# Ensemble of Randomized Linear Discriminant Analysis for Face Recognition with Single Sample per Person

Ying Li
Electrical Engineering, Mathematics and
Computer Science
York university
Toronto, ON,Canada
liying.yunran@gmail.com

Wei Shen
Faculty of Department of Communication
Engineering
Shanghai University
Shanghai, 200072, P. R. China
wei.shen@shu.edu.cn

Xun Shi
Magnum Semiconductor
Waterloo, ON, Canada
shixun@gmail.com

Zhijiang Zhang
Faculty of Department of Communication
Engineering
Shanghai University
Shanghai, 200072, P. R. China
zjzhang@staff.shu.edu.cn

*Abstract*— **Linear Discriminant Analysis (LDA) has been widely used in appearance-based face recognition. However, it requires lots of training samples for each person with respect to the large dimensionality of the image space, which is difficult to collect in reality. To overcome the severe constraint of training sample deficiency, approaches based on single training sample per person (SSPP) arise in the past decades. Though making great improvements for years, these methods still suffer from low accuracy when dealing with high dimensional image features. In this paper, we develop a new variant of LDA that addresses the SSPP problem especially and apply random projections to generate extra useful training samples on an ensemble of low-dimensional subspaces. A novel extension to kernel version is also presented. We demonstrate the functionality of the proposed methods that outperform the state-of-the-arts on several benchmarks of face recognition.**

## I. Introduction

Face recognition has been a hot topic in research area for decades due to its wide range of practical applicability, such as security validation, surveillance and facial recognition access. Many methods have been proposed to address this problem (See a detailed review in [28]). They can roughly be divided into two categories: geometric-based and appearance-based approaches. Nowadays, algorithms belonging to the last category attract most attention in the field. Though they show impressive results when dealing with multiple sample problem (MSP), their performances deteriorate rapidly when the number of training samples decreases [19]. But lacking of adequate training samples is a main technical obstacle for realistic problems, like passport recognition. Under this circumstance, many approaches focusing on single sample per person (SSPP) have been proposed.

All SSPP algorithms have to deal with a main obstacle: how to get as much information as possible from the single sample considering the large appearance variations caused by illumination, expression, pose and so on. It is well-known that Eigenfaces [20] and Fisher Linear Discriminant (FLD) [1] are widely used in MSP algorithms. But with single sample, they fail to work properly. Furthermore, the inter-person variation cannot be estimated accurately, and we can't even compute intra-person variation at all. Another challenge of SSPP lies in how to deal with high dimensional effect when the number of training samples is less than the dimension of instance features. Under this circumstance, the estimation of data variance is often sensitive to the noise or outliers. Even though, several approaches show optimistic results with insufficient samples. [19] provided a detail review of previous SSPP algorithms. Recent advances in SSPP problems include adaptive generic FLD [26] that learns pose or expressions variations from extra datasets and a *stringface* system that recognizes partial occluded faces through a string-based matching algorithm [6].

In this paper, we propose a novel method named as "Ensemble of Randomized Linear Discriminant Analysis"(ERLDA) based on ensemble of randomly selected features of the single training sample. Our intuition is simple: projection on randomly selected features for each sample can generate considerable representations for each person. Fig. 1 illustrates the pipeline of the proposed method. Based on the motivation of ensemble learning, our method utilized that fact that weak learners from randomized down sampled face features provide some interpretation of the signature of faces. Unlike the traditional ensemble learning algorithms, our method uses an unsupervised weight adaption method by solving a convex optimization problem.

Our contribution is four-fold: First, we develop a new one-sample LDA that is suitable for learning with only one sample per class. Theoretical analysis shows that our one-sample LDA is equivalent to a discriminant function that involves a novel *metric* which considers the statistics of all the training set; Second, the proposed method combines the technique of random projections and ensemble learning to make the one-sample LDA capable of capturing multiple views from projected subspaces. Third, we present an extension of Kernel ERLDA which replaces the *metric* as an inner product of two kernel vectors. Proper kernels can improve the performance of ERLDA. Last but not least, we provide extensive experimental results on four well-known benchmarks of face recognition. The functionality of our methods is illustrated by outperforming state-of-the-arts methods on these benchmarks.

## II. Related Work

On the basis of usage of training set, existing methods for the SSPP problem can be divided into three categories. Methods in the first category only make use of single training sample in training set, while approaches in the second category generate virtual images to obtain multiple samples per person. The last category collects an auxiliary generic training set including multiple samples per person along with the original single training set.

For methods in the first category, they enrich the information simply based on the single training sample. One of the most popular branches in this category is extensions of PCA, like 2DPCA [25] and $(PC)^2A$ [5], [23]. 2DPCA uses 2D image matrices directly rather than normal 1D vectors, then obtains the image covariance matrix based on the average image for all samples in the training set. $(PC)^2A$ uses traditional eigenface technique after synthesizing new images based on the horizontal and vertical projections of the original one. Some similar approaches, such as Kernel PCA [2] and singular vector decomposition (SVD) [27], fall into this category too.

As for the second category, extra virtual face images are generated in order to make general learning techniques applicable. In [13], local component features are constructed by moving the original image in various directions. In [5], each face image is partitioned into several sub-images which share the same label. In [8], Gabor filter is used to generate different local features. With these extra training samples, general supervised learning algorithms, e.g., LDA or FLD, can be applied directly. However, these methods above either need prior knowledge to guide the generation of virtual images or often result in generation of outliers, which make them unpractical to use.

For the last category, the basic intuition is that different persons could have very similar intra-person variations. They make use of a generic training set, which can contain as many images as required, besides the single sample training set [15], [26], [22]. Especially in [22], both within-class and between-class scatter matrices are learnt on the generic training set and applied directly for feature extraction. But this kind of methods increase the storage space for training samples and also requires more computation time, which violates a unique advantage of SSPP.

Our method vaguely falls into the second category, though it has two main differences : 1) ERLDA makes use of binary one-sample LDA while most methods in the second category focus to apply multi-class LDA directly. Within the framework of only one training sample per class, one-vs-the-rest scheme can exploit more information about the discriminative structure of different labels. 2) ERLDA randomly projects the down sampled features of a training sample and makes use of ensemble learning technique. This process is much more efficient than most of the methods in this category which need considerable efforts on designing new training samples.

## III. Problem Formulation and Background

In this section, we formulate the task of face recognition with single sample in each class and give a brief introduction to linear discriminant analysis (LDA). Assume there are $C$ different subjects, and in each subject, we only have one known training sample. So the training set consists of $C$ faces $\{\mathbf{x}_1, \ldots, \mathbf{x}_C\}$. Given any test instance $\mathbf{x}_t$, we aim to find the most likely subject that $\mathbf{x}_t$ belongs to:

$$k^* = \arg \max_{1 \le k \le C} P(\mathbf{x}_k | \mathbf{x}_t) \tag{1}$$

Possible data pre-processing includes dimension reduction and normalization. PCA is often used to reduce the high-dimensional features into low-dimensional space. Another common pre-process is to normalize each instance so that every feature has a unit norm. This procedure restricts each instance within a ball of unit norm.

### A. Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is widely used to identify the class labels in a discriminative manner. LDA aims to find the linear features that maximize the between-class separation of data, while minimizing the within-class scatter. Given a training set containing $N$ examples $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, where $\mathbf{x}_i \in \mathbb{R}^D$. Each training example belongs to one of the $C$ classes. Let $\mathcal{C}_k$ represents the set of all examples of class $k$ and let $n_k = |\mathcal{C}_k|$ be the number of examples in class $k$. In LDA, the within-class scatter matrix $S_W$ and the between-class scatter matrix $S_B$ are defined as,

$$S_W = \frac{\sum_{k=1}^{C} \sum_{i \in \mathcal{C}_k} (\mathbf{x}_i - m_k)(\mathbf{x}_i - m_k)^T}{N}, \tag{2}$$

$$S_B = \frac{\sum_{k=1}^{C} n_k (m_k - m)(m_k - m)^T}{N}. \tag{3}$$

where $m_k = \frac{1}{n_k} \sum_{i \in \mathcal{C}_k} \mathbf{x}_i$ is the mean of $k$th class, and $m = \frac{1}{N} \sum_i \mathbf{x}_i$ is the mean of the whole data set. The LDA is

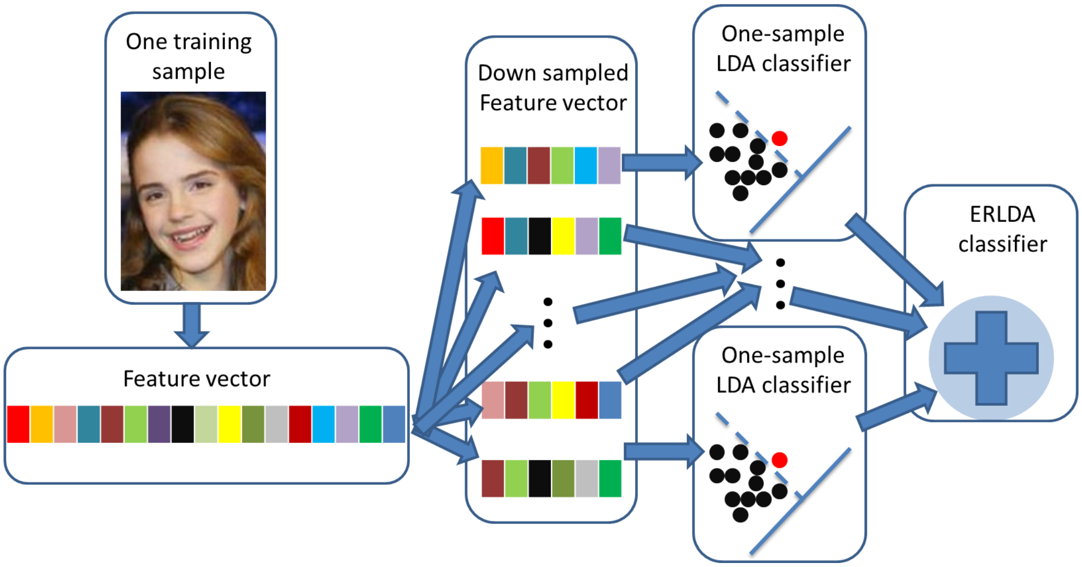

One training sample
Feature vector
Down sampled Feature vector
One-sample LDA classifier
One-sample LDA classifier
ERLDA classifier

Fig. 1. A schematic description of our method.

then formulated as the solution of the following optimization problem:

$$W_{opt} = \arg\max_{W} \frac{\|W^T S_B W\|}{\|W^T S_W W\|}. \tag{4}$$

It can be shown that each column of the optimal $W_{opt}$ is the generalized eigenvector $w$ such that $S_B w = \lambda S_W w$, where $\lambda$ is the eigenvalue of $S_B^{-1} S_W$. One consequence of this result is that LDA correlates the data both between and within classes.

However, for binary class LDA, we have a simple solution:

$$w \propto (S_W)^{-1}(m_1 - m_2). \tag{5}$$

This only applies for two-class cases.

## IV. Ensemble of Randomized LDA

### A. *One-Sample LDA*

The problem in applying LDA directly in face recognition with single sample in each class is the difficulty of calculating $S_W$ in (2) and $S_B$ in (3). Previous work tried to estimate them by using SVD technique to approximate each training sample and thus obtaining one additional sample. However, this line of methods suffer from the insufficiency and inaccuracy of the estimated $S_B$. Another problem lies in the imbalance of multi-class LDA with very few training samples. To overcome these shortcomings, we design an ensemble of randomized LDA to address the challenge of face recognition with single sample.

Binary LDA is used with the one-vs-the-rest technique. Given any test instance $\mathbf{x}_t$, we can apply binary LDA for $C$ times. For the $k$-th class, we can find an optimal projection $w^k$ by (5) as follows:

$$\begin{aligned} w^k &= \Sigma_k^{-1}(\mathbf{x}_k - m_{-k}), \\ w^k &\Leftarrow \frac{w^k}{\|w^k\|_2}, \end{aligned} \tag{6}$$

where $m_{-k} = \frac{\sum_{i \neq k} \mathbf{x}_i}{C-1}$ is the mean of all the examples except $\mathbf{x}_k$. The second step is a normalization of the optimal projection since binary LDA is homogenous in $w^k$.

The scatter matrix $\Sigma_k$ can be estimated as,

$$\Sigma_k = \frac{1}{C-1}\sum_{i \neq k}(\mathbf{x}_i - m_{-k})(\mathbf{x}_i - m_{-k})^T + \gamma \mathbf{x}_k \mathbf{x}_k^T; \tag{7}$$

This estimation is well-situated if we do a preprocess work by normalizing each instance feature. One may argue that $\Sigma_k$ in (7) is not an accurate form of the within-class scatter matrix in the $k$-th class. However, in one-vs-the-rest scheme, we have a lot more “nagative” example than positive one, so what we are facing is not a balanced problem. As shown in Section(V), if we choose a proper value of $\gamma$, we can construct a separable discriminative function.

The score of $s_k = (w^k)^T \mathbf{x}_t$ can be seen as a similarity between $\mathbf{x}_t$ and $\mathbf{x}_k$. By applying softmax technique, we can obtain a posterior form:

$$P(\mathbf{x}_k | \mathbf{x}_t) = \frac{\exp(s_k)}{\sum_{i=1}^{C} \exp(s_i)}. \tag{8}$$

Note that we assume here the prior of each class is uniform.

### B. *Ensemble of One-Sample LDA with Random Projections*

The use of one-vs-the-rest technique is not enough for such a challenging task with only one training sample in each class. Following the line of ensemble learning, we try to find to a set of weak learners. It is best that these weak learners can describe different views of the task. The one-vs-the-rest one-sample LDA is a good choice of weak learner. The only problem is how to make complementary LDAs. One approach is sub-sampling. We randomly sample a subset of dimensions in both training and test features. Any instance $\mathbf{x}_i = \{x_{i1}, x_{i2}, \ldots, x_{iD}\} \in \mathbb{R}^D$ can be sampled into a sparse form $x_i^{(S_j)} = \{x_{iS_j(1)}, x_{iS_j(2)}, \ldots, x_{iS_j(|S_j|)}\}$ where $S_j$ is the index vector which represents the index set of chosen dimensions. Different choices of $S_j$ can be interpreted as different subspaces in the initial feature space. We also apply random projection to make the chosen subspaces anisotropic to each other [11]. Orthogonal projection $\Phi \in \mathbb{R}^{|S_j| \times d}$ is often used in random projections. One advantage about orthogonal projection is the preservation of pair-wise distances. Another two common random projectors are following normal distribution or uniform distribution in the range $(-1, 1)$ [11]. Theoretical analysis in Section (V) shows that these two random projectors preserve the Euclidean distance in a tight bounded ball with a high probability.

By these two techniques, we can construct a set of complementary weak learners $h_j$ which apply one-sample LDA on features that are sub-sampled and randomly projected. So the overall strong classifier is:

$$P(\mathbf{x}_k|\mathbf{x}_t) = \sum_{j}^{J} P(\mathbf{x}_k|\mathbf{x}_t, h_j)P(h_j) \tag{9}$$

where $v_j = P(h_j)$ is the prior of the $j$-th weak learner and $J$ is the number of weak learners. However, it is difficult to learn the prior in a boosting-like way since we only have one training sample per class. Here we develop a novel method to estimate the prior of each weak learner. Denote

$$P(\cdot|\mathbf{x}_t, h_j) = \{P(\mathbf{x}_1|\mathbf{x}_t, h_j), P(\mathbf{x}_2|\mathbf{x}_t, h_j), \ldots, P(\mathbf{x}_C|\mathbf{x}_t, h_j)\}$$

as the distribution of the output by the $j$-th classifier $h_j$. Denote $c_{tj} = H(P(\cdot|\mathbf{x}_t, h_j))$, the entropy of $P(\cdot|\mathbf{x}_t, h_j)$, which can measure the discriminative power of the classifier $h_j$. So we can learn the prior distribution in an unsupervised manner:

$$\mathbf{v}^* = \arg\max_{\mathbf{v}} \sum_{j=1}^{J} v_j c_{tj}^{-1} - \beta H(\mathbf{v}) \tag{10}$$

$$s.t. \sum_{j=1}^{J} v_j = 1, v_j \geq 0. \tag{11}$$

where $\mathbf{v} = \{v_1, ..., v_J\}$, $H(\cdot)$ represents the entropy of a distribution and $\beta$ is a regularization parameter. This is a typical nonlinear convex optimization problem, and we use the package CVX [7].

The choice of prior function is easy to understand. If the entropy $H(P(\cdot|\mathbf{x}_t, h_j))$ is large, e.g. $P(\cdot|\mathbf{x}_t, h_j)$ is an uniform distribution, then it indicates that $h_j$ is unable to discriminate the difference between classes, so we can make the prior/weight of $h_j$ small. We summarize the algorithm of Ensemble of Randomized LDA (ERLDA) in Fig.(2).

## V. Theoretical Analysis

In this section, we provide some theoretical analysis about the proposed method: ERLDA.

*Theorem 5.1:* If $\gamma = 0$, one can find a discriminative function $f(x)$ such that for any $h_j$,

$$\ln(P(\mathbf{x}_k|\mathbf{x}_t, h_j)) - \ln(P(\mathbf{x}_l|\mathbf{x}_t, h_j)) = f(\mathbf{x}_t, \mathbf{x}_k) - f(\mathbf{x}_t, \mathbf{x}_l). \tag{12}$$

*Proof:* Denote $m = \frac{\sum_{i=1}^{C} \mathbf{x}_i}{C}$ , $A = \frac{\sum_{i=1}^{C} \mathbf{x}_i(\mathbf{x}_i)^T}{C}$ and $U_k = m + \mathbf{x}_k$. If $\gamma = 0$, by eqn.(7), we have

$$\begin{aligned}
\Sigma_k &= \frac{1}{C-1}\sum_{i \neq k}(\mathbf{x}_i - m_{-k})(\mathbf{x}_i - m_{-k})^T \\
&= \frac{1}{C-1}\sum_{i \neq k}(\mathbf{x}_i\mathbf{x}_i^T) - m_{-k}m_{-k}^T \\
&= \frac{1}{C-1}(CA - \mathbf{x}_k\mathbf{x}_k^T) - (\frac{Cm - \mathbf{x}_k}{C-1})(\frac{Cm - \mathbf{x}_k}{C-1})^T \\
&= \frac{C}{C-1}A - \frac{C^2}{(C-1)^2}mm^T \\
&\quad -\frac{C}{(C-1)^2}(m\mathbf{x}_k^T + m^T\mathbf{x}_k + \mathbf{x}_k\mathbf{x}_k^T) \\
&= \frac{C}{C-1}A - \frac{C^2}{(C-1)^2}mm^T - \frac{C}{(C-1)^2}(U_kU_k^T - mm^T) \\
&= \frac{C}{C-1}(A - mm^T) - \frac{C}{(C-1)^2}(U_kU_k^T) \\
&= B - \mu(U_kU_k^T)
\end{aligned} \tag{13}$$

where $B = \frac{C}{C-1}(A - mm^T)$ and $\mu = \frac{C}{(C-1)^2}$. Note that $B$ is dependent on all the training set. By Sherman-Morrison formula, we have

$$\Sigma_k^{-1} = B^{-1} + \mu B^{-1}U_k(1 - \mu U_k^T B^{-1}U_k)^{-1}U_k^T B^{-1} \tag{14}$$

By eqn.(8), we have

$$\begin{aligned}
&\ln P(\mathbf{x}_k|\mathbf{x}_t, h_j) - \ln P(\mathbf{x}_l|\mathbf{x}_t, h_j) \\
\propto\ & (w^{(k)} - w^{(l)})^T\mathbf{x}_t \\
\propto\ & (\Sigma_k^{-1}(\mathbf{x}_k - m_{-k}) - \Sigma_l^{-1}(\mathbf{x}_l - m_{-l}))^T\mathbf{x}_t \\
\propto\ & [\mathbf{x}_t^T B^{-1}\mathbf{x}_k + \frac{\mu\mathbf{x}_t^T B^{-1}U_kU_k^T B^{-1}(\mathbf{x}_k - m)}{1 - \mu U_k^T B^{-1}U_k}] \\
& -[\mathbf{x}_t^T B^{-1}\mathbf{x}_l + \frac{\mu\mathbf{x}_t^T B^{-1}U_lU_l^T B^{-1}(\mathbf{x}_l - m)}{1 - \mu U_l^T B^{-1}U_l}]
\end{aligned} \tag{15}$$

so we can obtain

$$\begin{aligned}
f(\mathbf{x}_t, \mathbf{x}_k) &\propto \mathbf{x}_t^T B^{-1}\mathbf{x}_k + \frac{\mu\mathbf{x}_t^T B^{-1}U_kU_k^T B^{-1}(\mathbf{x}_k - m)}{1 - \mu U_k^T B^{-1}U_k} \\
&\propto g(d_{\mathcal{B}}(\mathbf{x}_t, \mathbf{x}_k), d_{\mathcal{B}}(\mathbf{x}_t, m), d_{\mathcal{B}}(\mathbf{x}_k, \mathbf{x}_k), d_{\mathcal{B}}(m, \mathbf{x}_k))
\end{aligned} \tag{16}$$

where $d_{\mathcal{B}}(x, y) = x^T B^{-1}y$. The detailed form of function $g$ can be seen in Appendix. ■

Actually, if $\gamma > 0$, the theorem still holds, but the form of the function $g$ would be much more difficult to analyze. Empirical experience suggests that value of $\gamma$ is of little significance for the final recognition accuracy.

*Lemma 5.2:* For any vectors $x$ and $y$, $d_{\mathcal{B}}(x, y) = x^T B^{-1}y$ is a metric.

*Proof:* It is easy to verify that $B$ is symmetric and positive-definite. So $d_{\mathcal{B}}(x, y) = x^T B^{-1}y$ is symmetric and satisfying the triangular inequality. ■

By Theorem(5.1), our one-sample LDA is aimed at constructing a discriminative function for each subject, which is based on the metric in Lemma(5.2) over the whole training set. Actually, if we let $\mu = 0$, we can see that $f(\mathbf{x}_t, \mathbf{x}_k) = d_{\mathcal{B}}(\mathbf{x}_t, \mathbf{x}_k)$, then our method becomes Nearest-Neighbor (NN) classifier.

```
INPUT The training set $\{\mathbf{x}_1, \dots, \mathbf{x}_C\}$ and a test instance $\mathbf{x}_t$.
1. Sub-sample the training set and test instance on the index set $S_j$ and obtain a new training set $\{x_1^{(S_j)}, \dots, x_C^{(S_j)}\}$ and test instance $\mathbf{x}_t^{(S_j)}$.
2. Randomly choose an orthogonal projection $\Phi_j$.
3. Calculate $P(\mathbf{x}_k|\mathbf{x}_t, h_j)$ , $k = 1, \dots, C$ by applying one-sample LDA on the new training set $\{\Phi_j x_1^{(S_j)}, \dots, \Phi_j x_C^{(S_j)}\}$ and test instance $\Phi_j \mathbf{x}_t^{(S_j)}$:
    3.a Calculate the scatter matrix $\Sigma_k$ by (7).
    3.b Calculate the optimal projection vector $w^k$ by (6).
    3.c Obtain the posterior probability $P(\mathbf{x}_k|\mathbf{x}_t, h_j)$ by (8).
4. Calculate the weight of the $j$-th classifier by solving the optimization problem (10).
5. Calculate the overall posterior probability $P(\mathbf{x}_k|\mathbf{x}_t) = \frac{\sum_j v_j P(\mathbf{x}_k|\mathbf{x}_t, h_j)}{\sum_j v_j}$
OUTPUT The label of test instance $\mathbf{x}_t$ : $L(\mathbf{x}_t) = \arg\max_k P(\mathbf{x}_k|\mathbf{x}_t)$.
```

Fig. 2. Algorithm of Ensemble of Randomized LDA.

*Theorem 5.3:* Let $\mathfrak{Q}$ be the finite collection of training samples in $\mathbb{R}^D$, and $C = |\mathfrak{Q}|$. Fix $0 < \epsilon < 1$ and $\beta > 0$. Let $\Phi$ be a random projector from $\mathbb{R}^D$ to $\mathbb{R}^d$ with

$$d \geq (\frac{4 + 2\beta}{\epsilon^2/2 - \epsilon^3/3}) \ln C,$$

and every element of $\Phi$ follows either a normal Gaussian distribution or a uniform distribution in the region $[-1, 1]$. If $d \leq D$, then for any $x$ and $y$ that belong to $\mathcal{K}(\mathfrak{Q})$, we have

$$Pr\{(1-\epsilon)\sqrt{\frac{d}{D}} \leq \frac{\|\Phi x - \Phi y\|}{\|x - y\|} \leq (1-\epsilon)\sqrt{\frac{d}{D}}\} \geq 1 - C^{-\beta}.$$

This is a direct application of Johnson-Lindenstrauss Lemma [3]. This theorem states that with high probability the random projectors can preserve the pairwise distance within a certain bound.

## VI. Kernel ERLDA

In this section, we provide a kernel variant of ERLDA based on the theoretical analysis in Section(V). By Theorem(5.1) and Lemma(5.2), we extend ERLDA in the Hilbert space. The key idea is that $d_{\mathcal{B}}(x, y)$ is actually a metric (similarity) between $x$ and $y$, so by "*kernel trick*", we can replace $d_{\mathcal{B}}(x, y)$ as

$$d_{Ker}(x, y) = < Ker(x, \cdot), Ker(y, \cdot) >, \tag{17}$$

where $< \cdot, \cdot >$ means the inner product and $Ker(x, \cdot)$ is the kernel vector

$$Ker(x, \cdot) = [K(x, \mathbf{x}_1), K(x, \mathbf{x}_2), \dots, K(x, \mathbf{x}_C)], \tag{18}$$

where $\mathbf{x}_i, i = 1, \dots, C$ are the training samples. $K(x, \cdot)$ is a *kernel* function and a common choice would be "Gaussian":

$$K(x, y) = \exp(-\|x - y\|^2/\sigma), \tag{19}$$

where $\sigma$ is a hyper-parameter.

According to Theorem (5.1), we can make use of the separation function in (16) as a similarity between the test instance and the $k$-th person:

$$s_k = g(d_{Ker}(\mathbf{x}_t, \mathbf{x}_k), d_{Ker}(\mathbf{x}_t, m), d_{Ker}(\mathbf{x}_k, \mathbf{x}_k), d_{Ker}(m, \mathbf{x}_k)). \tag{20}$$

where $m$ is the mean of the training samples, and the detailed expression of the function $g(\cdot)$ is shown in Appendix.

We summarize the algorithm of Kernel ERLDA in Fig.(3).

## VII. Experimental results

### A. Results on gray scale datasets

In this section, we evaluate the proposed methods on four well-known face datasets: ORL [18], Yale [24],UMIST [10], and FERET [16]. The ORL database consists of samples from 40 individuals, each of which owns 10 different images. These images were taken with a tolerance for tilting and rotation up to 20°. The facial expressions are various: open or closed eyes, smiling or non-smiling and even occlusion of glasses. All images are grayscale and normalized to a resolution of $112 \times 92$ pixels. This dataset is mainly designed to test the performance under scale and rotation variations. The Yale database contains images from 15 subjects with 11 different samples for each individual. The images differ from lighting conditions (left-light, center-light, right-light), facial expressions (happy, sad, surprised, and so on), and occlusion (with/without glasses). We use the cropped images of size $32 \times 32$. This database is used to test the performance under expression and illumination variations. The UMIST database is a multi-view dataset, consisting of 575 images from 20 persons, each of which has a wide range of poses from profile to frontal views. This database mainly focuses on pose variations. The FERET database consists of a gallery which includes 1196 persons with only a single image for each person, and four probe sets: *fafb*,*fafc*,*dup1*, and *dup2*. The *fafb* set has expression variation, *fafc* set contains lighting variation while *dup1* and *dup2* sets were acquired a few days later. Samples from these datasets are shown in Fig.(4).

We set the parameter $d = 50$, the hyper-parameter $\sigma = 0.5$ and the regularization parameter $\beta = 1.0$ for all datasets. And we apply random orthoprojectors in our methods. Detailed comparison between different projectors are presented in section (VII-C). We run our methods for 10 independent trials, and average results are reported. We use recognition accuracy to measure the performance of the proposed methods.

```
INPUT The training set $\{\mathbf{x}_1,\dots,\mathbf{x}_C\}$ and a test instance $\mathbf{x}_t$.
1. Sub-sample the training set and the test instance on the index set $S_j$ and obtain a new training set $\{x_1^{(S_j)},\dots,x_C^{(S_j)}\}$ and test instance $\mathbf{x}_t^{(S_j)}$.
2. Randomly choose an orthogonal projection $\Phi_j$, and obtain the new training set $\{\Phi_j x_1^{(S_j)},\dots,\Phi_j x_C^{(S_j)}\}$ and test instance $\Phi_j\mathbf{x}_t^{(S_j)}$
3. Calculate $P(\mathbf{x}_k|\mathbf{x}_t,h_j)$ , $k=1,\dots,C$ as follows:
    3.a Calculate the kernel vectors $Ker(\mathbf{x}_t,\cdot)$,$Ker(\mathbf{x}_k,\cdot)$,$Ker(m,\cdot)$ by (18) and (19).
    3.b Calculate $d_{Ker}(\mathbf{x}_t,\mathbf{x}_k)$,$d_{Ker}(\mathbf{x}_t,m)$,$d_{Ker}(\mathbf{x}_k,\mathbf{x}_k)$,$d_{Ker}(m,\mathbf{x}_k)$ by (17).
    3.c Calculate the discriminative function $s_k$ by (20).
    3.d Obtain the posterior probability $P(\mathbf{x}_k|\mathbf{x}_t,h_j)$ by (8).
4. Calculate the weight of the $j$-th classifier by solving the optimization problem (10).
5. Calculate the overall posterior probability $P(\mathbf{x}_k|\mathbf{x}_t)=\frac{\sum_j v_j P(\mathbf{x}_k|\mathbf{x}_t,h_j)}{\sum_j v_j}$.
OUTPUT The label of test instance $\mathbf{x}_t$ : $L(\mathbf{x}_t)=\arg\max_k P(\mathbf{x}_k|\mathbf{x}_t)$.
```

Fig. 3. Algorithm of Kernel ERLDA.

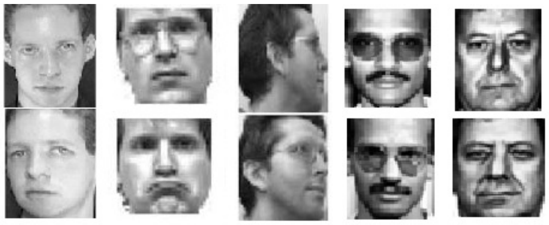


Fig. 4. Some sample images from the four datasets. The first three columns are for ORL, Yale, UMIST, respectively. The last two columns come from FERET.

We compare our methods with several existing state-of-the-arts in SSPP problem: Eigenface [20], $(PC)^2A$ [5], SVD+PCA [27], Block FLD [4], SVD-based FLD [9] and Adaptive FLD [14]. We implement the above methods except Adaptive FLD [14] and SVD-based FLD. Adaptive FLD needs extra database to train the model, so we just compare it with its own reported result. For SVD-based FLD [9], we use the public software[1]. As for $(PC)^2A$, we tune the only parameter $\alpha$ between 0.1 and 0.5 suggested by [5], and report the best result. As to the Block FLD [4], the sensitive parameter is the size of the block. We tried four different sizes ($10\times10$,$10\times25$,$20\times10$ and $20\times25$) and report the best result.

We show the comparison results of these methods on the four databases in Tab.(I). We can see that our methods outperform the existing works on all the four benchmarks. Especially in Yale dataset, due to large variations between images of the same person, one-sample training is hard to discriminate each subject, while our ERLDA makes use of different projections to try to capture the possible variations on different subspaces. Furthermore, it is noticeable that Kernel ERLDA can improve upon ERLDA by $2\% \backsim 6\%$. This is because the "*kernel trick*" makes the similarity between faces more reliable.

### *B. Results on face images in wild*

The face imegse of the four data sets tested in the previous section are obtain in experimental environments. In this section, an experiment is carried on the "Labeled Faces in the Wild" (LFW) dataset [12], which contains $13,233$ images of $5,749$ persons collected from web. The face images of LFW are collected in wild conditions, so it's much more challenging. Faces were detected by the Viola-Johns face detector [21]. In our experiments, a subset of LFW is used to facilitate the test in the framework of SSPP. We chose 24 persons who have more than 30 but less than 60 pictures. Examples of the chosen persons are shown in Fig.(5). We use the V1-like feature [17] to represent each person. On this dataset, we use the same parameter settings: $d=50$, $\sigma=0.5$ and $\beta=1.0$. Since the number of instances for each person varies, we use ROC curve to evaluate the performances of different algorithms. We randomly choose one image for each person as the training sample and the other images as testing samples. We repeat doing this experiments 50 times, then report the average ROC curve of the 50 independent runs. Results are shown in Fig.(6).

Fig. 5. Examples of the chosen 24 persons. Each of them has more than 30 but less than 60 different images.

It is noticeable from Fig.(6) that the proposed methods outperform to other similar algorithms on the task of SSPP. In addition, Kernel ERLDA is better than ERLDA due to the capability of capturing nonlinear relationship among the sub-sampled data.

[1] http://www4.comp.polyu.edu.hk/ cslzhang/code.htm

TABLE I

A COMPARISON OF ACCURACY ON FOUR DATABASES.

| Methods/Databases | ORL | Yale | UMIST | *fafb* | *fafc* | *dup1* | *dup2* |
|---|---|---|---|---|---|---|---|
| Eigenface | 44.5 | 15.7 | 17.9 | 87.4 | 10.3 | 38.9 | 12.8 |
| $(PC)^2A$ | 54.1 | 18.7 | 18.6 | 87.9 | 12.4 | 38.6 | 13.2 |
| SVD+PCA | 64.1 | 23.3 | 52.5 | 88.2 | 35.4 | 39.3 | 16.4 |
| Block FLD | 70.8 | 32.0 | 57.2 | 89.5 | 50.0 | 41.3 | 33.8 |
| SVD-based FLD | 75.6 | 34.7 | 61.4 | 90.5 | 61.3 | 45.7 | 36.8 |
| Adaptive FLD | - | - | - | 88.5 | 71.6 | 53.3 | 35.0 |
| ERLDA | **79.4** | **45.9** | **63.7** | **92.4** | **70.8** | **52.4** | **38.0** |
| Kernel ERLDA | **82.3** | **52.0** | **65.6** | **93.6** | **72.9** | **58.0** | **40.1** |

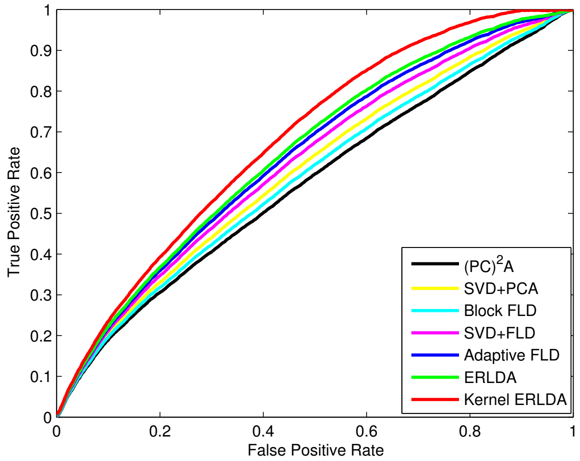
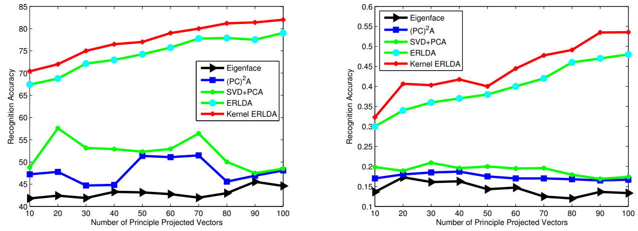


Fig. 6. Examples of the chosen 24 persons. Each of them has more than 30 but less than 60 different images.



(A) (B)

Fig. 7. A quantitative comparison with state-of-the-art methods on different number of principle projection vectors. (A) and (B) show the results on ORL and Yale databases, respectively.

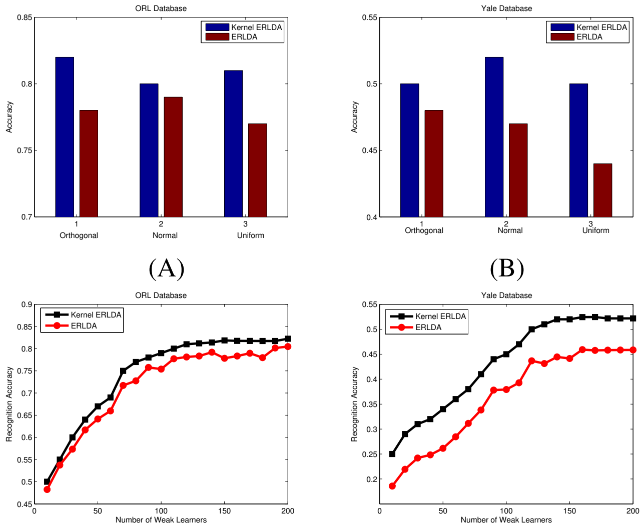
## C. Sensitivity Analysis

In this section, we discuss the sensitivity of some parameters in the proposed methods: the number of principle projected vectors, the effect of different random projections(orthogonal, normal gaussian, uniform), and the number of weak learners. Only two datasets (ORL and Yale databases) are used as illustrative examples.

First, we choose different number of principle projection vectors, and compare our methods with some other principle vector-based methods, such as Eigenface [20], $(PC)^2A$ [5], SVD+PCA [27]. Note that Block FLD [4] and SVD-based FLD [9] only perform well when a few principle projection vectors (less than 10) are used. The comparison results on ORL and Yale databases are shown in Fig.(7). We can see that, our methods outperform the others with the same principle vectors. Also, it is observed that when the number of principle vectors increases, the performances of ERLDA and Kernel ERLDA are better. This is easy to understand, since our one-sample LDA is built on sub-sampled data, and more choices of principle vectors make the proposed algorithm more capable to capture the structure between different classes.

Secondly, we test the effect of different choice of random projections. As discussed before, three kinds of projectors (orthogonal, normal, uniform) are adopted for random projections. We tried all these three projectors, and set the other parameters fixed. The results are reported in Fig.(8)(A,B). We can see that, three projectors are achieving almost the same performance. Moreover, orthogonal projection seems a slightly better than the other two.

Finally, we test the proposed methods with different numbers of ensembles on ORL and Yale Databases. Fig.(8)(C,D) illustrates the performances, from which, we can conclude that, as the ensemble number increases, the proposed methods can achieve better performance. However, there is a trade-off between accuracy and time cost.



(A) (B)

(C) (D)

Fig. 8. (A) and (B) show the comparison results of different types of random projection on ORL and Yale datasets. (C) and (D) are the comparison results of different number of weak learners on ORL and Yale datasets.

## VIII. Conclusion

We have presented a novel approach for face recognition with single sample per person. Our method takes advantage of one-sample LDA and random projections. The main principle is that ensemble of weak learners from each randomized small sampled face is capable to capture the signature of face discrimination and therefore improve the performance of face recognition. Theoretical analysis is given and a kernel extension is presented. Significant improvement over state-of-the-arts has been observed on various benchmark datasets.

## References

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

[2] S. Bernhard, S. Alexander, and M. Klaus-Robert. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.*, 10:1299–1319, 1998.

[3] A. Blum. Random projection, margins, kernels, and feature-selection. *LNCS*, 3940:52–68, 2005.

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

[5] S. Chen, D. Zhang, and Z. Zhou. Enhanced (pc)2a for face recognition with one training image per person. *Pattern Recognition Letters*, 25:1173–1181, 2004.

[6] W. Chen and Y. Gao. Recognizing partially occluded faces from a single sample per class using string-based matching. In *Proceedings of ECCV*, pages 496–509, 2010.

[7] I. CVX Research. CVX: Matlab software for disciplined convex programming, version 2.0 beta.

[8] R. Ebrahimpour, M. Nazari, M. Azizi, and A. Amiri. Single training sample face recognition using fusion of classifiers. *International Journal of Hybrid Information Technology*, 4:25 – 32, Jan. 2011.

[9] Q. Gao, L. Zhang, and D. Zhang. Face recognition using flda with single training image per person. *Applied Mathematics and Computation*, pages 726–734, 2008.

[10] D. Graham. Umist face database, 2005.

[11] C. Hegde, M. B. Wakin, and R. G. Baraniuk. Random projections for manifold learning. In *NIPS*, Dec. 2007.

[12] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[13] J. Huang, P. C. Yuen, W.-S. Chen, and J. H. Lai. Component-based lda method for face recognition with one training sample. In *Proceedings of the IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, AMFG '03, 2003.

[14] M. Kan, S. Shan, Y. Su, X. Chen, and W. Gao. Adaptive discriminant analysis for face recognition from single sample per person. In *Automatic Face, Gesture Recognition and Workshops*, pages 193 – 199, 2011.

[15] L.Zhang and D.Samaras. Face recognition from a single training image under arbitrary unknown lighting using spherical harmonics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28:1746–1762, 2006.

[16] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss. The feret evaluation methodology for face recognition algorithms. *TPAMI*, 22(3):1090–1103, 2000.

[17] N. Pinto, J. J. DiCarlo, and D. D. Cox. How far can you get with a modern face recognition test set using only simple features? In *IEEE Computer Vision and Pattern Recognition*, 2009.

[18] F. S. Samaria and A. Harter. Parameterisation of a stochastic model for human face identification, 1994.

[19] X. Tan, S. Chen, Z. Zhou, and F. Zhang. Face recognition from a single image per person: A survey. *Pattern Recognition*, 39:1725–1745, 2006.

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

[21] P. Viola and M. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57:137–154, 2004.

[22] J. Wang, K.N.Plataniotis, J. Lu, and A.N.Venetsanopoulos. On solving the face recognition problem with one training sample per subject. *Pattern Recognition*, pages 1746–1762, 2006.

[23] J. Wu and Z. Zhou. Face recognition with one training image per person. *Pattern Recognition Letters*, 23:1711–1719, 2002.

[24] U. Yale. Yale face database, 2002.

[25] J. Yang, D. Zhang, A. F. Frangi, and J. Yang. Two-dimensional pca: A new approach to appearance-based face representation and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26:131–137, 2004.

[26] X. Y.Su, S.Shan and W.Gao. Adaptive generic learning for face recognition from a single sample per person. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[27] D. Zhang, S. Chen, and Z. Zhou. A new face recognition method based on svd perturbation for single example image per person. *Applied Mathematics and Computation*, pages 895–907, 2005.

[28] W. Zhao, R. Chellappa, P. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Computing Surveys*, 35, 2003.

## IX. Appendix

**Derivation of** $g(\cdot)$.
Following Eqn.(16), we have

$$\begin{aligned} g(\cdot) &\propto \mathbf{x}_t^T B^{-1}\mathbf{x}_k + \\ &\quad \mu\mathbf{x}_t^T B^{-1} U_k (1-\mu U_k^T B^{-1} U_k)^{-1} U_k^T B^{-1}(\mathbf{x}_k - m) \\ &= \mathbf{x}_t^T B^{-1}\mathbf{x}_k + \\ &\quad \frac{\mu\mathbf{x}_t^T B^{-1}(m+\mathbf{x}_k)(m+\mathbf{x}_k)^T B^{-1}(\mathbf{x}_k - m)}{1-\mu(m+\mathbf{x}_k)B^{-1}(m+\mathbf{x}_k)^T} \\ &= (d_{\mathcal{B}}(\mathbf{x}_t,\mathbf{x}_k) - \mu(2d_{\mathcal{B}}(\mathbf{x}_t,\mathbf{x}_k)d_{\mathcal{B}}(\mathbf{x}_k,m) + \\ &\quad 2d_{\mathcal{B}}(\mathbf{x}_t,\mathbf{x}_k)d_{\mathcal{B}}(m,m) + d_{\mathcal{B}}(\mathbf{x}_t,m)d_{\mathcal{B}}(m,m) - \\ &\quad d_{\mathcal{B}}(\mathbf{x}_t,m)d_{\mathcal{B}}(\mathbf{x}_k,\mathbf{x}_k)))/ \\ &\quad (1-\mu(d_{\mathcal{B}}(\mathbf{x}_k,\mathbf{x}_k) + 2d_{\mathcal{B}}(\mathbf{x}_k,m) + d_{\mathcal{B}}(m,m)) \end{aligned} \tag{21}$$

where $d_{\mathcal{B}}(x,y) = x^T B^{-1} y = y^T B^{-1} x$, since $B$ is symmetric. Since $d_{\mathcal{B}}(m,m)$ is a constant, we can write the function $g(\cdot)$ as $g(d_{\mathcal{B}}(\mathbf{x}_t,\mathbf{x}_k), d_{\mathcal{B}}(\mathbf{x}_t,m), d_{\mathcal{B}}(\mathbf{x}_k,\mathbf{x}_k), d_{\mathcal{B}}(m,\mathbf{x}_k))$.