Linear Laplacian Discrimination for Feature Extraction

Deli Zhao Zhouchen Lin Rong Xiao Xiaoou Tang

Microsoft Research Asia, Beijing, China

delizhao@hotmail.com, {zhoulin,rxiao,xitang}@microsoft.com

Abstract

Discriminant feature extraction plays a fundamental role in pattern recognition. In this paper, we propose the Linear Laplacian Discrimination (LLD) algorithm for discriminant feature extraction. LLD is an extension of Linear Discriminant Analysis (LDA). Our motivation is to address the issue that LDA cannot work well in cases where sample spaces are non-Euclidean. Specifically, we define the within-class scatter and the between-class scatter using similarities which are based on pairwise distances in sample spaces. Thus the structural information of classes is contained in the within-class and the between-class Laplacian matrices which are free from metrics of sample spaces. The optimal discriminant subspace can be derived by controlling the structural evolution of Laplacian matrices. Experiments are performed on the facial database for FRGC version 2. Experimental results show that LLD is effective in extracting discriminant features.

1. Introduction

Discriminant feature extraction plays the central role in recognition and classification. Principal component analysis (PCA) is a classic linear method for unsupervised feature extraction. PCA [10] learns a kind of subspaces where the maximum covariance of all training samples are preserved. The Eigenfaces [19] method for face recognition applies PCA to learn an optimal linear subspace of facial structures. PCA also plays a fundamental role in face sketch recognition[18, 17]. Locality Preserving Projections (LPP) [7] is another typical approach for un-supervised feature extraction. LPP is the linearization of Laplacian Eigenmaps [4] which can find underlying clusters of samples. LPP shows the superiority in terms of image indexing and face recognition. The Laplacianfaces face recognition method [8] is based on the combination of PCA and LPP, in the sense that LPP is performed in the PCA-transformed feature space.

However, un-supervised learning cannot properly model underlying structures and characteristics of different classes. Discriminant features are often obtained by class-supervised learning. Linear discriminant analysis (LDA) is the traditional approach to learning discriminant subspaces where the between-class scatter of samples is maximized and the within-class scatter is minimized at the same time. The Fisherfaces algorithm [3] and many variants of LDA hae shown good performance in face recognition in complex scenarios. [24, 9, 20, 22, 11, 12, 28]. By defining the representations of intra-personal and extra-personal differences, Bayesian face recognition [2] proposes another way to explore discriminant features via probabilistic similarity measure. The inherent connection between LDA and Bayesian faces was unified by Wang and Tang [21] in a more general form.

LDA algorithm has the advantages of reasonable motivation in principle and the simplicity in form. The conventional LDA algorithm is formulated by the ratio of between-class scatter and the within-class scatter which are represented by norms measured with Euclidean metrics. So there is an underlying assumption behind LDA that it works in Euclidean spaces. However there are many scenarios where sample spaces are non-Euclidean in computer vision. For instance, distances between feature vectors yielded by histograms cannot be measured by Euclidean norms. In this case, some non-Euclidean measures are usually applied, such as the Chi squares statistic, the log-likelihood statistic, and the histogram intersection. The primary formulation of LDA does not hold in non-Euclidean spaces. As a consequence, LDA fails to find the optimal discriminant subspace.

We propose an improved method, named Linear Laplacian Discrimination (LLD), for discriminant feature extraction in this paper. We formulate the within-class scatter and the between-class scatter by means of similarity-weighted criterions. These criterions benefit from the advantages of Laplacian Eigenmaps and LPP. Similarities here are computed from the exponential function of pairwise distances in the original sample spaces, which is free from various forms of metrics. So, LLD can be applied to any linear space for

1-4244-1180-7/07/$25.00 ©2007 IEEE

classification. The structural information of classes is governed by the within-class Laplacian matrix and the between-class Laplcian matrix. These two matrices evolve with the time which is a free parameter in similarity measure. From this viewpoint, LDA is exactly a special case when the time approaches the positive infinity. Therefore, LLD not only overcomes the problems of non-Euclidean metrics but also presents an alternative way to find better discriminant subspaces.

Experiments are performed for face identification on a subset of facial database for FRGC version 2. We compare our LLD method with PCA, LPP, LBP, and the traditional LDA. Discriminant features are extracted on PCA and LBP expressive features [16], implying that LLD, LPP, and LDA are performed in the PCA and LBP transformed spaces, respectively. The PCA expressive features can be viewed Euclidean whereas the LBP expressive features are non-Euclidean. Experimental results show that LLD outperforms existing methods in terms of discrimination power.

2. Linear Lapacian Discrimination

2.1. Motivations

Since LDA encounters the problem of the metric measure, we consider similarity as the inherent characteristic of pairwise points instead of the distance in our work. Our work is inspired by the application of Laplacian Eigenmaps [4] in manifold learning and its linearization LPP [7] in clustering and recognition. The geometric distances between mapped points that lie on an underlying manifold can be controlled by similarities between corresponding points in the original space. Underlying clusters will appear automatically after non-linear maps. By extensive experiments, He *et al*. [8] concluded that the linearization of such criterions yields a good performance in image indexing, clustering, and face recognition. From above considerations, we propose the LLD algorithm.

2.2. Discriminant scatters

Specifically, let $\mathbf{x}_i^s$ denote the i-th sample in the s-th class, where $\mathbf{x}_i^s \in \mathcal{M}^D$ and $\mathcal{M}^D$ is the D-dimensional sample space. We obtain the associated discriminant feature $\mathbf{y}_i^s$ of $\mathbf{x}_i^s$ by projection

$$\mathbf{y}_i^s = \mathbf{U}^T \mathbf{x}_i^s, \tag{1}$$

where the d columns of the projection matrix $\mathbf{U}$ are the orthogonal bases of discriminant subspace. Let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]$ denote all original samples, where n is the number of all samples. Then we have $\mathbf{Y} = \mathbf{U}^T\mathbf{X}$, where $\mathbf{Y} = [\mathbf{y}^1, \ldots, \mathbf{y}^n]$. Given two points $\mathbf{x}_i^s$ and $\mathbf{x}_i^t$, the Euclidean distance between them is defined as

$$\|\mathbf{x}_i^s - \mathbf{x}_i^t\|^2_{\mathcal{R}^D} = \sum_{k=1}^{D} (x_{ik}^s - x_{ik}^t)^2, \tag{2}$$

where x_{ik}^s is the k-th component of $\mathbf{x}_{ik}^s$.

Let α_s denote the within-class scatter of class s. Define it as

$$\alpha_s = \sum_{i=1}^{c_s} w_i^s \|\mathbf{y}_i^s - \bar{\mathbf{y}}^s\|^2_{\mathcal{R}^D}, s = 1, \ldots, c, \tag{3}$$

where w_i^s is the weight, defined by

$$w_i^s = \exp\left(-\frac{\|\mathbf{x}_i^s - \bar{\mathbf{x}}^s\|^2_{\mathcal{M}^D}}{t}\right), i = 1, \ldots, c_s. \tag{4}$$

Here t is the time variable, and $\exp(\bullet)$ denotes the exponential function. It suffices to note that the distance between $\mathbf{y}_i^s$ and $\bar{\mathbf{y}}^s$ are measured by the Euclidean norm $\| \bullet \|_{\mathcal{R}^D}$, and the distance between $\mathbf{x}_i^s$ and $\bar{\mathbf{x}}^s$ are measured by the norm $\| \bullet \|_{\mathcal{M}^D}$ which depends on the metric of the original sample space. The space may be Euclidean or non-Euclidean. To obtain the compact expression of (3), let $\mathbf{W}_s = diag(w_1^s, w_2^s, \ldots, w_{c_s}^s)$ be a diagonal matrix and $\mathbf{Y}_s = [\mathbf{y}_1^s, \mathbf{y}_2^s, \ldots, \mathbf{y}_{c_s}^s]$. Besides, let $\mathbf{e}_{c_s}$ denote the all-one column vector of length c_s. Then $\bar{\mathbf{y}}^s = \frac{1}{c_s}\mathbf{Y}_s\mathbf{e}_{c_s}$. Rewriting (3) shows

$$\begin{aligned}
\alpha_s &= \sum_{i=1}^{c_s} w_i^s tr\left\{(\mathbf{y}_i^s - \bar{\mathbf{y}}^s)(\mathbf{y}_i^s - \bar{\mathbf{y}}^s)^T\right\} && (5)\\
&= tr\left\{\sum_{i=1}^{c_s} w_i^s \mathbf{y}_i^s (\mathbf{y}_i^s)^T\right\} && (6)\\
&\quad - 2tr\left\{\sum_{i=1}^{c_s} w_i^s \mathbf{y}_i^s \left(\frac{1}{c_s}\mathbf{Y}_s\mathbf{e}_{c_s}\right)^T\right\} && (7)\\
&\quad + tr\left\{\sum_{i=1}^{c_s} w_i^s \left(\frac{1}{c_s}\mathbf{Y}_s\mathbf{e}_{c_s}\right)\left(\frac{1}{c_s}\mathbf{Y}_s\mathbf{e}_{c_s}\right)^T\right\} && (8)\\
&= tr(\mathbf{Y}_s\mathbf{W}_s\mathbf{Y}_s^T) - \frac{2}{c_s} tr\left\{\mathbf{Y}_s\mathbf{W}_s\mathbf{e}_{c_s}(\mathbf{e}_{c_s})^T\mathbf{Y}_s^T\right\} && (9)\\
&\quad + \frac{\mathbf{e}_{c_s}\mathbf{W}_s(\mathbf{e}_{c_s})^T}{c_s^2} tr\left\{\mathbf{Y}_s\mathbf{e}_{c_s}(\mathbf{e}_{c_s})^T\mathbf{Y}_s^T\right\}. && (10)
\end{aligned}$$

Thus we obtain

$$\alpha_s = tr(\mathbf{Y}_s\mathbf{L}_s\mathbf{Y}_s^T), \tag{11}$$

where

$$\mathbf{L}_s = \mathbf{W}_s - \frac{2}{c_s}\mathbf{W}_s\mathbf{e}_{c_s}(\mathbf{e}_{c_s})^T + \frac{\mathbf{e}_{c_s}\mathbf{W}_s(\mathbf{e}_{c_s})^T}{c_s^2}\mathbf{e}_{c_s}(\mathbf{e}_{c_s})^T. \tag{12}$$

Let α denote the total within-class scatter of all samples. We have

$$\alpha = \sum_{s=1}^{c} \alpha_s = \sum_{s=1}^{c} tr(\mathbf{Y}_s \mathbf{L}_s \mathbf{Y}_s^T). \tag{13}$$

There is a 0-1 indicator matrix $\mathbf{S}_s$ satisfying $\mathbf{Y}_s = \mathbf{Y}\mathbf{S}_s$. Each column of $\mathbf{S}_s$ records the class information which is known for supervised learning. Substituting the expression of $\mathbf{Y}_s$ to (13) gives

$$\alpha = \sum_{s=1}^{c} tr(\mathbf{Y}\mathbf{S}_s \mathbf{L}_s \mathbf{S}_s^T \mathbf{Y}^T) = tr(\mathbf{Y}\mathbf{L}_{\mathfrak{w}} \mathbf{Y}^T), \tag{14}$$

where

$$\mathbf{L}_{\mathfrak{w}} = \sum_{s=1}^{c} \mathbf{S}_s \mathbf{L}_s \mathbf{S}_s^T, \tag{15}$$

is the *within-class Laplacian matrix*. If the matrix $\mathbf{X}$ is ordered such that samples appear by class $\mathbf{X} = [\mathbf{x}_1^1, \ldots, \mathbf{x}_{c_1}^1, \ldots, \mathbf{x}_1^c, \ldots, \mathbf{x}_{c_c}^c]$, then the within-class Laplacian matrix $\mathbf{L}_{\mathfrak{w}}$ reads the diagonal block form of $\mathbf{L}_{\mathfrak{w}} = diag(\mathbf{L}_1, \mathbf{L}_2, \ldots, \mathbf{L}_c)$. Such alignment technique is applicable for problems that can be formulated as the similar expression (11) [25, 26].

Plugging the expression of $\mathbf{Y}$ into (14), we arrive at the final form of the total within-class scatter, showing

$$\alpha = tr(\mathbf{U}^T \mathbf{D}_{\mathfrak{w}} \mathbf{U}), \tag{16}$$

where $\mathbf{D}_{\mathfrak{w}} = \mathbf{X}\mathbf{L}_{\mathfrak{w}}\mathbf{X}^T$ is the *within-class scatter matrix*.

Next, the between-class scatter β of all classes is defined as

$$\beta = \sum_{s=1}^{c} w^s \|\bar{\mathbf{y}}^s - \bar{\mathbf{y}}\|_{\mathcal{R}^D}^2, s = 1, \ldots, c, \tag{17}$$

where w^s is defined by

$$w^s = \exp\left(-\frac{\|\bar{\mathbf{x}}^s - \bar{\mathbf{x}}\|_{\mathcal{M}^D}^2}{t}\right). \tag{18}$$

Let $\bar{\mathbf{Y}} = [\bar{\mathbf{y}}^1, \ldots, \bar{\mathbf{y}}^c]$ denote the matrix consisting of all center vectors of classes and $\mathbf{W}_{\mathfrak{b}} = diag(w^1, w^2, \ldots, w^c)$. Following similar formulations from (5) to (12), we can rewrite (17) as

$$\beta = tr(\bar{\mathbf{Y}}\mathbf{L}_{\mathfrak{b}}\bar{\mathbf{Y}}^T), \tag{19}$$

where

$$\mathbf{L}_{\mathfrak{b}} = \mathbf{W}_{\mathfrak{b}} - \frac{2}{c}\mathbf{W}_{\mathfrak{b}}\mathbf{e}_c(\mathbf{e}_c)^T + \frac{\mathbf{e}_c \mathbf{W}_{\mathfrak{b}} (\mathbf{e}_c)^T}{c^2}\mathbf{e}_c(\mathbf{e}_c)^T \tag{20}$$

is the *between-class Laplacian matrix*. Let $\bar{\mathbf{X}} = [\bar{\mathbf{x}}^1, \ldots, \bar{\mathbf{x}}^c]$. We have $\bar{\mathbf{Y}} = \mathbf{U}^T\bar{\mathbf{X}}$. Rewriting (19) yields

$$\beta = tr(\mathbf{U}^T \mathbf{D}_{\mathfrak{b}} \mathbf{U}), \tag{21}$$

where $\mathbf{D}_{\mathfrak{b}} = \bar{\mathbf{X}}\mathbf{L}_{\mathfrak{b}}\bar{\mathbf{X}}^T$ is called the *between-class scatter matrix*.

2.3. Finding the optimal projection

Like LDA, to make projected samples favor of classification in feature space, we expect that samples within the same classes cluster as close as possible and samples between classes seperate as far as possible. Let us examine the formulations of the within-class scatter (3) and the between-class scatter (17), respectively. One can see that the smaller the distance between $\mathbf{x}_i^s$ and $\bar{\mathbf{x}}^s$ is, the larger the similarity w_i^s is. If the within-class scatter α_s keeps constant, we know from (3) that $\|\mathbf{y}_i^s - \bar{\mathbf{y}}^s\|_{\mathcal{R}^D}$ will be small if the weight w_i^s is large, implying that $\mathbf{y}_i^s$ will be close to its center $\bar{\mathbf{y}}^s$. So, $\mathbf{y}_i^s$ will approach its center $\bar{\mathbf{y}}^s$ as α_s approaches the minimum. Therefore, our expectation on within-class samples will be fulfilled if the total within-class scatter α is minimized. By the similar analysis, our expectation on between-class samples being far apart will be realized if the between-class scatter β is maximized. To summary, we get the following dual-objective optimization model [27]

$$\begin{cases} \arg\min\limits_{\mathbf{U}} tr(\mathbf{U}^T \mathbf{D}_{\mathfrak{w}} \mathbf{U}) \\ \arg\max\limits_{\mathbf{U}} tr(\mathbf{U}^T \mathbf{D}_{\mathfrak{b}} \mathbf{U}) \\ \mathbf{U}^T\mathbf{U} = \mathbf{I}_d. \end{cases} \tag{22}$$

To simplify the optimization, we construct the following Fisher criterion

$$\mathfrak{J}(\mathbf{U}) = \frac{\beta}{\alpha} = \frac{tr(\mathbf{U}^T \mathbf{D}_{\mathfrak{b}} \mathbf{U})}{tr(\mathbf{U}^T \mathbf{D}_{\mathfrak{w}} \mathbf{U})}. \tag{23}$$

Then the optimization reduces to the similar fashion of the conventional LDA

$$\begin{cases} \arg\max\limits_{\mathbf{U}} \mathfrak{J}(\mathbf{U}) \\ \mathbf{U}^T\mathbf{U} = \mathbf{I}_d. \end{cases} \tag{24}$$

To solve for $\mathbf{U}$, the above optimization can be done on Grassmann manifolds [6] where $\mathbf{U}$ is viewed as a point on geodesic flows. However, the formal discussion pertaining to this topic is beyond the scope of this paper. Here, we take the similar approach used in the traditional LDA to solve the above optimization problem. We take the d eigenvectors from the following generalized eigen-analysis

$$\mathbf{D}_{\mathfrak{b}} \mathbf{u}_i = \lambda_i \mathbf{D}_{\mathfrak{w}} \mathbf{u}_i \tag{25}$$

that are associated with the d largest eigen-values $\lambda_i, i = 1, \ldots, d$.

2.4. Computational considerations

Like LDA, LLD encounters the computational trouble as well when $\mathbf{D}_{\mathfrak{w}}$ is singular. $\mathbf{D}_{\mathfrak{w}}$ is not invertible when $\mathbf{L}_{\mathfrak{w}}$ is not of full rank. Such case frequently occurs in computer vision since images have large dimensions whereas the number of classes is usually small. However, the generalized

eigen-analysis (25) needs a positive definite $\mathbf{D}_{\mathfrak{w}}$. Several strategies exist to address the issue. Here we propose two approaches.

2.4.1 Approach I: PCA subspace

When the original sample space is Euclidean, discriminant features can be extracted from expressive features yielded by PCA. Namely LLD can be performed in the PCA-transformed space. Fisherfaces [3] successfully employed this strategy for recognition. Specifically, let $\mathbf{U}_{PCA}$ denote the matrix whose columns are a set of orthogonal base of the principal subspace. We first project $\mathbf{D}_{\mathfrak{w}}$ and $\mathbf{D}_{\mathfrak{b}}$ into the PCA-transformed space to give $\tilde{\mathbf{D}}_{\mathfrak{w}} = (\mathbf{U}_{PCA})^T \mathbf{D}_{\mathfrak{w}} \mathbf{U}_{PCA}$ and $\tilde{\mathbf{D}}_{\mathfrak{b}} = (\mathbf{U}_{PCA})^T \mathbf{D}_{\mathfrak{b}} \mathbf{U}_{PCA}$. Then perform the generalized eigen-analysis (25) using $\tilde{\mathbf{D}}_{\mathfrak{w}}$ and $\tilde{\mathbf{D}}_{\mathfrak{b}}$ instead. Let $\mathbf{U}_{LLD}$ denote the discriminant subspace. Then the final transformation is given by $\mathbf{U}_{PCA}\mathbf{U}_{LLD}$.

2.4.2 Approach II: dual subspaces

The ideas of dual spaces were proposed in [20]. Specifically, let the eigen-decomposition of $\mathbf{D}_{\mathfrak{w}}$ be $\mathbf{D}_{\mathfrak{w}} = \mathbf{V}\Lambda\mathbf{V}^T$, where $\mathbf{V}$ is the eigen-vector matrix and Λ is the diagonal eigenvalue matrix. Suppose $\mathbf{V}$ is split into $\mathbf{V} = [\mathbf{V}_1, \mathbf{V}_2]$, where $\mathbf{V}_1$ consists of eigenvectors corresponding to the r non-zeros eigenvalues and $\mathbf{V}_2$ consists of eigenvectors associated with the d zero eigenvalues, where r is the rank of $\mathbf{D}_{\mathfrak{w}}$. The method of Wang and Tang's dual-subspace [20] is to project $\mathbf{D}_{\mathfrak{b}}$ into $\mathbf{V}_1$ and $\mathbf{I} - \mathbf{V}_1\mathbf{V}_1^T$ respectively, then perform eigen-analysis on the projected between-class scatter matrices, which is equivalent to projecting the center of each class in this two spaces and performing PCA respectively. Next compute $\mathbf{D}_{\mathfrak{b}}^1 = \mathbf{V}_1^T \mathbf{D}_{\mathfrak{b}} \mathbf{V}_1$ and $\mathbf{D}_{\mathfrak{b}}^2 = \mathbf{V}_2^T \mathbf{D}_{\mathfrak{b}} \mathbf{V}_2$. Let $\mathbf{Q}_1$ and $\mathbf{Q}_2$ denote the principal eigenvector matrices of $\mathbf{D}_{\mathfrak{b}}^1$ and $\mathbf{D}_{\mathfrak{b}}^2$, respectively. Then we get two dual projection matrices $\mathbf{W}_1 = \mathbf{V}_1\mathbf{Q}_1$ and $\mathbf{W}_2 = \mathbf{V}_2\mathbf{Q}_2$.

Given two samples $\mathbf{x}_i$ and $\mathbf{x}_j$, the distance between their feature vectors $\mathbf{y}_i$ and $\mathbf{y}_j$ is determined by

$$d(\mathbf{y}_i, \mathbf{y}_j) = \|\mathbf{W}_1^T(\mathbf{x}_i - \mathbf{x}_j)\|_{\mathbb{R}^d} + \|\mathbf{W}_2^T(\mathbf{x}_i - \mathbf{x}_j)\|_{\mathbb{R}^d}. \quad (26)$$

Note that, for LDA, projecting samples only on the subspace spanned by $\mathbf{W}_1$ is essentially the approach of tackling the singular problem of the within-class scatter matrix by simultaneous diagonalization [16, 20]. In the following, the dual LLD and the dual LDA means that LLD and LDA are performed by dual subspaces.

2.5. Connections with existing methods

It is not hard to see that LLD is exactly LDA if t approaches the positive infinity in the similarity functions (4) and (18). So, the discriminant subspace of LDA is the stable state of the evolution of that of LLD with respect to the time t. Therefore, LLD is a more general version of LDA. LLD inherits the strengthes of LDA.

The LLD method also has connections with Chen *et al*.'s LDE [5] and Yan *et al*.'s MFA [23]. Overall, these methods can be viewed as specific forms of graph embedding formulated in [23]. However in principle, they are essentially different. LDE and MFA are more complicated. They take advantage of the partial structural information of classes and neighborhoods of samples at the same time while LDA and LLD purely explore the information of classes for discrimination.

3. Experiments

3.1. Dataset

We focus our attention on the problem of face identification. Given a novel face, the identification problem is that the system is asked to find the identity of the person in the gallery where the portrait of the person is presented. The motivation of this task comes from the current trends of performing face recognition or retrieval based on the facial images on the web or photos in digital family albums. In such cases, one is usually interested in finding the most similar faces of a given sample, which can be converted to be the face identification problem.

We perform the related experiments on a subset of facial data in experiment 4 of FRGC version 2 [14]. The query set for experiment 4 in this database consists of single uncontrolled still images which contains all the diverse factors of quality presented in the preceding subsection. There are 8014 images of 466 subjects in the set. However, there are only two facial images available for some persons. To guarantee the meaningful reference of our work for the tasks given above, we select a subset in the query set for our experiments.

To ensure the reproductivity of the tests, we clearly present the procedures. First, we search all images of each person in the set and take the first ten facial images if the number of facial images is not less than ten. Thus we get 3160 facial images of 316 subjects. Then we divide the 316 subjects into three subsets. First, the first 200 subjects are used as the gallery and probe set and the remaining 116 subjects are used as the training set. Second, we take the first five facial images of each person as the gallery set and the remaining five images as the probe set. Therefore, the set of persons for training is disjoint with that of persons for the gallery and the probe. Table 1 contains the information of facial data for experiments. We align the facial images according to the positions of eyes and mouth. Each facial image is cropped to a size of 64×72. Figure 1 shows ten images of one subject.

Set	Number of subjects	Number of images
Training	116	1160
Gallery	200	1000
Probe	200	1000

Table 1. Information of facial data for the experiments. These sets are selected from the query set for experiment 4 of FRGC version 2.

Figure 1. Facial images of one subject for the experiment in FRGC version 2. The facial images in the first row are in the gallery set and the second row is in the probe set.

3.2. Experimental results

We perform the disciminant feature extraction on the expressive features yielded by PCA and LBP [13, 1], respectively. This means that LLD, LPP, and LDA are performed in the PCA and LBP transformed spaces, respectively. PCA is the classic and well-recognized method for expressive feature extraction. LBP is a new approach which is proved effective for un-supervised feature extraction [13, 1]. The PCA feature space is Euclidean. The distances in this space are measured by the Euclidean norm (2). The LBP feature space is non-Euclidean. A distance measure in such a space is the Chi square, defined as

$$\chi^2(\mathbf{x}_i^s, \mathbf{x}_i^t) = \sum_{k=1}^{D} \frac{(x_{ik}^s - x_{ik}^t)^2}{x_{ik}^s + x_{ik}^t}. \qquad (27)$$

PCA and LBP are the baselines, respectively.

3.2.1 Based on PCA features

For the PCA-based two step strategy, the number of principal components is a free parameter. Wang and Tang [22, 21] showed that the dimension of principal subspaces significantly affects the performance of recognition for the PCA plus LDA strategy. Besides, they confirmed by experiments that the optimal number lies in the interval $[50, 200]$. Based on their work, we search the optimal number of principal components in this interval. We find that PCA performs best when the dimension of feature vectors is 190. So we take 190 as the number of principal components.

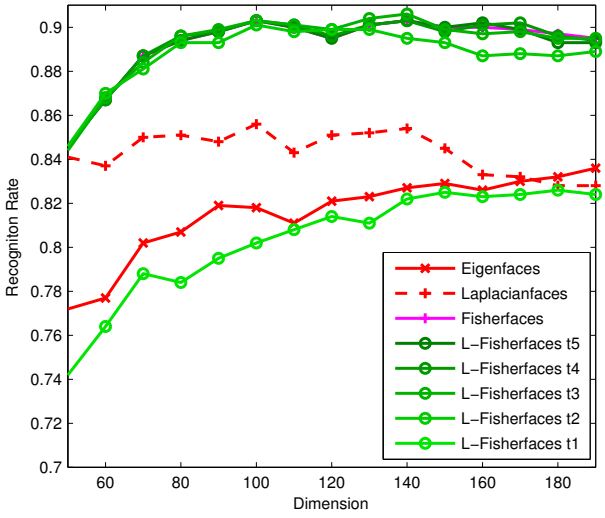

Figure 2. Recognition rates based on PCA features. Eigenfaces is the baseline. Eigenfaces (PCA), Laplacianfaces (PCA plus LPP), Fisherfaces (PCA plus LDA), and Lapalcian Fisherfaces (L-Fisherfaces in short, PCA plus LLD) are tested. We take 190 principal components in the PCA step due to that PCA performs best in this dimension. In the figure, $t_1 = 0.01$, $t_2 = 0.1$, $t_3 = 1$, $t_4 = 10$, and $t_5 = 100$. Laplacian Fisherfaces converge to Fisherfaces with a fast speed in the Euclidean feature space.

We name our method Laplacian Fisherfaces (L-Fisherfaces in short) due to that it is Laplacian-kernelized and formulated by Fisher criterion. As shown in Figure 2, L-Fisherfaces converge to Fisherfaces with a fast speed. The best performance of LLD is approximately achieved when LLD arrives at its stable state where each $\mathbf{W}_s$ is essentially the identity matrix when $t \geq 100$. This result means that the principal subspace of LLD yields the best discriminant performance in the Euclidean feature space when it approaches the stable state. Figure 3 shows various eigenfaces and their evolution across time.

3.2.2 Based on LBP features

We perform LBP on each facial image and then sub-divide each facial image by 7×7 grids. Histograms with 59 bins are performed on each sub-block. A LBP feature vector is obtained by concatenating the feature vectors on sub-blocks. Here we use 58 uniform patterns for LBP and each uniform pattern accounts for a bin. The remaining 198 binary patterns are put in another bin, resulting in the 59-bin histogram. So, the number of tuples in a LBP feature vector is $59 \times (7 \times 7) = 2891$. The $(8, 2)$ LBP is adopted. Namely, the number of circular neighbors for each pixel is 8 and the radius of the circle is 2. The above steps are consistent with that in [1].

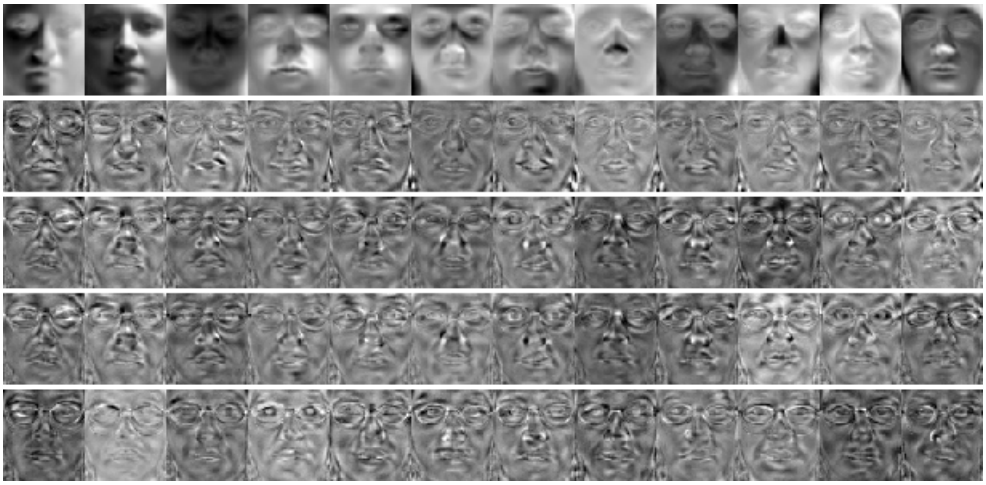

Figure 3. Eigenfaces of three different methods. The first row shows eigenfaces of PCA, the second row Laplacianfaces, the third row Fisherfaces (L-Fisherfaces with $t = 100$), the fourth row L-Fisherfaces with $t = 1$, and the fifth row L-Fisherfaces with $t = 0.01$.

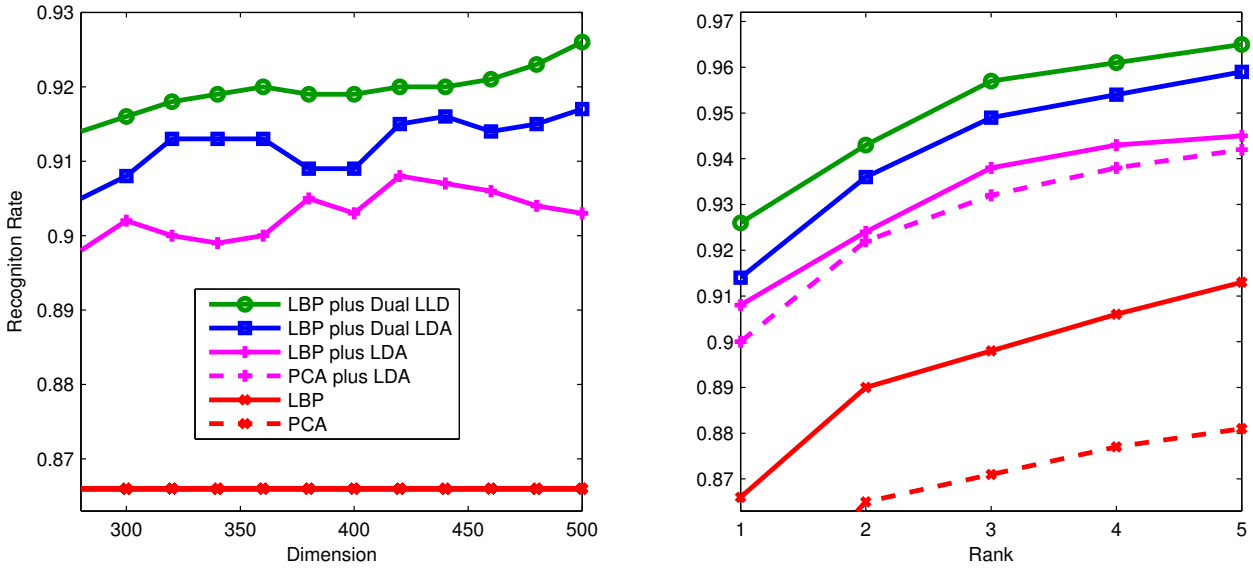

Figure 4. Recognition Rates based on LBP features. The performance of the LBP algorithm is the baseline. The left figure shows recognition rates with respect to dimension and the right one with respect to rank. The rank here is the top n criterion applied in the FERET evaluation methodology [15]. The legend also applies to the right figure. The time is $t = 500$ in Dual LLD.

As illustrated in Figure 4, Dual LLD ($t = 500$) consistently outperforms other methods with the recognition rate of 92.6%. The baseline (LBP) is 86.6%. The performance of LLD is equivalent to that of LDA for Euclidean features. However, LLD shows superiority to LDA for non-Euclidean features. The performance of LDA is limited when the feature space is non-Euclidean. LLD performs better in this case and is less limited by the change of attributions of feature spaces.

4. Conclusion

We develop a novel method (LLD) for pattern classification and discriminant feature extraction. Using similarity-weighted discriminant criterions, we define the within-class Laplacian matrix and the between-class Laplacian matrix. Thus LLD has the flexibility of finding optimal discriminant subspaces.

Experiments are performed on a subset in FRGC version 2. Experimental results show that LLD is equivlant

to the traditional LDA when the feature space is Euclidean and is superior to LDA when the feature space is non-Euclidean. In addition, LLD can significantly improve the discriminant performance of expressive features yielded by PCA and LBP. These results indicate that discriminant criterions formulated in LLD are more suitable for discriminant feature extraction. Whether the sample space is Euclidean or non-Euclidean, LLD is capable of capturing the discriminant characteristics of samples The performance of LLD will be further enhanced by trying other improved LDA methods.

Acknowledgement

We thank reviewers for their insightful comments and constructive suggestions. The first author would like to thank Xiaogang Wang, Dahua Lin, and Yuandong Tian for their generous help in experiments, and Xiaodi Hou for his suggestion on presentation.

References

[1] T. Ahonen, A. Hadid, and M. Pietikäinen. Face decription with local binary patterns: application to face recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(12):2037–2041, 2006.

[2] T. J. B. Moghaddam and A. Pentland. Bayesian face recognition. *Pattern Recognition*, 33:1771–1782, 2000.

[3] P. Belhumeur, J. Hespanha, and D. Kriegman. Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(1):711–720, 1997.

[4] M. Belkin and P. Niyogi. Laplacian Eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15:1373–1396, 2003.

[5] H. Chen, H. Chang, and T. Liu. Local discriminant embedding and its variants. *Proc. International Conf. on Computer Vision and Pattern Recognition*, 2, 2005.

[6] A. Edelman, T. Arias, and S. Smith. The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.*, 20(2):303–353, 1998.

[7] X. He and P. Niyogi. Locality preserving projections. *Proc. Neural Information Processing Systems Conf.*, 2003.

[8] X. He, S. Yan, P. N. Y.X. Hu, and H. Zhang. Face recognition using Laplacianfaces. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(3):328–340, 2005.

[9] P. Howland and H. Park. Generalizing discriminant analysis using the generalized singular value decomposition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 26(8):995–1006, 2004.

[10] T. Jollie. Principal component analysis. *Springer-Verlag, New York*.

[11] C. Liu. Capitalize on dimensionality increasing techniques for improving face recognition grand challenge performance. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(5):725–737, 2006.

[12] A. Martinez and M. Zhu. Where are linear feature extraction methods applicable? *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(12):1934–1944, 2006.

[13] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-sclae and rotation invariant texture classification with local binary patterns. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.

[14] P. Philips, P. Flynn, T. Scruggs, and K. Bowyer. Overview of the face recognition grand challenge. *Proc. International Conf. on Computer Vision and Pattern Recognition*, 2005.

[15] P. Phillips, H. Moon, S.A.Rizvi, and P. Rauss. The FERET evaluation methology for face-recogniton algorithms. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(10):1091–1103, 2000.

[16] D. Swets and J. Weng. Using discriminant eigenfeatures for image retrieval. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 18(8):831–836, 1996.

[17] X. Tang and X. Wang. Face sketch synthesis and recognition. *Proc. IEEE International Conf. on Computer Vision*, pages 687–694, 2003.

[18] X. Tang and X. Wang. Face sketch recognition. *IEEE Tran. on Circuits and Systems for Video Technology-*, 14(1):50–57, 2004.

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

[20] X. Wang and X. Tang. Dual-space linear discriminant analysis for face recognition. *Proc. International Conf. on Computer Vision and Pattern Recognition*, pages 564–569, 2004.

[21] X. Wang and X. Tang. A unified framework for subspace face recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 26(9):1222–1228, 2004.

[22] X. Wang and X. Tang. Random sampling for subspace face recognition. *International Journal of Computer Vision*, 70(1):91–104, 2006.

[23] S. Yan, B. Z. D. Xu, and H. Zhang. Graph embedding: a general framework for dimensionality reduction. *Proc. International Conf. on Computer Vision and Pattern Recognition*, 2, 2005.

[24] J. Yang, A. Frangi, J. Yang, D. Zhang, and Z. Jin. KPCA plus LDA: a complete kernel Fisher discriminant framework for feature extraction and recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(2):230–244, 2005.

[25] D. Zhao. Formulating LLE using alignment technique. *Pattern Recognition*, 39:2233–2235, 2006.

[26] D. Zhao. Numerical geomtry of data manifolds. *Shanghai Jiao Tong University*, Master Thesis, Jan. 2006.

[27] D. Zhao, C. Liu, and Y. Zhang. Discriminant feature extraction using dual-objective optimization model. *Pattern Recognition Letters*, 27:929–936, 2006.

[28] M. Zhu and A. Martinez. Selecting principal components in a two-stage LDA algorithm. *Rroc. International Conf. on Computer Vision and Pattern Recognition*, 1:132–137, 2006.