Sparsity Preserving Discriminant Analysis for Single Training Image Face Recognition

Lishan Qiao[1,2], Songcan Chen[1,*], Xiaoyang Tan[1]

[1] *Department of Computer Science and Engineering, Nanjing University of Aeronautics & Astronautics, 210016, Nanjing, P.R. China*

[2] *Department of Mathematics Science, Liaocheng University, 252000, Liaocheng, P.R. China*

Abstract: Single training image face recognition is one of main challenges to appearance-based pattern recognition techniques. Many classical dimensionality reduction methods such as LDA have achieved success in face recognition field, but can not be directly used to the single training image scenario. Recent graph-based semi-supervised dimensionality reduction (SSDR) provides a feasible strategy to deal with such problem. However, most of the existing SSDR algorithms such as Semi-supervised Discriminant Analysis (SDA) are locality-oriented and generally suffer from the following issues: 1) they need a large number of unlabeled training samples to estimate the manifold structure in data, but such extra samples may not be easily obtained in a given face recognition task; 2) they model the local geometry of data by the nearest neighbor criterion which generally fails to obtain sufficient discriminative information due to the high-dimensionality of face image space; 3) they construct the underlying adjacency graph (or data-dependent regularizer) using a fixed neighborhood size for all the sample points without considering the actual data distribution. In this paper, we develop a new graph-based SSDR algorithm called ***Sparsity Preserving Discriminant Analysis*** (SPDA) to address these problems. More specifically, 1) the graph in SPDA is constructed by sparse representation, and thus the local structure in data is automatically modeled instead of being manually predefined. 2) With the natural discriminative power of sparse representation, SPDA can remarkably improve recognition performance only resorting to very few extra unlabeled samples. 3) A simple ensemble strategy is developed to accelerate graph construction, which results in an efficient ensemble SPDA algorithm. Extensive experiments on both toy and real face data sets are provided to validate the feasibility and effectiveness of the proposed algorithm.

Key words: Linear discriminant analysis, Semi-supervised discriminant analysis, sparse representation, graph construction

[*] Corresponding author: Tel: +86-25-84896481 Ext. 12221; Fax: +86-25-84892400; E-mail: s.chen@nuaa.edu.cn (S. Chen)

1 Introduction

One of the major challenges to appearance-based face recognition is the small sample size (SSS) problem (Duda, Hart et al. 2001). In particular, in many practical applications such as law enforcement, driver license or passport card identification, usually only one labeled sample per person is available. Under such scenario, most of the traditional methods including Eigenface (Turk and Pentland 1991) and Fisherface (Belhumeur, Hespanha et al. 1997) will suffer serious performance drop or even fail to work. Therefore, special tricks, such as synthesizing virtual sample (Beymer and Poggio 1995) and localizing the training image (Chen, Liu et al. 2004), are generally required to deal with the single training sample problem. One can refer to a recent survey (Tan, Chen et al. 2006) for more details on this topic.

Although much success has been achieved by synthetic sample techniques, such artificial process has trouble in capturing the real face data distribution due to the variations of pose, illumination and facial expression (Tan, Chen et al. 2006). An alternative and more natural way to deal with such problem is semi-supervised dimensionality reduction (SSDR) if considerable unlabeled samples are available. For example, the recent Semi-supervised Discriminant Analysis (SDA) (Cai, He et al. 2007) , a semi-supervised extension of typical Linear Discriminant Analysis (LDA), has been successfully applied to the *single training image face recognition problem*[1]. Besides SDA, researchers have developed some special SSDR algorithms, such as SSLDA (Song, Nie et al. 2008), SSMMC (Song, Nie et al. 2008), lapLDA (Chen, Ye et al. 2007), etc., and reported that the semi-supervised extensions can generally improve the performance over their supervised counterparts like LDA and MMC (Li, Jiang et al. 2006; Liu, Chen et al. 2007). Despite being independently proposed, these SSDR algorithms share similar starting point and can be unified under a graph-based dimensionality reduction framework (Yan, Xu et al. 2007; Song, Nie et al. 2008). We will give a brief review on these methods in the next section.

Despite the success of many graph-based SSDR algorithms in dealing with partially labeled face recognition problem (Cai, He et al. 2007; Song, Nie et al. 2008), there are still some problems that are not properly addressed, especially under the single labeled training image scenario. In

[1] Strictly speaking, it should be called "single ***labeled*** image face recognition problem". We abuse the terminology, i.e., single ***training*** image face recognition problem, just for keeping consistent with the expression in (Cai, He et al. 2007).

particular,

1) Many existing graph-based SSDR algorithms are based on manifold assumption, implying that sufficiently many samples are required to characterize the data distribution (Belkin, Niyogi et al. 2006). For example, with a large number of auxiliary unlabeled training samples, SDA can remarkably improve the performance of LDA. However, it is generally uneasy to obtain a sufficient sampling for intrinsic high-dimensional data such as face images[2]. Therefore, a natural question is: ***can we improve the performance of LDA just with very few extra unlabeled samples?***
2) As pointed out in (Zhu 2008), although graph is at the heart of the graph-based semi-supervised methods, its construction has not been studied extensively. Most of the current algorithms such as SDA and lapLDA construct their adjacency graphs by the nearest neighbor criterion on raw data set. However, the nearest neighbor criterion generally fails to obtain sufficiently discriminative information due to its poor performance in the original high-dimensional face space.
3) The underlying adjacency graphs (or data-dependent regularizers) involved in many SSDR algorithms are artificially defined beforehand and use a fixed neighborhood size for all the sample points. Not only does this ignore the actual data distribution, but also bring the difficulty of parameter selection, especially when only few labeled samples are available as in single training image face recognition.

To address the above issues, in this paper, we present a new graph-based SSDR algorithm called ***Sparsity Preserving Discriminant Analysis*** (SPDA) which is motivated by the recent progress in sparse representation (Qiao, Chen et al. 2009; Wright, Yang et al. 2009). Concretely, we highlight the favorable properties of SPDA and main contributions of this paper:

1) SPDA can remarkably improve the performance of typical LDA ***only resorting to very few extra unlabeled samples***, because it does not based on manifold assumption, but mainly focuses on the discriminative power which can be naturally achieved by minimizing a ℓ_1-regularization objective function. We will give a detailed discussion on this point in

[2] A recent research (Meytlis and Sirovich 2007) has shown that the face space is estimated to have at least 100 dimensions.

section 3.

2) Graph construction involved in SPDA relies on sparse representation classification criterion (Wright, Yang et al. 2009) which is generally superior to the nearest neighbor criterion, especially for high-dimensional data.
3) The "neighborhood" size and edge weight for each sample are automatically obtained in one single step by a ℓ_1 optimization problem. As a result, different sample will get different neighborhood sizes, which is more adaptive to complex data distribution.
4) Alternatively, we develop a simple ensemble SPDA algorithm to reduce the computational complexity involved in obtaining sparse representation for graph construction when a large number of unlabeled samples are provided. Also, as a byproduct, we formulate the kernelized version of SPDA.
5) The idea behind SPDA is quite general, and can potentially be extended to other graph-based semi-supervised learning algorithms by integrating with different discriminant criteria or loss functions.

The rest of the paper is organized as follows. Section 2 briefly reviews several existing graph-based SSDR algorithms. In section 3, we develop a new data-dependent regularizer and SPDA algorithm. In Section 4, we extend SPDA to kernel and ensemble versions. Section 5 shows the experimental results, followed by the conclusion and future work in Section 6.

2 Brief review of semi-supervised dimensionality reduction (SSDR)

Firstly, we want to make clear that why we employ SSDR instead of other Semi-supervised learning (SSL) algorithms for the single training image face recognition problem. Indeed, various SSL algorithms have been developed in the past few years. One can refer to (Zhu 2008) for a detailed literature survey. However, as pointed out in (Cai, He et al. 2007), many of the existing SSL algorithm can only work on ***transductive*** setting, which requires both the training and test set are available during the learning process. Therefore, they are not always suitable for face recognition applications where the test set is generally not available during the training phrase. In contrast, SSDR first learns a subspace from the available training set (containing labeled and unlabeled samples), and then the forthcoming test sample is projected onto the subspace for further decision.

2.1 Semi-supervised Discriminant Analysis (SDA) (Cai, He et al. 2007)

SDA extends LDA to incorporate the manifold structure illustrated by both labeled and unlabeled data. Therefore, SDA aims to best preserve the discriminative information as well as the geometric structure in data. Given a set of data points $X = [x_1, x_2, \cdots, x_n]$ including both labeled and unlabeled samples, the SDA objective function is defined as follows:

$$\max_w \frac{w^T S_b w}{w^T S_t w + \lambda_1 w^T w + \lambda_2 J_{\mathrm{MR}}(w)} \tag{1}$$

where, S_b and S_t are respectively the inter-class and total scatter matrix calculated using the labeled training samples. $w^T w$ is the Tikhonov regularizer, and $J_{\mathrm{MR}}(w)$ is a data-dependent manifold regularizer (Belkin, Niyogi et al. 2006). λ_1 and λ_2 are two parameters, controlling the balance among the three terms in denominator. Obviously, if $\lambda_1 = \lambda_2 = 0$, SDA becomes the standard LDA; if $\lambda_1 \neq 0, \lambda_2 = 0$, it becomes the Regularized Discriminant Analysis (RDA) (Hastie 2009).

The data-dependent regularizer in SDA plays a role in preserving the manifold structure in data. It is constructed using both labeled and unlabeled training samples as follows:

$$J_{\mathrm{MR}}(w) = w^T XLX^T w = \sum_{i,j} (w^T x_i - w^T x_j)^2 p_{ij} \tag{2}$$

where *L* is the graph laplacian (Belkin and Niyogi 2003), p_{ij} is the edge weight between data point x_i and x_j. In particular,

$$p_{ij} = \begin{cases} \exp\{\| x_i - x_j \| / 2\sigma^2\}, x_i \in N_k(x_j) \vee x_j \in N_k(x_i) \\ \quad 0 \qquad\qquad , \qquad otherwise \end{cases} \tag{3}$$

Since SDA shares the similar objective function to LDA, one can solve SDA by the following generalized eigenvalue problem:

$$S_b w = \eta (S_t + \lambda_1 I + \lambda_2 XLX^T) w \tag{4}$$

2.2 Other SSDR algorithms

Although, in the recent years, many graph-based SSDR algorithms have been proposed independently, most of them share the same idea: the labeled sample points are used to maximize the discriminative power, while the unlabeled sample points are used to best preserve the geometric structure in data. As a result, they are similar to each other with different choices of discriminant criterion and regularization term. Table 1 gives several popular examples of those methods.

Table 1. Several special SSDR algorithms proposed recently

algorithms	Discriminant criterion		Regularization term	
	Fisher	MMC*	Tiknonov	Manifold
SDA (Cai, He et al. 2007)	√		√	√
LapLDA (Chen, Ye et al. 2007)	√			√
SSLDA(Song, Nie et al. 2008)	√		√	√
SSMMC (Song, Nie et al. 2008)		√	√	√

*Maximum Margin Criterion

Since the discriminant criteria (e.g., Fisher criterion and MMC) are usually off-the-shelf, the data-dependent regularizer naturally plays an important role in the graph-based SSDR algorithms. Also, we notice that the data-dependent regularizer is generally determined by a graph constructed based on both labeled and unlabeled samples. For example, in SDA, the manifold regularizer roots in the above mentioned *k*-neighborhood graph (3). Therefore, in the next section, we will start to introduce our SPDA algorithm from constructing a novel graph.

3 Sparsity preserving discriminant analysis

3.1 Graph construction based on sparse representation

3.1.1 Motivation from sparsity

We first give the reasons why the sparse representation is suitable to graph construction.

1) ***Sparsity plays an important role in typical k-neighborhood graph.*** On one hand, sparsity implicitly characterizes the locality of data distribution; on the other hand, it can effectively save computational cost and storage space. However, for the typical *k*-neighborhood graph constructed by eqn. (3), its sparsity depends on artificially fixed neighborhood size. It seems to be unreasonable that all data points share an identical *k*, which may not characterize the manifold structure well, especially in undersampling case. This motivates us to consider

whether we can automatically learn the sparsity from the data instead of artificial predefinition.

2) ***The sparsest representation is naturally discriminative.*** Since our ultimate goal is classification, we expect that the graph can contain as much discriminative information as possible. That is, two data points are linked by an edge if they are likely from the same class. For the typical *k*-neighborhood graph, this desirability depends heavily on how well the nearest neighbor criterion works in original space (Chen, Chang et al. 2005). Unfortunately, the nearest neighbor criterion does not generally achieve good performance for raw high-dimensional data, e.g. face images (Meytlis and Sirovich 2007). In contrast, the recent researches (Wright, Yang et al. 2009) showed that sparse representation has natural discriminative power and can work well under high-dimensional scenario. Moreover, the discriminative power is closely related to the class numbers rather than the sample numbers (Wright, Yang et al. 2009). ***As a result, we might construct a graph which contains considerable discriminative information without requiring abundant unlabeled samples.***

3.1.2 The objective function for graph construction

Instead of considering *k*-neighborhood and the pairwise similarity as in typical graph construction, we attempt to automatically construct a graph G and make it well preserve discriminative information based on sparse representation.

Given a set of sample points $X = [x_1, x_2 \cdots, x_n]$, where $x_i \in R^m, i = 1,2,\cdots,n$, we expect to reconstruct each sample point x_i using as few data points in X as possible. This can be expressed by the following ℓ_0-minimization problem:

$$\begin{aligned} &\min_{s_i} \| s_i \|_0 \\ &s.t.\ \ x_i = Xs_i \end{aligned} \qquad (5)$$

where $s_i = [s_{i1},\cdots,s_{i,i-1},0,s_{i,i+1},\cdots,s_{in}]^T$ is a *n*-dimensional column vector in which the *i*-th element is equal to zero, implying the x_i is removed from X, and the element $s_{ij}, j \neq i$ denotes the contribution of x_j for reconstructing x_i. It is well known, (5) is a NP-hard problem. Here, we bypass this difficulty by solving the following ℓ_1 optimization problem[3]:

$$\begin{aligned} &\min_{s_i} \| s_i \|_1 \\ &s.t. \;\; x_i = Xs_i \end{aligned} \tag{6}$$

where ℓ_1 is used instead of ℓ_0. It can be effectively solved by linear programming. Recent researches showed that if the optimal solution sought is sparse enough, the solution of ℓ_0 minimization problem is equal to the solution of ℓ_1 minimization problem (Baraniuk 2007). After obtaining all of the optimal reconstruction coefficient $\hat{s}_i$ for each x_i, we construct a sparse weight matrix S by

$$S = [\hat{s}_1, \hat{s}_2, \cdots, \hat{s}_n] \tag{7}$$

Then, the new constructed graph $G = \{X, S\}$, where X is the training sample set, S is the edge weight matrix.

In practice, the constraint $x_i = Xs_i$ in (6) does not always hold due to noise or insufficient training samples. We extend it by incorporating a reconstructive compensation term t_i as follows:

$$\begin{aligned} &\min_{s_i} \| s_i \|_1 + \lambda \| t_i \|_p \\ &s.t. \;\; x_i = Xs_i + t_i \end{aligned} \tag{8}$$

which is equivalent to

$$\min_{s_i} \| s_i \|_1 + \lambda \| x_i - Xs_i \|_p \tag{9}$$

where $t_i = x_i - Xs_i \in R^m$ can be seen as a compensation (or error tolerance) for reconstructing x_i. $\| t_i \|_p$ denotes the ℓ_p-norm, a special measure of the compensation t_i. From the Bayesian viewpoint, $\| t_i \|_p$ essentially corresponds to different prior distribution (or assumption) about t_i. For example, $\| t_i \|_2$ is related to Gaussian prior[4], while $\| t_i \|_1$ is related to Laplacian prior[5]. λ

[3] In fact, suboptimal solutions can be found by a variety of strategies such as greedy-based (Mallat and Zhang 1993) and Bayesian-based (Ji, Xue et al. 2008) methods. Here, we consider the ℓ_1 strategy simply due to that the equivalence of the ℓ_0 and ℓ_1 problem has been studied deeply from a mathematical perspective.

[4] In this sense, the (9) has the same mathematical expression as the popular LASSO (Tibshirani 1996) in statistics.

is a regularized parameter to control the trade-off between the sparsity of reconstructive coefficient and the reconstructive compensation. Although the parameter selection problem has been studied in-depth (Mallat and Zhang 1993; Ji, Xue et al. 2008; Hastie 2009), there is currently no reliable method in theory to assign optimal value for λ. Therefore, we simply set $\lambda = 1$ in all our experiments.

It is worthwhile to point out that the above graph construction manner differs from the one in our previous SPP algorithm (Qiao, Chen et al. 2009), though both are motivated by sparse representation (Wright, Yang et al. 2009). More specifically, 1) in SPP we used two ***independent*** sparse representation models, which are directly developed from (Wright, Yang et al. 2009). In contrast, here we reveal the inherent relationship between these models and unify them in one single objective function (8). As a result, the graph construction models behind SPP are just two special instances of (8), and we can develop new graph construction model from the unified objective according to different priors. 2) In SPP, we require sum-to-one constraint as in LLE (Roweis and Saul 2000). However, we ignore such constraint in SPDA, since we mainly concern discrimination. Not only does this save computational cost, but also, more fortunately, we achieve higher recognition rate than SPP (see table 4 in section 5 for details).

3.2 Sparsity preserving regularization

Now we propose the new data-dependent regularizer based on the previously-constructed graph G=$\{X, S\}$. Revisiting the manifold regularizer (2) in SDA, it implies that if x_i and x_j are "close" to each other, then their low-dimensional representation $y_i = w^T x_i$ and $y_j = w^T x_j$ should be close to each other as well. However, for the newly constructed graph *G*, its edge weight $\hat{s}_{ij}$ is not a rigorous similarity measure, and thus we can not construct the data-dependent regularizer as in SDA.

Note that the relationship between x_i and x_j is characterized by $x_i \approx \sum_{j=1}^{n} \hat{s}_{ij} x_j$ instead of

[5] It has been validated applicable to face images with partial occlusion (Wright, Yang et al. 2009). In this paper, we use this prior by empirically modeling the variations of expression and illumination as partial corruption on clear face images.

simple “closeness”, and hence we expect that their low-dimensional representations y_i and y_j preserve such relationship as well, i.e., $y_i \approx \sum_{j=1}^{n} \hat{s}_{ij} y_j$, which is motivated by LLE (Roweis and Saul 2000). Therefore, we propose the data-dependent regularizer by minimizing the following objective function:

$$J_{Sparsity}(w) = \sum_{i=1}^{n} \| y_i - Y\hat{s}_i \|^2 = \sum_{i=1}^{n} \| w^T x_i - w^T X\hat{s}_i \|^2 \tag{10}$$

where $Y = [y_1, y_2, \cdots, y_n]$ is the low-dimensional representation of the original data. Since the regularizer aims to preserve the sparse reconstructive relationship, we call it ***sparsity preserving regularizer.*** Then, with simple algebraic formulation (see **appendix**), it can be rewritten as

$$w^T X L_s X^T w \tag{11}$$

where $L_s = I - S - S^T + SS^T$. Although, the data-dependent regularizer can potentially be incorporated into many semi-supervised learning algorithms, we only focus on SSDR in this paper.

3.3 Sparsity preserving discriminant analysis (SPDA)

Similar to SDA, we extend LDA[6] to semi-supervised version based on the newly proposed data-dependent regularizer. Naturally, the objective function can be defined as follows:

$$\max_{w} \frac{w^T S_b w}{w^T (S_t + \lambda_1 I + \lambda_2 X L_s X^T) w} \tag{12}$$

where, S_b and S_t are respectively the inter-class and total scatter matrices, which are calculated just using the labeled training samples. I is an identity matrix related to Tikhonov regularizer, and $w^T X L_s X^T w$ is the sparsity preserving regularizer. The solution of (12) can be easily achieved by the following generalized eigenvalue problem.

$$S_b w = \eta (S_t + \lambda_1 I + \lambda_2 X L_s X^T) w \tag{13}$$

The algorithmic procedure is shown as follows. Concretely, we assume the training sample

[6] Of course, we can consider other discriminant criteria such as MMC if necessary.

set $X=[x_1,\cdots,x_l,x_{l+1},\cdots,x_{l+u}]=[X_L,X_U]$, where the first *l* training samples $\{x_i\}_{i=1}^{l}$ are labeled and from *c* classes (there are l_k samples in the *k*-th class), the last *u* training samples $\{x_i\}_{i=l+1}^{l+u}$ are unlabeled. Without loss of generality, the sample points in X_L are ordered according to their labels.

Algorithm 1. Sparsity Preserving Discriminant Analysis

Step1. Calculate $S_b=X_L H_L X_L^T$ and $S_t=X_L X_L^T$ based on the labeled training samples in X_L, where, $H_L=diag(H^1,H^2,\cdots,H^c)$ is a block-diagonal matrix, and H^k is a $l_k \times l_k$ matrix with all elements equal to $1/l_k$.

Step2. Construct graph *G*=$\{X,S\}$. The weight matrix S is calculated based on all training samples in X using (6) or (8).

Step3. Calculate the data-dependent (sparsity preserving) regularizer $w^T X L_s X^T w$, where $L_s=I-S-S^T+S^T S$.

Step4. Calculate the projections by the generalized eigenvalue problem (13), and the projection matrix $W=[w_1,w_2,\cdots,w_d]$, where w_i are the eigenvectors corresponding to the largest *d* eigenvalues.

4 Extensions of SPDA

In this section, we extend the proposed algorithm to its kernelized version (for improving the flexibility of SPDA) and ensemble version (for reducing the computational complexity), respectively.

4.1 Kernel SPDA

As described above, SPDA only focuses on linear dimensionality reduction, and thus it may fail to deal with the highly nonlinear structure in data. Fortunately, we can easily extend SPDA to perform in Reproducing Kernel Hilbert Space (RKHS) like other graph-based dimensionality reduction algorithms.

Let $\phi: x \rightarrow F$ be a function mapping the data points in the input space to the feature space.

According to the kernel trick, we expect to replace the explicit mapping with the inner product $K(x_i, x_j) = (\phi(x_i) \cdot \phi(x_j))$. Furthermore, we assume $\phi_L = [\phi(x_1), \phi(x_2), \cdots, \phi(x_l)]$, $\phi_U = [\phi(x_{l+1}), \phi(x_{l+2}), \cdots, \phi(x_n)]$ and $\phi = [\phi_L, \phi_U]$, then the inter-class scatter matrix and the total scatter matrix in the feature space can respectively be denoted as

$$S_b^F = \phi_L H_L \phi_L^T = \phi \begin{pmatrix} H_L & 0 \\ 0 & 0 \end{pmatrix} \phi^T = \phi H \phi^T, \quad S_t^F = \phi_L \phi_L^T = \phi \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} \phi^T = \phi \widetilde{I} \phi^T \tag{14}$$

where H_L is defined as in the SPDA algorithm.

According to the Representer Theorem (Scholkopf, Herbrich et al. 2001), the projection w^F sought in feature space can be expressed as $w^F = \phi\alpha$, where $\alpha = [\alpha_1, \alpha_2, \cdots, \alpha_n]^T$ is a coefficient vector that represents w^F in the feature space. Let $K = \phi^T \phi$ be the kernel matrix, the objective function of kernel SPDA can be expressed as follows:

$$\max_{w} \frac{\alpha^T KHK^T \alpha}{\alpha^T (K\widetilde{I}K^T + \lambda_1 K + \lambda_2 KL_s K^T)\alpha} \tag{15}$$

The optimal solution $\hat{\alpha}$ can be obtained by solving the following generalized eigenvalue problem:

$$KHK^T \alpha = \eta (K\widetilde{I}K^T + \lambda_1 K + \lambda_2 KL_s K^T)\alpha \tag{16}$$

Thus, given a new data point x, its low-dimensional representation is $(w^F)^T \phi(x) = \hat{\alpha}^T K(\cdot, x)$, where $K(\cdot,\cdot)$ is a kernel function.

4.2 Ensemble SPDA

According to (Wright, Yang et al. 2009), the sparsity of the ideal solution sought is mainly related to the class numbers rather than the sample numbers. Therefore, intuitively, a large number of unlabeled samples do not necessarily help improve the performance of SPDA significantly, and conversely incur high computational burden since SPDA constructs graph based on all the training samples. Here, we introduce a very simple ensemble strategy to speed up the proposed SPDA algorithm based on the above observation.

In particular, given a set of training samples $X = [x_1, \cdots, x_l, x_{l+1}, \cdots, x_{l+u}] = [X_L, X_U]$ as mentioned before, we randomly partition the unlabeled sample set X_U into q small sample sets $X_{U1}, X_{U2}, \cdots, X_{Uq}$, and thus we can generate a series of new training sets $X_1 = [X_L, X_{U1}], \cdots, X_q = [X_L, X_{Uq}]$. Then, we perform SPDA on each new training sets and the test sample is classified by voting strategy. Concretely, the ensemble SPDA algorithmic procedure is shown as follows.

Algorithm 2. ensemble SPDA

```
Step1.  Partition the training set $X$ into q sub-sets $X_1, X_2, \cdots, X_q$.
Step2.  Implement SPDA on each sub-set, and get q subspaces.
Step3.  Project the test sample $x$ onto each subspace, and then implement classification (e.g.,
        1NN) on each subspace.
Step4.  Vote to decide the class label of the test sample.
```

5 Experiments

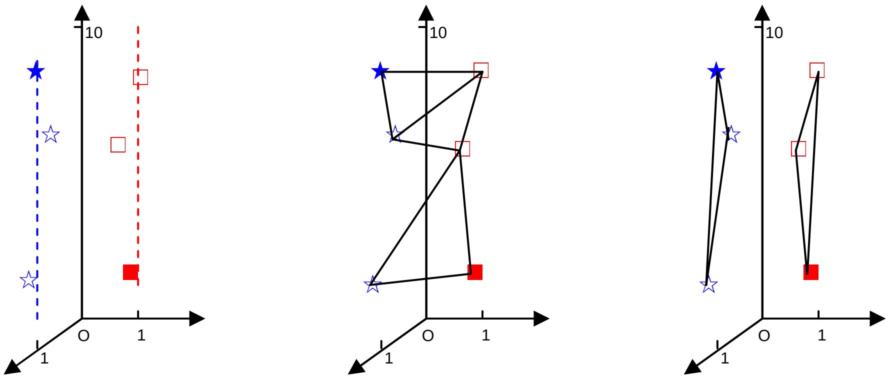
5.1 Illustrative examples

In this subsection, we intuitively illustrate why the proposed algorithm might work well through two illustrative experiments on toy (5.1.1) and face (5.1.2) data sets, respectively.

5.1.1 Illustrative experiment on toy data

For simplicity of our illustration here, we only consider binary classification problem and assume that each class lies in a 1-dimensional subspace embedding in 3-dimensional ambient space. We randomly sample 3 (1 labeled and 2 unlabeled) data points from per class for training. Fig. 1(a) gives an instance of so-generated training sample points which are respectively signed with pentacle and square. In order to approximate practical problem, the data points are corrupted by Gaussian additive white noise with standard deviation 0.1.

Based on the training data, we construct the typical neighborhood graph and the sparse reconstruction graph, respectively. In particular, Fig. 1(b) gives the neighborhood graph, where the neighborhood size *k*=2. It is easy to see that the edges on the graph link the data points which are close to each other, yet from different class. Obviously, other locality-oriented graph construction manners such as the one involved in LLE may also suffer from the fact that the samples from different class are close to each other. In contrast, Fig. 1(c) gives the sparse reconstruction graph behind SPDA. With the sparsity constraint, the non-zero reconstructive coefficients for a given data point more possibly match the data points in the same class. As a result, so-constructed graph tends to contain more discriminative information than typical neighborhood graph. More specifically, we classify 100 randomly generated test samples using 1-Nearest Neighbor (1NN) classifier on the obtained 1-dimensional subspace by SDA and SPDA, respectively. The average classification accuracies corresponding to SDA and SPDA are 70.34% and **87.22%,** respectively.



(a) The data distribution (b) Graph for SDA (c) Graph for SPDA

Fig. 1 Toy problem

5.1.2 Illustrative experiment on face data

The previous toy problem showed that locality-oriented graph construction manners may affect the performance significantly. How about on real-world data sets? Here, we take AR database[7] as an example to compare the proposed algorithm with LLE, since the graph behind SPDA is constructed by ℓ_1-minimization optimization problem which is closely related to the least square graph construction hidden in LLE.

More specifically, we assume the face data set $X = [x_1, x_2, \cdots, x_n]$, where the samples are ordered according to their labels for the convenience of illustration. Then, given a face image $x_i \in X$, by solving (8) or (9), we obtain a sparse reconstruction coefficient s_i in which

[7] See the next subsection for the description about this database. Here we just used the face images taken in the first session.

the nonzero values model the contribution of each support point[8] or support face to represent x_i. In other words, simply by solving (8) or (9), we get both the graph and its corresponding edge weights simultaneously, which is contrary to the scheme of LLE, where the graph and its edge weights are estimated separately. In particular, for a specific face image, Fig. 2 gives an illustration of support faces and corresponding coefficient found by SPDA and those by LLE.

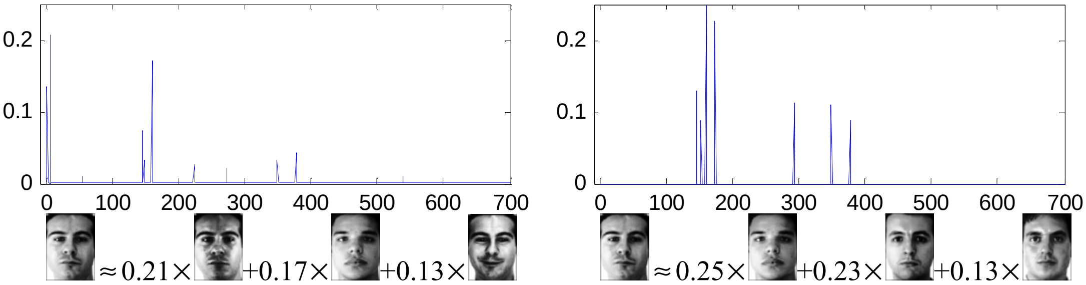


Figure 2: Illustration of three support faces with corresponding coefficients (bottom row) for a given face image (the first image on the left side of the approximately equal mark in bottom row) and the reconstruction coefficient distribution (upper row) using SPDA (left) and LLE (right), respectively.

From the experiment result, we note that the support faces found by SPDA with a ℓ_1-minimization criterion are more discriminative than those by LLE with the least square criterion – two of three faces with the same identity as the prototype are correctly found by SPDA scheme.

5.2 One (labeled) training image face recognition

In this subsection, we perform one training image face recognition experiments on three publicly available face databases: CMU PIE, Extended Yale B and AR databases.

5.2.1 Database description

CMU PIE face database contains 68 subjects with 41,368 face images as a whole. The face images were captured under varying pose, illumination and expression. Similar to (Cai, He et al. 2007), in the experiment, we choose the frontal pose (C27) with varying lighting which leaves us 43 images per person. The size of each face image is cropped to have 32x32 pixels as shown in Fig. 3(top).

[8] Here, support point denotes the face image which contributes to represent the given face image.

Extended Yale B database (Lee, Ho et al. 2005) contains 2414 front-view face images of 38 individuals. For each individual, about 64 pictures were taken under various laboratory-controlled lighting conditions. In our experiments, we simply use the cropped images[9] with the resolution of 32x32 as shown in Fig. 3(middle). The database may be substantially more challenging than the above PIE database due to much larger illumination variations.

AR database consists of over 4000 face images of 126 individuals. For each individual, 26 pictures were taken in two sessions (separated by two weeks) and each section contains 13 images. These images include front view of faces with different expressions, illuminations and occlusions. In our experiments, we only use the images without occlusion in the AR face database provided and preprocessed by (Martinez and Kak 2001). This sub-dataset contains 1400 face images corresponding to 100 person (50 men and 50 women), where each person has 14 different images taken in two sessions. The original resolution of these image faces is 165x120. Here, for computational convenience, we resize them to 66x48 as shown in Fig. 3(bottom).

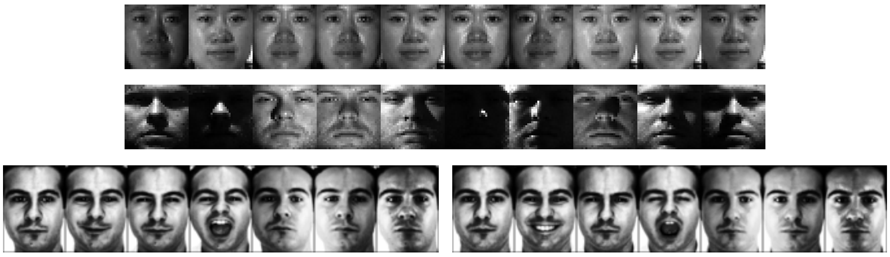


Fig. 3 Some face images from PIE (top), Yale B (middle) and AR (bottom) databases.

5.2.2 Experimental setting

On each face database, we perform two groups of experiments with different unlabeled training sample numbers. Table 2 gives the specific experimental setting. For example, for PIE database, experiment 1 denotes that 3 images are randomly selected from each class as the training set, and the rest images as the testing set. Among the 3 training images, only 1 image is randomly selected and labeled, which leaves the rest 2 images unlabeled; while, in experiment 2, the labeled training samples keep the same, but the number of the total training samples per subject increases to 30. In fact, the experiment 2 is also considered in (Cai, He et al. 2007) where the authors justified that their proposed SDA algorithm achieved better performance than some popular algorithms, e.g. LPP (He and Niyogi 2003) and LapSVM (Belkin, Niyogi et al. 2006). For all the experiments here,

[9] We directly download the cropped image data from http://www.cs.uiuc.edu/homes/dengcai2.

we report the averaged results over 30 random training/test splits.

Table 2 Data set description and partition

Database	Sample sizes per class	Class numbers	Experiment 1		Experiment 2	
			Train	Labeled	Train	Labeled
PIE	64	68	3	1	30	1
Yale B	43	38	3	1	30	1
AR	14	100	3	1	10	1

Based on each data partition, we compare SPDA with Baseline, unsupervised SPP, supervised LDA[10] and semi-supervised SDA. The baseline approach denotes the 1NN classifier on the original face space without dimensionality reduction. For LDA, the face subspace is learnt only using the labeled samples; for SPP, the face subspace is learnt using all the training samples without label information; for SDA and SPDA, the face subspace is learnt using both labeled and unlabeled samples. Then, based on the learnt subspace, 1-NN classifier is employed to evaluate the recognition rate on the test data. As descript previously, SPDA suffers from high computational cost when a large number of unlabeled samples are considered. Therefore, for the experiment 2 on each database, we adopt the ensemble version of SPDA where the unlabeled samples are simply and randomly partitioned into q=10 small subsets.

5.2.3 Parameter selection

LDA and SPP are both parameter-free. SDA contains 4 parameters: two regularized parameters and two free parameters for graph construction. Here, we use the same parameter values for SDA as in (Cai, He et al. 2007). For convenience of comparison, the two regularized parameters in SPDA are assigned the same values as in SDA. In addition, for all the above algorithms, the subspace dimension is set to c-1, where c is the class number. Table 3 gives specific parameter values for SDA and SPDA.

Table 3. Parameter setting for SDA and SPDA

Algorithms	Reg. para. λ_1	Reg. para. λ_2	Neighbor k	Edge weights
SDA	0.01	0.1	2	Cosine
SPDA	0.01	0.1	Auto	Auto

5.2.4 Experimental results and overall observations

[10] Strictly speaking, under the single training sample case, typical LDA fails to work since the intra-class variation cannot be obtained. Here, we simply replace the intra-class scatter matrix using a constant matrix as in (Zhao, Chellappa et al. 1999).

Based on the above experimental setting, table 4 reports the classification accuracies corresponding to different algorithms and databases, where E1 and E2 denote the first and second groups of experiments, respectively.

Table 4 Performance comparison for single training image face recognition problem

		Baseline (%)	LDA (%)	SPP (%)	SDA (%)	SPDA (%)
PIE	E1	25.88 ± 1.2	25.88 ± 1.2	62.55 ± 2.0	30.91 ± 2.1	**67.47** ± 1.8
	E2	26.21 ± 1.6	26.21 ± 1.6	51.29 ± 3.1	59.57 ± 3.2	**70.44** ± 3.0
Yale B	E1	12.60 ± 1.2	12.60 ± 1.2	17.95 ± 3.1	16.10 ± 1.6	**31.27** ± 3.6
	E2	13.01 ± 1.4	13.01 ± 1.4	14.28 ± 3.2	26.77 ± 2.5	**35.44** ± 3.3
AR	E1	24.55 ± 1.4	24.55 ± 1.4	44.57 ± 2.6	22.48 ± 1.6	**58.46** ± 2.0
	E2	24.69 ± 2.4	24.69 ± 2.4	55.06 ± 3.1	26.22 ± 2.1	**61.23** ± 2.5

From the experimental results on the three popular face databases, we can achieve several observations as follows:

1) Among the discussed dimensionality reduction methods, LDA generally achieve relatively low accuracies due to the fact that only one labeled sample per class is used to learning the face subspace.
2) Despite its unsupervised nature, SPP can outperform LDA with the help of extra training samples. However, the performance of SPP does not always be improved with the increase of training samples. Interestingly, SPP can even achieve better performance than SDA in some of the experiments, which benefits from the natural discriminative power of sparse representation.
3) Semi-supervised SDA and the proposed SPDA always outperform LDA if considerable unlabeled training samples are available. That is, the extra unlabeled training samples can generally help improve the performance.
4) SPDA consistently outperforms SPP and SDA on all the used face databases. This illustrates both label information and well-constructed graph (or equivalently, data-dependent regularizer) play important roles in the ultimate recognition rates. More importantly, the proposed SPDA algorithm can remarkably improve the performance of LDA even when only few unlabeled training samples are available.

6 Conclusion and future works

In this paper, we developed a new semi-supervised dimensionality reduction method called

Sparsity Preserving Discriminant Analysis (SPDA). The newly proposed algorithm does not only model the "locality" automatically, but also remarkably improves the performance of typical LDA only resorting to very few additional unlabeled samples. As a result, SPDA algorithm is more applicable to face recognition problem with only a few training samples.

From the experimental results, we can find that SPDA is more effective than the popular SDA algorithm, but has still a big gap from practical face recognition applications. Therefore, in the future work, we will attempt to integrate the typical strategies (e.g. synthesizing virtual samples, localizing the training images) with the proposed algorithm and expect to further improve the performance.

Acknowledgement

This work was partly supported by National Natural Science Foundation of China (60973097, 60773061 and 60773060) and the Natural Science Foundations of Jiangsu Province (BK2009369 and BK2008381). Also thank Dr. Deng Cai for providing the code and related parameters used in SDA.

References

Baraniuk, R. G. (2007). "Compressive Sensing [Lecture Notes]." Signal Processing Magazine, IEEE **24**(4): 118-121.

Belhumeur, P. N., J. P. Hespanha and D. J. Kriegman (1997). "Eigenfaces vs. Fisherfaces: recognition using class specific linear projection." IEEE Transactions on Pattern Analysis and Machine Intelligence **19**(7): 711-720.

Belkin, M. and P. Niyogi (2003). "Laplacian eigenmaps for dimensionality reduction and data representation." Neural Computation **15**(6): 1373-1396.

Belkin, M., P. Niyogi and V. Sindhwani (2006). "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples." Journal of Machine Learning Research **7**: 2399-2434.

Beymer, D. and T. Poggio (1995). Face recognition from one example view. International Conference on Computer Vision (ICCV).

Cai, D., X. F. He and J. W. Han (2007). Semi-supervised discriminant analysis. IEEE International Conference on Computer Vision (ICCV).

Chen, H. T., H. W. Chang and T. L. Liu (2005). Local discriminant embedding and its variants. IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Chen, J. H., J. P. Ye and Q. Li (2007). Integrating global and local structures: A least squares framework for dimensionality reduction. IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Chen, S. C., J. Liu and Z. H. Zhou (2004). "Making FLDA applicable to face recognition with one sample per person." Pattern Recognition **37**(7): 1553-1555.

Duda, R. O., P. E. Hart and D. G. Stork (2001). Pattern classification. New York, Wiley.

Hastie, T. (2009). The elements of statistical learning, second edition : data mining, inference, and prediction. New York, Springer.

He, X. F. and P. Niyogi (2003). Locality preserving projections. Neural Information Processing Systems (NIPS).

Ji, S. H., Y. Xue and L. Carin (2008). "Bayesian compressive sensing." IEEE Transactions on Signal

Processing **56**(6): 2346-2356.

Lee, K. C., J. Ho and D. J. Kriegman (2005). "Acquiring linear subspaces for face recognition under variable lighting." IEEE Transactions on Pattern Analysis and Machine Intelligence **27**(5): 684-698.

Li, H. F., T. Jiang and K. S. Zhang (2006). "Efficient and robust feature extraction by maximum margin criterion." IEEE Transactions on Neural Networks **17**(1): 157-165.

Liu, J., S. C. Chen, X. Y. Tan and D. Q. Zhang (2007). "Comments on "Efficient and robust feature extraction by maximum margin criterion"." IEEE Transactions on Neural Networks **18**(6): 1862-1864.

Mallat, S. G. and Z. Zhang (1993). "Matching pursuits with time-frequency dictionaries." IEEE Transactions on Signal Processing **41**(12): 3397-3415.

Martinez, A. M. and A. C. Kak (2001). "PCA versus LDA." IEEE Transactions on Pattern Analysis and Machine Intelligence **23**(2): 228-233.

Meytlis, M. and L. Sirovich (2007). "On the dimensionality of face space." IEEE Transactions on Pattern Analysis and Machine Intelligence **29**(7): 1262-1267.

Qiao, L. S., S. C. Chen and X. Y. Tan (2009). Sparsity Preserving Projections with Applications to Face Recognition. Pattern Recognition.

Roweis, S. T. and L. K. Saul (2000). "Nonlinear dimensionality reduction by locally linear embedding." Science **290**(5500): 2323-2326.

Scholkopf, B., R. Herbrich and A. J. Smola (2001). A generalized representer theorem. Computational Learning Theory.

Song, Y. Q., F. P. Nie, C. S. Zhang and S. M. Xiang (2008). "A unified framework for semi-supervised dimensionality reduction." Pattern Recognition **41**(9): 2789-2799.

Tan, X. Y., S. C. Chen, Z. H. Zhou and F. Y. Zhang (2006). "Face recognition from a single image per person: A survey." Pattern Recognition **39**(9): 1725-1745.

Tibshirani, R. (1996). "Regression shrinkage and selection via the LASSO." Journal of the Royal Statistical Society: Series B **58**(1): 267-288.

Turk, M. and A. Pentland (1991). "Eigenfaces for Recognition." Journal of Cognitive Neuroscience **3**(1): 71-86.

Wright, J., A. Y. Yang, A. Ganesh, S. S. Sastry and Y. Ma (2009). "Robust face recognition via sparse representation." IEEE Transactions on Pattern Analysis and Machine Intelligence **31**(2): 210-227.

Yan, S. C., D. Xu, B. Y. Zhang, H. J. Zhang, Q. Yang and S. Lin (2007). "Graph embedding and extensions: A general framework for dimensionality reduction." IEEE Transactions on Pattern Analysis and Machine Intelligence **29**(1): 40-51.

Zhao, W., R. Chellappa and P. J. Phillips (1999). Subspace linear discriminant analysis for face recognition. Technical Report CAR-TR-914. University of Maryland.

Zhu, X. (2008). Semi-supervised learning literature survey. Univ. of Wisconsin, Madison.

Appendix: The formulation for sparsity preserving regularizer

$$J_{Sparsity}(w)=\sum_{i=1}^{n}\| y_i - Y\hat{s}_i \|^2 = \sum_{i=1}^{n}\| w^T x_i - w^T X\hat{s}_i \|^2 = w^T\left(\sum_{i=1}^{n}(x_i - X\hat{s}_i)(x_i - X\hat{s}_i)^T\right)w$$

Let e_i be a n-dimensional unit vector with the i-th element 1, 0 otherwise, then the above Equ. is equal to

$$w^T\left(\sum_{i=1}^{n}(Xe_i - X\hat{s}_i)(Xe_i - X\hat{s}_i)^T\right)w$$

$$= w^T X\left(\sum_{i=1}^{n}(e_i - \hat{s}_i)(e_i - \hat{s}_i)^T\right)X^T w$$

$$= w^T X \left(\sum_{i=1}^{n} e_i e_i^T - \hat{s}_i e_i^T - e_i \hat{s}_i^T + \hat{s}_i \hat{s}_i^T \right) X^T w$$

$$= w^T X L_s X^T w$$

where $L_s = I - S - S^T + SS^T$ and $S = [\hat{s}_1, \hat{s}_2, \cdots, \hat{s}_n]$.