.9±2.3	43.6±2.4
Consistency [26]	48.0±1.8	–
LapSVM [3]	43.5±1.6	43.1±2.6
LapRLS [3]	42.5±1.6	42.1±2.6
SDA	**41.0±2.0**	**40.5±2.7**

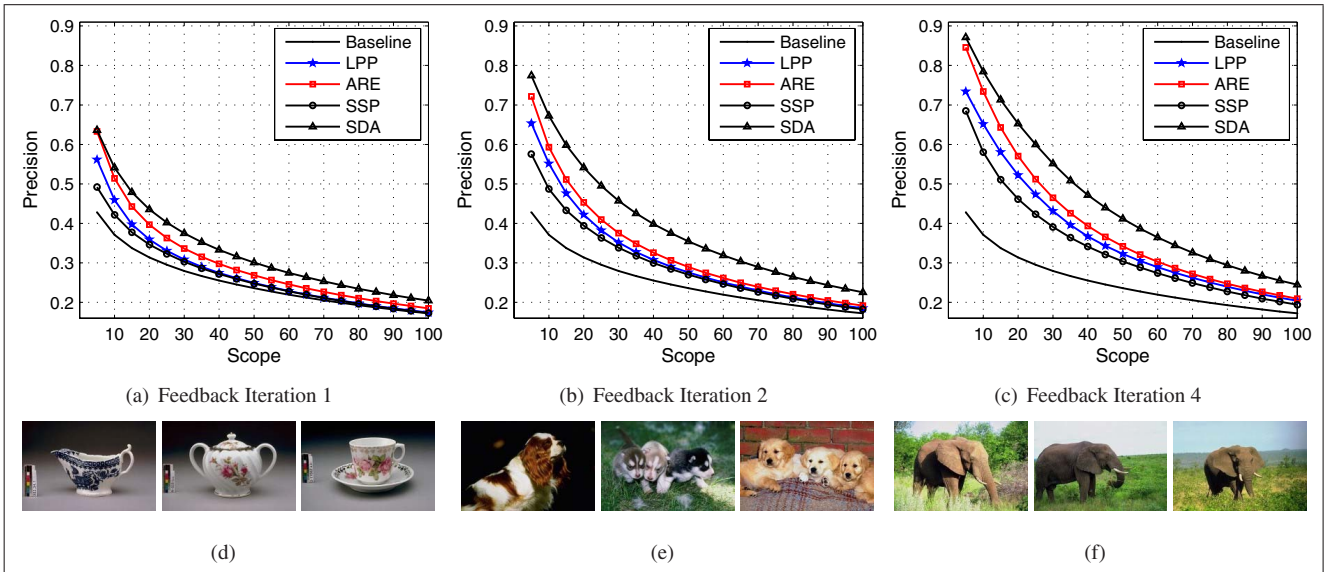
Table 1 shows the performance comparison of different algorithms. The Baseline approach is simply the nearest neighbor classification on the original image space. For other approaches, all the training images (labeled and unlabeled) are used to learn either a subspace or a classifier. The nearest neighbor classifier is then performed in the subspace[3]. The Baseline and Eigenface approaches do not consider the manifold structure and get a very poor performance due to the illumination change. All the other semi-supervised learning approaches make use of the manifold structure and achieved significant improvements. Particularly, our SDA method achieved the best performance among all the compared algorithms.

4.2. Relevance Feedback Image Retrieval

Relevance feedback is a well established and effective framework for narrowing down the gap between low-level visual features and high-level semantic concepts in Content-Based Image Retrieval (CBIR) [18]. Due to the limitation of the user's feedbacks and the high dimensionality of the feature space, one hopes to find a subspace with certain dimensionality reduction algorithms. The semantic relationship between images can be better revealed in this subspace. The relevance feedback setting is certainly a semi-supervised setting, with a large number of unlabeled data (images in the database) and a small number of labeled data (feedbacks provided by the user).

Recently, there are considerable interests on developing semi-supervised dimensionality reduction algorithms for CBIR. Some popular ones include incremental Local-

[3]Since we have only one labeled sample per class, nearest neighbor classifier and the nearest centroid method are the same.



(a) Feedback Iteration 1 (b) Feedback Iteration 2 (c) Feedback Iteration 4

(d) (e) (f)

Figure 2. Compare the retrieval performance of different algorithms. (a)-(c) Via illustrating with the precision-scope curves, we plot the results in the 1st, 2nd, and 4th feedback iteration, respectively. Our SDA algorithm performs the best on the entire scope for all the three feedback iterations. (d)−(f) Sample images from category 24, 25, and 30, respectively.

ity Preserving Projection (LPP) [12], Augmented Relation Embedding (ARE) [15] and Semantic Subspace Projection (SSP) [25]. In this experiment, we compare our SDA with these three algorithms for relevance feedback image retrieval.

4.2.1 Image Database and Low Level Features

The COREL data set is widely used in many CBIR systems, such as [12, 15, 25]. For the sake of evaluations, we also choose this data set for testing. 79 categories of color images were selected, where each consists of 100 images. Some sample images are shown in Figure 2.

We combine 64-dimensional color histogram and 64-dimensional Color Texture Moment (CTM, [24]) to represent the images. The color histogram is calculated using $4 \times 4 \times 4$ bins in HSV space. The Color Texture Moment is proposed by Yu et al. [24], which integrates the color and texture characteristics of the image in a compact form. CTM adopts local Fourier transform as a texture representation scheme and derives eight characteristic maps for describing different aspects of co-occurrence relations of image pixels in each channel of the (SVcosH, SVsinH, V) color space. Then CTM calculates the first and second moment of these maps as a representation of the natural color image pixel distribution. Please see [24] for details.

4.2.2 Evaluation Settings

To exhibit the advantages of using our approach, we need a reliable way of evaluating the retrieval performance and the comparisons with other systems. Different aspects of the experimental design are described below.

Evaluation Metrics: We use *precision-scope curve* [15] to evaluate the effectiveness of the image retrieval algorithms. The scope is specified by the number (N) of top-ranked images presented to the user. The precision is the ratio of the number of relevant images presented to the user to the scope N. The precision-scope curve describes the precision with various scopes and thus gives an overall performance evaluation of the algorithms.

In a real image retrieval system, a query image is usually not in the image database. To simulate such environment, we use *five-fold cross validation* to evaluate the algorithms which is also adopted in the paper [15]. More precisely, we divide the whole image database into five subsets with equal size. Thus, there are 20 images per category in each subset. At each run of cross validation, one subset is selected as the query set, and the other four subsets are used as the database for retrieval. The precision-scope curve and precision rate are computed by averaging the results from the five-fold cross validation.

Automatic Relevance Feedback Scheme: We designed an automatic feedback scheme to model the retrieval process. For each submitted query, our system retrieves and ranks the images in the database. The top 10 ranked images were selected as the feedback images, and their label information (relevant or irrelevant) is used for re-ranking. Note that, the images which have been selected at previous iterations are excluded from later selections. For each query, the automatic relevance feedback mechanism is performed for four iterations. The similar scheme was used in [12], [15], [25].

4.2.3 Image Retrieval Results

In real world, it is not practical to require the user to provide many rounds of feedbacks. The retrieval performance after the first several rounds of feedbacks is the most important. Figure 2 shows the average *precision-scope* curves of the different algorithms for the 1st, 2nd and 4th feedback iterations. The *baseline* curve describes the initial retrieval result without feedback information. Specifically, at the beginning of retrieval, the Euclidean distances in the original 128-dimensional space are used to rank the images in the database. After the user provides relevance feedbacks, the LPP, ARE, SSP, and SDA algorithms are then applied to rerank the images in the database. Our SDA algorithm significantly outperforms the other three algorithms on the entire scope. ARE performs better than the other two, especially with a small scope. All these four algorithms are significantly better than the baseline, which indicates that the user provided relevance feedbacks are very helpful for improving the retrieval performance.

5. Conclusion

In this paper, we propose a new linear dimensionality reduction algorithm called Semi-supervised Discriminant Analysis. It can make efficient use of both labeled and unlabeled data points. The labeled data points are used to maximize the discriminating power, while the unlabeled data points are used to maximize the locality preserving power. Experimental results on single training image face recognition and relevance feedback image retrieval demonstrate the effectiveness of our algorithm.

References

[1] P. N. Belhumeur, J. P. Hepanha, and D. J. Kriegman. Eigenfaces vs. fisherfaces: recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997.

[2] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in Neural Information Processing Systems 14*. 2001.

[3] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from examples. *Journal of Machine Learning Research*, 2006.

[4] D. Beymer and T. Poggio. Face recognition from one example view. In *Proceedings of the Fifth International Conference on Computer Vision (ICCV'95)*, 1995.

[5] D. Cai, X. He, and J. Han. Spectral regression: A unified subspace learning framework for content-based image retrieval. In *Proceedings of the ACM Conference on Multimedia*, 2007.

[6] O. Chapelle, J. Weston, and B. Schölkopf. Cluster kernels for semi-supervised learning. In *Advances in Neural Information Processing Systems 16*, 2003.

[7] F. R. K. Chung. *Spectral Graph Theory*, volume 92 of *Regional Conference Series in Mathematics*. AMS, 1997.

[8] J. H. Friedman. Regularized discriminant analysis. *Journal of the American Statistical Association*, 84(405):165–175, 1989.

[9] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, 2nd edition, 1990.

[10] G. H. Golub and C. F. V. Loan. *Matrix computations*. Johns Hopkins University Press, 3rd edition, 1996.

[11] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer-Verlag, 2001.

[12] X. He. Incremental semi-supervised subspace learning for image retrieval. In *Proceedings of the ACM Conference on Multimedia*, New York, October 2004.

[13] X. He and P. Niyogi. Locality preserving projections. In *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2003.

[14] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang. Face recognition using laplacianfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3):328–340, 2005.

[15] Y.-Y. Lin, T.-L. Liu, and H.-T. Chen. Semantic manifold learning for image retrieval. In *Proceedings of the ACM Conference on Multimedia*, Singapore, November 2005.

[16] K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis*. Academic Press, 1980.

[17] A. Y. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems 14*, pages 849–856. MIT Press, Cambridge, MA, 2001.

[18] Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra. Relevance feedback: A power tool for interative content-based image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 8(5), 1998.

[19] T. Sim, S. Baker, and M. Bsat. The CMU pose, illuminlation, and expression database. *IEEE Transactions on PAMI*, 25(12):1615–1618, 2003.

[20] V. Sindhwani, P. Niyogi, and M. Belkin. Beyond the point cloud: from transductive to semi-supervised learning. In *Proc. 2005 Int. Conf. Machine Learning (ICML'05)*, 2005.

[21] A. N. Tikhonov. Regularization of incorrectly posed problems. *Soviet Math.*, (4), 1963 (English Translation).

[22] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[23] V. N. Vapnik. *Statistical learning theory*. John Wiley & Sons, 1998.

[24] H. Yu, M. Li, H.-J. Zhang, and J. Feng. Color texture moments for content-based image retrieval. In *International Conference on Image Processing*, pages 24–28, 2002.

[25] J. Yu and Q. Tian. Learning image manifolds by semantic subspace projection. In *Proceedings of the ACM Conference on Multimedia*, Santa Barbara, October 2006.

[26] D. Zhou, O. Bousquet, T. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. In *Advances in Neural Information Processing Systems 16*, 2003.

[27] X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *Proc. of the twentieth Internation Conference on Machine Learning*, 2003.