Semi-Supervised Discriminant Analysis via Spectral Transduction

Deming Zhai Student[1]
dmzhai@jdl.ac.cn

Hong Chang Advisor[2]
hchang@jdl.ac.cn

Bo Li Collaborator[1]
bli@jdl.ac.cn

Shiguang Shan Professor[2]
sgshan@jdl.ac.cn

Xilin Chen Professor[2]
xlchen@jdl.ac.cn

Wen Gao Professor[31]
wgao@jdl.ac.cn

[1] School of Computer Science and Technology,
Harbin Institute of Technology,
Harbin, China

[2] Key Laboratory of Intelligent Information Processing,
Chinese Academy of Sciences,
Beijing,China

[3] Institute of Digital Media,
Peking University,
Beijing, China

Abstract

Linear Discriminant Analysis (LDA) is a popular method for dimensionality reduction and classification. In real-world applications when there is no sufficient labeled data, LDA suffers from serious performance drop or even fails to work. In this paper, we propose a novel method called Spectral Transduction Semi-Supervised Discriminant Analysis (STSDA), which can alleviate such problem by utilizing both labeled and unlabeled data. Our method takes into consideration both label augmenting and local structure preserving. First, we formulate label transduction with labeled and unlabeled data as a constrained convex optimization problem and solve it efficiently with a closed-form solution by using orthogonal projector matrices. Then, unlabeled data with reliable class estimations are selected with a balanced strategy to augment the original labeled data set. At last, LDA with manifold regularization is performed. Experimental results on face recognition demonstrate the effectiveness of our proposed method.

1 Introduction

During the past decades, Linear Discriminant Analysis (LDA) [11] has been one of the most popular dimensionality reduction methods in pattern recognition field. It has been applied to a wide range of classification tasks with great success, such as face recognition, text classification, etc. However, when there are only few labeled data samples relative to the number of dimensionality, which is the so-called small sample size (SSS) problem [10], the performance of LDA will seriously deteriorate. Under such situation, the generalization capability of LDA on test samples cannot be guaranteed. In order to address this problem, several numerical solutions have been proposed, e.g., PseudoLDA [13], PCA+LDA [1], LDA/QR [21], NullLDA [10], and DualLDA [19].

Another possible solution for SSS problem is to learn with both labeled and unlabeled data. It is more natural and reasonable since in reality we usually have a large supply of unlabeled data and comparatively insufficient labeled data. Assigning labels requires laborious human effort, so it is expensive or hard to obtain. Moreover, accurate labeling may need some expert knowledge. Therefore, it is desirable to make good use of unlabeled data to improve the classification performance. In the past few years, semi-supervised learning has aroused a great deal of interest in the machine learning community. Some representative methods include: Co-training [5], transductive support vector machines [3], graph-based methods [6]. A good survey of semi-supervised learning can be found in [24].

Recently, some Linear Discriminant Analysis algorithms under semi-supervised setting have been proposed in the literature. Cai *et al.* first put forward semi-supervised LDA algorithm called SDA [8], which exploits the local neighborhood information of data points in performing dimensionality reduction. Later, Zhang and Yeung presented another method called SSDA [22] in which a robust path-based similarity measure is used to capture global manifold structure of the data. Similarly, SMDA [20] and UDA [15] also perform LDA under semi-supervised setting through manifold regularization. All the methods stated above can be considered as one class since they share the similar idea of exploiting the local or global geometric structure of both labeled and unlabeled data. Methods of this class have a main drawback in that no class-wise information, which is essential for LDA in scatter matrices estimation, is explored from unlabeled data.

More recently, a new algorithm called SSDA$_{CCCP}$ [23] has been proposed, which explores label information from unlabeled data and uses the augmented labeled data to perform LDA. In this method, unlabeled data is involved in maximizing the optimality criterion of LDA. The class labels for unlabeled data are estimated by solving the optimization problem through the constrained concave-convex procedure (CCCP). Compared with previous approaches mentioned above, SSDA$_{CCCP}$ leads to significant performance improvement. However, this method suffers from the disadvantage of the non-convex optimization, where CCCP works in an iterative way and the final solution generally depends on its initial value. Moreover, although manifold assumption can be adopted (as in $\mathrm{M}-\mathrm{SSDA}_{CCCP}$) in the label estimation step, no structure preserving strategy is used in performing LDA with the augmented labeled data, which may not be the best choice for the still unlabeled samples.

It has been proved that the number and quality of labeled samples have an exponential effect on reducing the classification error [9]. Therefore, in this paper, we continue to pursue in the direction of exploring label information from unlabeled training samples for LDA. Our proposed method, Spectral Transduction Semi-Supervised Discriminant Analysis (STSDA), works in a different way from [23]. It comprises three stages. First, we formulate label transduction with labeled and unlabeled data as a convex optimization problem with pairwise constraints and solve it efficiently with a closed-form solution. Then, some unlabeled data with reliable class estimations are selected through a balanced strategy to augment the original labeled data set. At last, LDA with manifold regularization is performed. Compared with previous related methods, our work has advantages in the following aspects: 1) We take into account both label augmenting and local structure preserving. 2) The optimization problem is convex which could be solved effectively in an analytical manner with a global optimal solution. 3) The balanced data selection strategy is more effective than the preliminary method.

The rest of this paper is organized as follows. In Section 2, some background knowledge is introduced. In Section 3, we present our proposed semi-supervised discriminant analysis algorithm in detail. Experiments and analysis are presented in Section 4. Finally, Section 5 gives some concluding remarks.

2 Background

Since the proposed method is a further development of LDA [11] and Normalized Cuts [16], we start by summarizing and discussing these two approaches.

2.1 Linear Discriminant Analysis

Suppose we are given a data set of n labeled samples $\mathscr{X} = \{\mathbf{x}_i\}_{i=1}^n$, where $\mathbf{x}_i \in \mathbb{R}^d$. $\mathscr{X}$ can be divided into C disjoint classes $\Pi_k, k = 1, \ldots, C$, where class Π_k contains n_k samples. Linear Discriminant Analysis (LDA) tries to seek for an optimal transform $\mathbf{W}$ by maximizing the following objective function

$$J(\mathbf{W}) = \text{trace}\left(\frac{\mathbf{W}^T \mathbf{S}_b\, W}{\mathbf{W}^T \mathbf{S}_w \mathbf{W}}\right), \tag{1}$$

where $\mathbf{S}_b$ is the between-class scatter matrix, and $\mathbf{S}_w$ is the within-class scatter matrix. The definitions of $\mathbf{S}_b$ and $\mathbf{S}_w$ are

$$\begin{aligned} \mathbf{S}_b &= \sum_{k=1}^{C} n_k (\overline{\mathbf{m}}_k - \overline{\mathbf{m}})(\overline{\mathbf{m}}_k - \overline{\mathbf{m}})^T, \\ \mathbf{S}_w &= \sum_{k=1}^{C} \sum_{\mathbf{x}_i \in \Pi_k} (\mathbf{x}_i - \overline{\mathbf{m}}_k)(\mathbf{x}_i - \overline{\mathbf{m}}_k)^T, \end{aligned} \tag{2}$$

where $\overline{\mathbf{m}} = \left(\sum_{i=1}^n \mathbf{x}_i\right)/n$ is the sample mean of the whole data set $\mathscr{X}$ and $\overline{\mathbf{m}}_k = \left(\sum_{\mathbf{x}_i \in \Pi_k} \mathbf{x}_i\right)/n_k$ is the mean of class Π_k. When the dimensionality of data is larger than the sample size, the scatter matrices of $\mathbf{S}_b$ and $\mathbf{S}_w$ are singular, thus the small size problem occurs.

2.2 Normalized Cuts

From the perspective of graph model, the training data can be represented by an undirected graph $G = (V, E)$ with weight matrix $\mathbf{S}$. $V = \{1, ..., n\}$ is the vertex set corresponding to all data samples in $\mathscr{X}$ and the edge set $E \subseteq V \times V$ represents the relationship between these samples. Specifically, we put an edge between node i and j if $\mathbf{x}_i$ and $\mathbf{x}_j$ are among k nearest neighbors of each other. Each edge is assigned a weight which is defined by

$$\mathbf{S}_{ij} = \begin{cases} e^{\frac{-\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{2\sigma^2}} & , \text{ if } \mathbf{x}_i \in \mathscr{N}_k(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in \mathscr{N}_k(\mathbf{x}_i) \\ 0 & , \text{ otherwise.} \end{cases} \tag{3}$$

$\mathscr{N}_k(\mathbf{x}_i)$ denotes the set of k nearest neighbors of $\mathbf{x}_i$ and σ is a scaling parameter which controls the decreasing speed of $\mathbf{S}_{ij}$ with the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$.

Normalized Cuts is a kind of graph partition criterion that can be formulated as a convex optimization problem with closed-form solution. If we partition the vertex set into two disjoint sets A and B satisfying $A \cup B = V$ and $A \cap B = \phi$, the Normalized Cuts of this graph is defined as

$$Ncut(A, B) = \frac{cut(A, B)}{asso(A, V)} + \frac{cut(A, B)}{asso(B, V)}, \tag{4}$$

where $cut(A, B) = \sum_{i \in A, j \in B} \mathbf{S}_{ij}$ and $asso(A, V) = \sum_{i \in A, j \in V} \mathbf{S}_{ij}$. Let $\mathbf{z} \in \{-1, 1\}^n$ be the class indicator vector, $\mathbf{d} = \mathbf{S}\mathbf{1}_n$ be the $n \times 1$ vector containing the row sums of $\mathbf{S}$ where $\mathbf{1}_n$ denotes the $n \times 1$ vector of all 1's. Let $\mathbf{D} = \text{diag}(\mathbf{d})$ be the $n \times n$ diagonal matrix, and $\mathbf{L} = \mathbf{D} - \mathbf{S}$ is called Laplacian matrix. With these notations, we can rewrite Eq. (4) as :

$$Ncut(A,B) = \frac{\mathbf{z}^T\mathbf{L}\mathbf{z}}{2\mathbf{d}^T(\mathbf{1}_n + \mathbf{z})} + \frac{\mathbf{z}^T\mathbf{L}\mathbf{z}}{2\mathbf{d}^T(\mathbf{1}_n - \mathbf{z})} = \frac{(\mathbf{z}^T\mathbf{L}\mathbf{z})\mathbf{d}^T\mathbf{1}_n}{\mathbf{z}^T(\mathbf{1}_n^T\mathbf{d}\mathbf{D} - \mathbf{d}\mathbf{d}^T)\mathbf{z}}, \tag{5}$$

where $\mathbf{d}^T\mathbf{1}_n$ is a constant with no influence on the solution. To be concise, denote $\mathbf{Q} = \mathbf{1}_n^T\mathbf{d}\mathbf{D} - \mathbf{d}\mathbf{d}^T$. The final optimization problem becomes:

$$\min Ncut(A,B) = \min \frac{\mathbf{z}^T\mathbf{L}\mathbf{z}}{\mathbf{z}^T\mathbf{Q}\mathbf{z}}. \tag{6}$$

The above method can be easily generalized to multi-class cases through $\alpha - \beta$ swap, which is used in many graph cuts applications [7].

3 Spectral Transduction Semi-Supervised Discriminant Analysis

In this section, we present our semi-supervised discriminant analysis algorithm in three phases: spectral transduction via constrained Normalized Cuts, labeled data set augmenting and LDA with local structure preserving. The overall algorithm is then summarized followed by a short discussion.

3.1 Spectral Transduction via Constrained Normalized Cuts

Suppose we have n training data samples, l samples of them $\mathbf{x}_1, \ldots, \mathbf{x}_l$ are with class labels from C classes $\Pi_k, k = 1, \ldots, C$, and the rest samples $\mathbf{x}_{l+1}, \ldots, \mathbf{x}_n$ are unlabeled. Note that $l \ll n$. To explore more label information from the training set, we try to predict the labels of unlabeled data as accurately as possible. More specifically, we utilize the global structure of labeled and unlabeled data to carry out spectral transduction based on Normalized Cuts method, and formulate the transduction procedure as a pairwise constrained convex optimization problem.

Let us consider the supervisory information in the form of pairwise similarity and dissimilarity constraints, which are included in $\mathscr{S}$ and $\mathscr{D}$, respectively.

$$\begin{aligned} \mathscr{S} &= \{(\mathbf{x}_i, \mathbf{x}_j) | \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ belong to the same class}\}, \\ \mathscr{D} &= \{(\mathbf{x}_m, \mathbf{x}_n) | \mathbf{x}_m \text{ and } \mathbf{x}_n \text{ belong to different classes}\}. \end{aligned} \tag{7}$$

Pairwise constraints are weaker than label information since labeled data can be converted to pairwise constraints but not vice versa.

We denote each pairwise similarity constraint $(\mathbf{x}_i, \mathbf{x}_j) \in \mathscr{S}$ by an n-dimensional indicator vector $\mathbf{u}_k$ (k-th, $k = 1, \ldots, |\mathscr{S}|$), which has only two non-zero elements: $\mathbf{u}_k(i) = 1$ and $\mathbf{u}_k(j) = -1$. Since the class indicators $\mathbf{z}_i$ and $\mathbf{z}_j$ are equal (both 1 or -1 for two class problem), we have $\mathbf{u}_k^T\mathbf{z} = 0$. Let $\mathbf{U} = [\mathbf{u}_1, ..., \mathbf{u}_{|\mathscr{S}|}]$ be the positive constraints matrix, where $|\mathscr{S}|$ denotes the cardinality of $\mathscr{S}$. Then, the pairwise similarity constraints can be expressed as $\mathbf{U}^T\mathbf{z} = \mathbf{0}$. Similarly, each dissimilarity constraint $(\mathbf{x}_m, \mathbf{x}_n) \in \mathscr{D}$ can be represented by an indicator vector $\mathbf{v}_k$ with only two non-zero elements: $\mathbf{v}_k(m) = 1$ and $\mathbf{v}_k(n) = 1$. Since $\mathbf{z}_m + \mathbf{z}_n = 0$, we have $\mathbf{v}_k^T\mathbf{z} = 0$. Define $\mathbf{V} = [\mathbf{v}_1, ..., \mathbf{v}_{|\mathscr{D}|}]$ as the negative constraints matrix.

The pairwise dissimilarity constraints can be expressed as $\mathbf{V}^T\mathbf{z} = \mathbf{0}$. Consequently, we have the following constrained Normalized Cuts formulation which is an extension of Eq. (6):

$$\min_{\mathbf{z}} \frac{\mathbf{z}^T\mathbf{L}\mathbf{z}}{\mathbf{z}^T\mathbf{Q}\mathbf{z}} \text{ s.t. } \mathbf{U}^T\mathbf{z} = \mathbf{0}\,;\, \mathbf{V}^T\mathbf{z} = \mathbf{0}. \tag{8}$$

Solving this constrained optimization problem can be facilitated by using orthogonal projection matrices. Let $\mathbb{U} \in \mathbb{R}^n$ be a subspace spanned by columns of $\mathbf{U}$ with $\mathbb{U}^\perp$ as its null orthogonal space. $\mathbf{P_U}$ and $\mathbf{P}_{\mathbf{U}^\perp}$ are the orthogonal projection matrices onto $\mathbb{U}$ and $\mathbb{U}^\perp$, respectively. From the definition above, $\mathbf{P}_{\mathbf{U}^\perp}\mathbf{z}$ is the projection of $\mathbf{z}$ onto $\mathbb{U}^\perp$ and it satisfies the property as follows:

$$\forall\, \mathbf{z} \in \mathbb{U}^\perp\,,\, \mathbf{P_U}\mathbf{z} = \mathbf{0}\,;\, \mathbf{P}_{\mathbf{U}^\perp}\mathbf{z} = \mathbf{z}. \tag{9}$$

According to [12], $\mathbf{P_U}$ can be calculated directly from $\mathbf{U}$ as $\mathbf{P_U} = \mathbf{U}(\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T$. It can be easily verified that $\mathbf{P}_{\mathbf{U}^\perp} = \mathbf{I} - \mathbf{P_U}$, where $\mathbf{I}$ is the identity matrix. Therefore, if $\mathbf{z}$ is a feasible solution, it must satisfy $\mathbf{P}_{\mathbf{U}^\perp}\mathbf{z} = \mathbf{z}$. In the same sense, $\mathbf{P}_{\mathbf{V}^\perp}$ is defined as the orthogonal projection matrix on the null orthogonal space spanned by $\mathbf{V}$ and it can be computed in a similar way. With these transformations, Eq. (8) can be expressed as:

$$\min_{\mathbf{z}} \frac{\mathbf{z}^T\mathbf{L}\mathbf{z}}{\mathbf{z}^T\mathbf{Q}\mathbf{z}} \text{ s.t. } \mathbf{P}_{\mathbf{U}^\perp}\mathbf{z} = \mathbf{z}\,;\, \mathbf{P}_{\mathbf{V}^\perp}\mathbf{z} = \mathbf{z}. \tag{10}$$

The solution to this optimization problem can be finally obtained by solving a generalized eigenvalue problem which be formulated as Eq. (11). This spectral transduction approach is more efficient than the iterative optimization approach and can get a global optimal solution.

$$\mathbf{L}\mathbf{P}_{\mathbf{U}^\perp}\mathbf{P}_{\mathbf{V}^\perp}\mathbf{z} = \lambda\mathbf{Q}\mathbf{z} \tag{11}$$

3.2 Labeled Data Set Augmenting

By applying constrained Normalized Cuts based spectral transduction, we estimate the class labels of all the unlabeled data points in the training set. In this subsection, we aim to select some unlabel data points whose label estimation are with sufficiently high confidence.

We first perform LDA using all the training data with their estimated labels. Then, in the embedding space, the label confidence is defined according to local data distribution. Specifically, for each unlabeled data $\mathbf{x}_i$ $(i = l+1,\ldots,n)$, its label confidence is defined as the proportion of data points with the same estimated label as $\mathbf{x}_i$ among its k nearest neighbors. Different from the previous method which makes use of a single confidence threshold for all unlabeled data points [23], we design an alternative strategy to have a balanced data augmenting results. For all the unlabeled data with the same estimated class labels, we sort their confidence values in descending order and select the first m samples to augment the original labeled data set. m is a selection scale factor which is usually chosen to be less than the minimum of class cardinalities.

The previous confidence threshold method sometimes suffers from the disadvantage that for some classes the number of augmented labeled data is large while for some ones very small or even zero. In contrast, our proposed augmenting strategy is more balanced since the augmented labeled data number for each class is equal. Consequently, the recognition error rates of all the classes are expected to decrease, which can bring some benefits for subsequent classification task.

3.3 LDA with Local Structure Preserving

With the augmented labeled data set, we seek to find a global projection that can not only improve class discriminative ability but also preserve local data structure. Based on the cluster assumption that nearby data points are likely to be in the same class, we take into account local structure preserving through Laplacian regularization [2]. The optimization problem of the regularized LDA can be written as:

$$\mathbf{W}^* = \max_{\mathbf{W}} \operatorname{trace}\left(\frac{\mathbf{W}^T \mathbf{S}_b \mathbf{W}}{\mathbf{W}^T \mathbf{S}_{\mathrm{w}} \mathbf{W} + \alpha K(\mathbf{W})} \right). \tag{12}$$

The scatter matrixes $\mathbf{S}_b$ and $\mathbf{S}_w$ are computed with the augmented labeled data set. $K(\mathbf{W})$ is the Laplacian regularization term and the coefficient α controls the relative importance of the discrimination and regularization. $K(\mathbf{W})$ is defined as:

$$K(\mathbf{W}) = \sum_{i,j=1}^{n} (\mathbf{W}^T \mathbf{x}_i - \mathbf{W}^T \mathbf{x}_j)^2 \mathbf{S}_{ij} = \mathbf{W}^T \mathbf{X} \mathbf{L} \mathbf{X}^T \mathbf{W}, \tag{13}$$

where $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_n]$ is the matrix form of the whole data set.

With the Laplacian regularizer, all data points are involved in the optimization. Thus the local geometric structure of both labeled and unlabeled data tends to be preserved with the transformation $\mathbf{W}$. The discriminant projector $\mathbf{W}$ can be computed efficiently by solving the following generalized eigendecomposition problem:

$$\mathbf{S}_b \mathbf{W} = \lambda (\mathbf{S}_w + \alpha \mathbf{X} \mathbf{L} \mathbf{X}^T) \mathbf{W}. \tag{14}$$

3.4 The Algorithm

The algorithm of Semi-Supervised Discriminant Analysis via Spectral Transduction (STSDA) is summarized in Table 1.

```
Input: Labeled data samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^{l}$ belonging to $C$ classes
       and unlabeled data samples $\{\mathbf{x}_i\}_{i=l+1}^{n}$
Output: Discriminant projector $\mathbf{W}$
Phase 1: Label Transduction
    Step 1: Construct the adjacency graph $G$;
    Step 2: Solve a generalized eigenvalue problem: $\mathbf{L}\mathbf{P}_{\mathbf{U}^\perp}\mathbf{P}_{\mathbf{V}^\perp}\mathbf{z} = \lambda \mathbf{Q}\mathbf{z}$;
Phase 2: Labeled Data Set Augmenting
    Step 3: Perform LDA on all data with estimated labels $\mathbf{z}$;
    Step 4: Compute label confidence in embedding space;
    Step 5: Select reliable unlabeled data for each class;
Phase 3: LDA with Local Structure Preserving
    Step 6: Compute $S_b$ and $S_w$ with the augmented labeled data set;
    Step 7: Solve a generalized eigenvalue problem: $\mathbf{S}_b\mathbf{W} = \lambda(\mathbf{S}_w + \alpha\mathbf{X}\mathbf{L}\mathbf{X}^T)\mathbf{W}$.
```

Table 1: Algorithm of STSDA

In our method, two optimization problems are defined (in Phase 1 and 3) and both make use of unlabeled data. Actually, the usage of unlabeled data in the two phases corresponds to the global and local meanings of the cluster assumption, respectively. In Phase 1, spectral transduction based on Normalized Cuts conveys the global cluster assumption that data points on the same manifold are likely to have the same label. In Phase 3, LDA with Laplacian regularization makes use of the local assumption that nearby data points are likely to be in the same class.

Our method belongs to the semi-supervised learning paradigm. The learned discriminant projector $\mathbf{W}$ can be applied to not only the unlabeled training data but also any unseen test data. The experiments below are conducted to solve face recognition problems under the semi-supervised settings.

4 Experiments

In this section, we perform some experiments on real-world face recognition problems to demonstrate the effectiveness of the proposed method.

4.1 Experimental Settings

In our experiments, we consider two public databases : CMU PIE [17] and AR [14], in order to achieve an extensive evaluation of different methods against database changes including illumination variations, facial expressions, occlusions and the database sizes.

The CMU PIE (Pose, Illumination and Expression) Database contains a total of 41,368 images of 68 individuals. Face images of each individual are captured under different poses, illumination conditions, and with different expressions. In our experiment, we choose the frontal pose (C27) with only lighting and illumination changes. For each subject, 30 images are randomly selected for the training and the rest 13 images for testing. The second database, AR Face Database, consists of 4,000 color frontal view faces of 135 individuals (70 men and 56 women). In our experiment, we select 100 individuals of 50 men and 50 women. Of the 26 images for each person, we split them equally to form training and testing sets.

Before the experiments, all face images are converted to gray images with histogram equalization and then resized to 32×32 pixels according to the positions of two eyes. On each training set, we randomly label $q = 1, 2, 3$, or 4 images for each class. For each configuration, we perform 20 random trails and report the average recognition error rate and standard derivation.

We compare our proposed method with some baseline and previous related methods. More specifically, six approaches are included in our comparative study: 1) Baseline method which directly uses original space for recognition; 2) Eigenface [18]; 3) Fisherface [1]; 4) SDA [8]; 5) SSDA$_{CCCP}$ [23]; 6) STSDA (our approach). For all methods stated above, after finding a embedding space, a simple nearest neighbor (NN) classifier is then performed.

There are a few parameters involved in our experiments. In Fisherface, PCA is applied first to avoid singularity problem and the target dimensionality is set to preserve 95% of data energy. The regularization coefficient α in Eq. (12) is fixed to 0.1 for both SDA and STSDA. The width parameter σ in Eq. (3) is empirically set between $4\bar{D} \sim 15\bar{D}$ for constrained Normalized Cuts and $\bar{D}/10 \sim \bar{D}/20$ for Laplacian regularization, with $\bar{D}$ representing the mean derivations in each class.

4.2 Recognition Results and Analysis

The recognition results on both databases are summarized in Table 2 and 3. As shown in these two tables, the baseline method and Eigenface (based on PCA), which don't use the supervisory information, yield poor performance. Fisherface (based on LDA) method, which is completely supervised, still cannot provide satisfied results since the number of labeled data (⩽4) is far less than that of the feature dimensionality (1024) and the so-called SSS problem occurs. Especially, when there is only one labeled image for each subject in training set, LDA-based method fails to work because the intra-person variance cannot be obtained.This

Method	♯ labels/class =1		♯ labels/class=2	
	Unlabeled Error	Test Error	Unlabeled Error	Test Error
Baseline	0.6751±0.0095	0.6711±0.0151	0.5351±0.0165	0.5381±0.0165
Eigenface	0.6867±0.0103	0.6851±0.0134	0.5165±0.0132	0.5161±0.0144
Fisherface	-	-	0.1615±0.0136	0.1619±0.0176
SDA	0.3577±0.0128	0.3567±0.0169	0.1134±0.0163	0.1149±0.0171
SSDA$_{CCCP}$	0.2631±0.0235	0.2602±0.0250	0.1195±0.0136	0.1164±0.0147
STSDA	**0.0664±0.0155**	**0.0673±0.0185**	**0.0421±0.0104**	**0.0429±0.0120**
Method	♯ labels/class =3		♯ labels/class=4	
	Unlabeled Error	Test Error	Unlabeled Error	Test Error
Baseline	0.3961±0.0203	0.3989±0.0249	0.3147±0.0169	0.3148±0.0144
Eigenface	0.4251±0.0203	0.4264±0.0251	0.3448±0.0167	0.3471±0.0165
Fisherface	0.0876±0.0165	0.0855±0.0174	0.0538±0.0087	0.0518±0.0100
SDA	0.0651±0.0147	0.0643±0.0153	0.0492±0.0120	0.0463±0.0124
SSDA$_{CCCP}$	0.0604±0.0109	0.0541±0.0129	0.0365±0.0094	0.0331±0.0123
STSDA	**0.0408±0.0107**	**0.0351±0.0121**	**0.0309±0.0085**	**0.0285±0.0102**

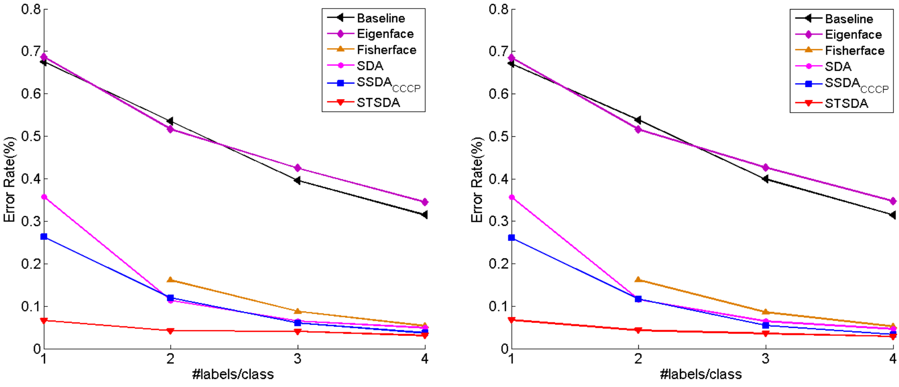
Table 2: Recognition error rates on PIE (mean±std-dev)

Method	♯ labels/class =1		♯ labels/class=2	
	Unlabeled Error	Test Error	Unlabeled Error	Test Error
Baseline	0.8902±0.0089	0.8915±0.0089	0.8311±0.0065	0.8285±0.0098
Eigenface	0.8928±0.0097	0.8959±0.0088	0.8359±0.0076	0.8371±0.0109
Fisherface	-	-	0.6702±0.0183	0.6633±0.0161
SDA	0.8897±0.0085	0.8911±0.0098	0.5711±0.0188	0.5677±0.0169
SSDA$_{CCCP}$	0.7056±0.0253	0.7236±0.0226	0.5549±0.0249	0.5782±0.0216
STSDA	**0.3557±0.0122**	**0.4507±0.0106**	**0.3177±0.0142**	**0.4141±0.0200**
Method	♯ labels/class =3		♯ labels/class=4	
	Unlabeled Error	Test Error	Unlabeled Error	Test Error
Baseline	0.7794±0.0103	0.7808±0.0082	0.7359±0.0146	0.7442±0.0097
Eigenface	0.7861±0.0096	0.7929±0.0085	0.7476±0.0139	0.7588±0.0083
Fisherface	0.5651±0.0143	0.5518±0.0179	0.4688±0.0182	0.4606±0.0143
SDA	0.4683±0.0176	0.4602±0.0156	0.4166±0.0175	0.3981±0.0135
SSDA$_{CCCP}$	0.4508±0.0175	0.4778±0.0179	0.3795±0.0187	0.4063±0.0155
STSDA	**0.3344±0.0143**	**0.3939±0.0151**	**0.3271±0.0206**	**0.3728±0.0131**

Table 3: Recognition error rates on AR (mean±std-dev)

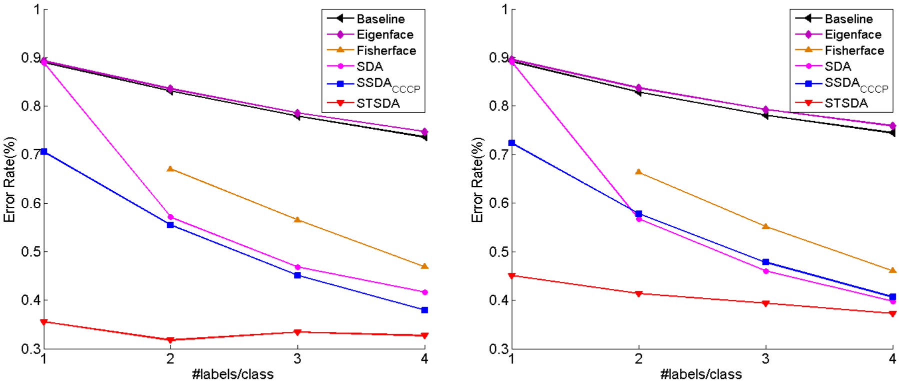
is a classical challenge in face recognition called recognition from single training image problem [4], which have not been completely solved yet. The SDA method, which performs better than Fishface, exploits unlabeled data to preserve their manifold structure while without exploring class-wise information. As for SSDA$_{CCCP}$, it mainly augments labeled data in a iterative way while no structure preserving strategy is used in performing LDA. Among all presented methods, the proposed STSDA achieves the best results. By exploring discriminative knowledge and preserving local data structure, our STSDA method leads to significant performance benefits.

Moreover, the changes of recognition error rates with the number of labeled data are studied. As illustrated in Figure 1 and 2, the error rates of baseline and Eigenface methods



(a) (b)

Figure 1: Recognition error rates as a function of the number of labeled data on PIE. (a) unlabeled data error rates, (b) test data error rates.



(a) (b)

Figure 2: Recognition error rates as a function of the number of labeled data on AR. (a) unlabeled data error rates, (b) test data error rates.

decrease significantly with the increasing number of labeled data for PIE database where the intra-person variability is small. While for AR database where the intra-person variance is lager than the inter-person variance, the performance changes of these two methods are much slower. As for Fisherface method, the recognition error rates decrease rapidly with increasing number of labeled data. When we add one more labeled data for each subject, the overall recognition error rate has a reduction of nearly 10%. The change trends of SDA and SSDA$_{CCCP}$ are similar. The only difference between them is that when there is only one labeled image for each class, SSDA$_{CCCP}$ method gives better performance since additional discriminative knowledge is explored from unlabeled data. Compared with all above methods, our proposed STSDA method achieves fairly low and stable recognition error rates. This is owe to the effectiveness of spectral transduction and balanced label augmenting strategy. In particular, when we confront with single training image recognition problem, the reduction of recognition error rate for STSDA is very remarkable.

5 Concluding Remarks

In this paper, we propose a novel spectral based discriminant analysis approach under semi-supervised setting. Different from some previous work, our method considers both label augmenting and local structure preserving. On one hand, spectral transduction is utilized for label estimation and is formulated as a convex optimization problem with pairwise constraints. This optimization problem can be solved efficiently with a closed-form solution. In addition, unlabeled data with reliable class estimations are selected to augment the labeled data set through the proposed balanced data selection strategy. On the other hand, both labeled and unlabeled data are utilized to preserve local structure by using manifold regularization. Experimental results on real-world face recognition tasks demonstrate the effectiveness of our method.

6 Acknowledgements

This paper is partially supported by Natural Science Foundation of China under contract Nos. 60803084, 60872124, and 60872077; and Grand Program of International S&T Cooperation of Zhejiang Province S&T Department (No.2008C14063).

References

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

[2] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization : A geometric framework for learning from examples. *Journal of Machine Learning Research*, pages 2399–2434, 2004.

[3] K.P. Bennett and A. Demiriz. Semi-supervised support vector machines. In *Advances in Neural Information Processing Systems 11*, pages 368–374. MIT Press, Cambridge, MA, USA, 1999.

[4] D. Beymer and T. Poggio. Face recognition from one example view. In *Proceedings of the Fifth IEEE International Conference on Computer Vision*, pages 500–507, Jun 1995.

[5] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pages 92–100, Madison, WI, 1998.

[6] A. Blum, J. Lafferty, M.R. Rwebangira, and R. Reddy. Semi-supervised learning using randomized mincuts. In *Proceedings of the Twenty-First International Conference on Machine Learning*, pages 97–104, 4–8 August 2004.

[7] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239, 2001.

[8] D. Cai, X. He, and J. Han. Semi-supervised discriminant analysis. In *Proceedings of the Eleventh IEEE International Conference on Computer Vision*, pages 1–7, 2007.

[9] V. Castelli and T.M. Cover. On the exponential value of labeled samples. *Pattern Recogntion Letter*, 16(1):105–111, 1995.

[10] L. Chen, H. Liao, M. Ko, J. Lin, and G. Yu. A new LDA-based face recognition system which can solve the small sample size problem. *Pattern Recognition*, 33(10): 1713–1726, 2000.

[11] R.A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936.

[12] G. H. Golub and C. F. Van Loan. *Matrix Computations.* Johns Hopkins Studies in Mathematical Sciences, 1996.

[13] W.J. Krzanowski, P. Jonathan, W.V. McCarthy, and M.R. Thomas. Discriminant analysis with singular covariance matrices: methods and applications to spectroscopic data. *Applied Statistics*, 44(1):101–115, 1995.

[14] A.M. Martinez and R. Benavente. The AR face database. *CVC Technical Report*, (24), June. 1998.

[15] H. Qiu, J. Lai, J. Huang, and Y. Chen. Semi-supervised discriminant analysis based on UDP regularization. In *Proceedings of the Nineteenth International Conference on Pattern Recognition*, 2008.

[16] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

[17] T. Sim, S. Baker, and M. Bsat. The CMU pose, illumination, and expression database. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(12):1615–1618, Dec. 2003.

[18] M. Turk and A. Pentland. Eigenfaces for recognition. In *Journal of Cognitive Neuroscience*, volume 3, pages 71–86, 1991.

[19] X. Wang and X. Tang. Dual-space linear discriminant analysis for face recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 564–569, 2004.

[20] R. Xiao and P. Shi. Semi-supervised marginal discriminant analysis based on QR decomposition. In *Proceedings of the Nineteenth International Conference on Pattern Recognition*, 2008.

[21] J.P. Ye and Q. Li. A two-stage linear discriminant analysis via QR-decomposition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(6):929–941, 2005.

[22] Y. Zhang and D.Y. Yeung. Semi-supervised discriminant analysis using robust path-based similarity. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2008.

[23] Y. Zhang and D.Y. Yeung. Semi-supervised discriminant analysis via CCCP. In *Proceedings of the European Conference on Machine Learning*, 2008.

[24] X. Zhu. Semi-supervised learning literature survey. Technical Report 1530, Departmant of Computer Sciences, University of Wisconsin at Madison, Madison, WI, 2006.