Orthogonal Locality Fisher Discriminant Projections for face recognition

Xiaojing Hou¹, Guoqiang Wang¹*, Mengluo Ji¹, Huanlong Zhang²

¹School of Computer and Information Engineering, Luoyang Institute of Science and Technology, Henan, China
²College of Electric and Information Engineering, Zhengzhou University of Light Industry, Henan, China

Abstract

A subspace learning method, orthogonal locality Fisher discriminant projections (OLFDP), is proposed. The goal of OLFDP is to seek a projection transformation such that geometry information and discriminant information in high dimension space are still preserved in low dimension subspace after dimensionality reduction. OLFDP not only inherits the advantages of manifold learning which has property of preserving local geometry structure, but also makes full use of class label information. In order to improve the discriminating power, we use Schmidt orthogonalizing to orthogonalize the basis eigenvectors. On the base of OLFDP method, orthogonalized kernel locality Fisher discriminant projections method (OKLFDP) is proposed. OLFDP is extended to the nonlinear method by the set of OKLFDP’s vectors. Experiment results on the BAC2005 face image database demonstrate the effectiveness of the proposed method.

Keywords: Subspace Learning, Orthogonal Locality Fisher Discriminant Projections, Manifold Learning, Kernel Trick, Face Recognition

1. INTRODUCTION

Face recognition has become an active research topic in computer vision, owing to its wide application in many areas such as public safety, financial security and human-computer interaction. In the past few decades, many face recognition methods have been presented, among which subspace learning is one of the most popular methods and proved to work well (Zhang & Ding et al., 2015; Hu, 2008). Eigenface and fisher face are well-known in subspace learning. Their main idea is using Principal Component Analysis (PCA) (Turk and Pentland, 1991) or Linear Discriminant Analysis (LDA) (Belhumeur and Hespanha, 1997). However, both of them work with the assumption that face images are located in linear embedded manifold, and goal of the algorithms is to keep the global Euclidean space structure of image space. It is shown by recent research that face images may be distributed in nonlinear manifold space which is nested in some high-dimensional space (Lin and Wang, 2010; Jin and Ruan, 2009; Tenenbaum and De Silva, 2000; Roweis and Saul, 2000). When we use PCA and LDA for dimensional reduction, intrinsic structure of face image hidden in nonlinear submanifolds will not be found. By study work in sample local geometry, some manifold learning methods are proposed to find nonlinear manifold structure, such as Isometrical Mapping (ISOMAP) (Tenenbaum and De Silva, 2000), Locally Linear Embedding (LLE) (Roweis and Saul, 2000), Laplacian Eigenmap (Belkin and Niyogi, 2003) and Local Tangent Space Alignment (LTSA) (Zhang and Zha, 2004). However, all of them can only map training data but can’t directly map new test points. Dealing with the problem, Locality Preserving Protections (LPP) is proposed by He. at el. LPP can not only map new test samples to low-dimensional space established by learning, but also inherit attribute of Laplacian Eigenmap to preserve local structure (He & Yan et al., 2005). It is valid, but it doesn’t take class label information into account which is important to face recognition (Wang and Shu, 2013). In addition, basis function of LPP is nonorthogonal. Nonorthogonality makes it fail to keep measure structure of high-dimensional space and estimate connotation dimension of high-dimensional data (Lei and Qi, 2014). In fact, wrong estimates will reduce the discriminant performance.
A subspace learning method, orthogonal locality Fisher discriminant projections (OLFDP), is proposed in this paper. It integrates manifold learning and classification ability into the algorithm. OLFDP not only inherits the advantages of manifold learning to preserve local geometry structure, but also makes full use of class label information. In order to elevate the discriminating power, we use Schmidt orthogonalizing to orthogonalize the basis eigenvectors. However, OLFDP is linear learning method. Based on OLFDP, a nonlinear method named OKLFDP (Orthogonalized Kernel Locality Fisher Discriminant Projections) is proposed in this article.

2. ORTHOGONAL LOCALITY FISHER DISCRIMINANT PROJECTIONS

2.1 Locality Fisher Discriminant Projections (LFDP)

Suppose we have N data points \( \{x_i\}_{i=1}^N \subset \mathbb{R}^D \) in submanifold \( M \), and data matrix is described as \( X=[x_1, x_2, ..., x_N] \). Then the proposed LFDP can be realized by the following three steps.

(1) Construct neighborhood graphs. Let \( G_w \) denote within-class neighborhood graph and \( G_b \) denote between-class neighborhood graph. In order to find geometry and discrimination information between data points on submanifold \( M \), we construct \( G_w \) and \( G_b \). For each sample point \( x_i \) on \( M \), specify \( N(x_i) = \{x'_1, x'_2, ..., x'_N\} \) is the set of \( k \) nearest neighbors to \( x_i \) and \( l(x_i) \) is the class label of \( x_i \). \( N_w(x_i) \) refers to the set of points having same class label with \( x_i \) in \( N(x_i) \), \( N_b(x_i) \) refers to the set of points having different class labels with \( x_i \) in \( N(x_i) \), and they can be given by

\[
N_w(x_i) = \{x'_j \mid l(x'_j) = l(x_i), 1 \leq j \leq k \} \quad (1)
\]

\[
N_b(x_i) = \{x'_j \mid l(x'_j) \neq l(x_i), 1 \leq j \leq k \} \quad (2)
\]

Then we can distinguish between \( G_w \) and \( G_b \) according to the definition. Consider each \( x_i \) in \( G_w \), edges are added between \( x_i \) and each point in \( N_w(x_i) \). Likewise, for each \( x_i \) in \( G_b \), edge is added between \( x_i \) and each point in \( N_b(x_i) \).

(2) Compute weights. Specify \( W_w \) and \( W_b \) are weight matrix of \( G_w \) and \( G_b \) respectively, they can be given by

\[
w_{b,ij} = \begin{cases} 
\exp\left(-\frac{\|x_i - x_j\|^2}{t}\right), & \text{if } x_i \in N_b(x_j) \text{ or } x_j \in N_b(x_i) \\
0, & \text{otherwise} 
\end{cases} \quad (3)
\]

\[
w_{w,ij} = \begin{cases} 
\exp\left(-\frac{\|x_i - x_j\|^2}{t}\right), & \text{if } x_i \in N_w(x_j) \text{ or } x_j \in N_w(x_i) \\
0, & \text{otherwise} 
\end{cases} \quad (4)
\]

(3) Compute projection transformation matrix \( U \). Projection transformation matrix can be obtained via solving the problem of generalized eigenvector in

\[
S_u = \lambda S_w u \quad (5)
\]

Where \( S_b = X(D_b - W_b)X^T \), \( S_w = X(D_w - W_w)X^T \), \( D_b \) and \( D_w \) are both diagonal matrix, whose diagonal elements are respectively \( D_{b,ii} = \sum_j W_{b,ij} \) and \( D_{w,ii} = \sum_j W_{w,ij} \). Thus embedding of \( x_i \) can be obtained by eq.(6):

\[
y_i = U^T x_i \quad (6)
\]

Where \( U = [u_1, u_2, ..., u_d] \).
We now explain the derivation process for above steps of LFDP.

The general problem of subspace learning is to find a linear transformation $U \in \mathbb{R}^{D \times d}$. Techniques originated from manifold learning base on different theories and cost function. Criteria of PCA is to minimize reconstruction error, whereas optimization objective of LDA is to maximize ratio of between-class divergence and within-class divergence. Both of them can’t find internal nonlinear structure of manifold. Different from PCA and LDA, LFDP makes use of locality relation between data points. It integrates Local information between points from different classes and neighbor information between points from same class into algorithm, which helps to find discrimination information more accurately. Thus classification is more effective. Inspired by LPP, LFDP models had local structural relationship by constructing neighborhood graphs. Hence LFDP and LPP have some similar attributes, e.g. explicitly considering nonlinear manifold structure of data sets. The aim of LFDP is to find projection transformation which maximize the distance between points from different classes and at the same time minimize the distance between points in same class. As a result, points from same class will stay as close as possible in low-dimensional space after transformation, and points close in original space will stay as far as possible after dimensionality reduction. The objective function is defined as follows.

$$\max_{U} \frac{\sum_{g} \|y_i - y_j\|^2 W_{b,ij}}{\sum_{g} \|y_i - y_j\|^2 W_{w,ij}}$$ (7)

Maximizing eq.(7) means maximizing numerator and moreover minimizing denominator. Numerator measures the distance between points from different classes. Maximizing it will make points from different classes dispersed as much as possible. In other words, if $x_i$ is near to $x_j$, where $x_i$ and $x_j$ belong to different classes, maximizing numerator will produce a punish value to make sure that the projection $y_i$ stay away from $y_j$. On the other hand, denominator measures the distance between points from same class. Minimizing denominator will make points in same class stay close as much as possible and meanwhile preserve local structure of each class. That is, if $x_i$ is near to $x_j$, $y_i$ will still stay near to $y_j$. Consequently minimizing denominator will produce a punish value if mapping of $x_i$ stays away from $x_j$. In short, numerator reflects within-class distance and denominator reflects between-class distance in eq.(7). The goal of maximizing objective function is to find max distance between different classes, at the same time keep the structure of each class. As the objective function of proposed method is similar with that of Fisher, and local structure is taken into considered, we name it locality Fisher discriminant projections (LFDP). When nearest neighbor classifier is adopted to discriminate, it is helpful to keep the structure of each class.

By Algebra, denominator of eq.(7) is simplified as follows:

$$\frac{1}{2} \sum_{g} \|y_i - y_j\|^2 W_{w,ij} = \frac{1}{2} \sum_{g} tr\{(U^T x_i - U^T x_j)(U^T x_i - U^T x_j)^T\}W_{w,ij} = tr(U^T S_b U)$$ (8)

Where $D_w$ is Diagonal matrix , whose input is sum of column $W_w$ (or sum of raw, because $W_w$ is symmetric), $D_{w,ij} = \sum_f W_{w,ij}$.

In same way, numerator of objective function is simplified as:

$$\frac{1}{2} \sum_{g} \|y_i - y_j\|^2 W_{b,ij} = tr(U^T S_b U)$$ (9)

Notice that traditional LDA uses ratio of determinant. Thus we use determinant paradigm instead of trace paradigm, i.e. $[U^T S_b U]$ instead of trace $(U^T S_b U)$, the objective function is rewritten as:
\[
\max J(U) = \frac{U^T S_b U}{U^T S_w U} \tag{10}
\]

At last, transformation matrix will be gained via finding feature vector corresponding to max-eigen of eq.(11).

\[
S_b u = \lambda S_w u \tag{11}
\]

In face recognition, characteristic dimension is always far greater than sample number, thus the matrix \( S_w \) is singular. To address this issue, we map the data by PCA without lose any information, which make \( S_w \) nonsingular.

### 2.2 Orthogonal Eigenvector

Because \( S_w^{-1} S_b \) is a asymmetric matrix in general, eigenvectors given by LFD is nonorthogonal.

With eigenvectors, transformation matrix is then described as \( U = [u_1, u_2, \ldots, u_d] \). So the Euclidean distance between two points in low-dimensional space after dimensionality reduction is as follows:

\[
dist(y_i, y_j) = \|y_i - y_j\| = \|U^T x_i - U^T x_j\| = \|U^T (x_i - x_j)\| = \sqrt{(x_i - x_j)^T U U^T (x_i - x_j)} \tag{12}
\]

The nonorthogonality prevents from keeping estimated structure of high-dimensional space. In fact, wrong estimation of high-dimensional space will greatly hurt the discriminant performance. On the contrary, if \( U \) is orthogonal, i.e. \( U U^T = I \), estimated structure will be preserved.

Let \( u_1, u_2, \ldots, u_d \) be all the discriminant vectors of LFD, if \( v_1 \) is equal to \( u_1 \), with assumption that we have got k vectors of OLF, of course the \( k+1 \) vector will be as follows:

\[
v_{k+1} = u_{k+1} - \sum_{i=1}^{k} \frac{v_i^T u_{k+1}}{v_i^T v_i} v_i \tag{13}
\]

Known by Linear algebra, \( v_1, v_2, \ldots, v_d \) is actually the form of \( u_1, u_2, \ldots, u_d \) after Schmidt orthogonalization.

We change eq.(13) as the form \( v_k = u_k - \sum_{i=1}^{k-1} h_i u_i \), where \( h_i \) (i=1, 2, ..., k-1, k=2, 3, ..., d) are real numbers.

Let \( V = [v_1, v_2, \ldots, v_d] \), then \( V = UH \), where \( H \) is an upper triangular matrix whose Lord diagonal elements are 1. We get:

\[
J(V) = \left[ \begin{array}{c} (UH)^T S_b (UH) \\ (UH)^T S_w (UH) \end{array} \right] = \left[ \begin{array}{c} H^T(U^T S_b U)H \\ H^T(U^T S_w U)H \end{array} \right] = \left[ \begin{array}{c} H^T U^T S_b U H^T \\ H^T U^T S_w U H^T \end{array} \right] = \left[ \begin{array}{c} U^T S_b U \\ U^T S_w U \end{array} \right] = J(U) \tag{14}
\]

It means that orthogonal vectors of OLFDP maximize objective function of eq.(10). That’s why performance of OLFDP is superior to LFD on experiments.

### 3. ORTHOGONALIZED KENEL LOCALITYFISHERDISCRIMINANTPROJECTIONS

OLFDP is thus a linear algorithm. Actually, Face recognition is always a nonlinear problem because that image of a face always changes with facial expression, illumination conditions, viewpoint, etc. Especially when severe illumination change (such as shadow) and wide view change occur, recognition problem becomes serious nonlinear. research work in (Adini and Moses, 1997) showed Distances between images of the same individual with different facial expression or illumination conditions tend to be larger than the distances between images of different individuals with same facial expression or illumination conditions. Obviously, linear method has some limitations in this application.
This section, we extend OLFDP method proposed above to Orthogonalized Kernel Locality Fisher Discriminant Projection (hereafter referred to as OKLFDP) (Scholkopf et al., 1998; Yang, 2002; Li et al., 2007). Derivation is given below.

Suppose $X$ is a dataset of $D$ dimensional vector, $X=\{x_1, x_2, \ldots, x_N\}$. Input vector space $\mathbb{R}^D$ is mapped to a high-dimensional feature space by a nonlinear mapping function $\Phi$. And then LFDP transform is implemented in high-dimensional feature space $\mathbb{R}^F$.

$$V_{opt} = \arg \max_v \left[ \frac{V^T S_b \Phi V}{V^T S_w \Phi V} \right]$$  \hspace{1cm} (15)

$$S_b = \Phi(D_b - W_b)\Phi^T$$  \hspace{1cm} (16)

$$S_w = \Phi(D_w - W_w)\Phi^T$$  \hspace{1cm} (17)

Where $\Phi = [\Phi(x_1), \Phi(x_2), \ldots, \Phi(x_N)]$

The corresponding characteristic equation is as follows.

$$S_b v = \lambda S_w v$$  \hspace{1cm} (18)

Usually, dimension of the feature space is very high, sometimes even infinite. Therefore, direct calculation of LFDP in high dimensional space is often very difficult, even impossible. Similar to KPCA, by introducing kernel function of support vector machine (SVM), we will use kernel function instead of the inner product of nonlinear function in the high-dimensional feature space. The advantage is that we can solve the problem just in input space without knowing the specific form of nonlinear function.

Because eigenvectors can be expressed in a linear combination of the elements in $F$, the coefficient vector $a= [a_1, a_2, \ldots, a_N]^T$ satisfies eq.(19):

$$v = \sum_{i=1}^N a_i \Phi(x_i) = \Phi a$$  \hspace{1cm} (19)

Then we get:

$$[\Phi(x_1), \Phi(x_2), \ldots, \Phi(x_N)](D_b - W_b)[\Phi(x_1), \Phi(x_2), \ldots, \Phi(x_N)]^T \sum_{i=1}^N a_i \Phi(x_i)$$

$$= [\Phi(x_1), \Phi(x_2), \ldots, \Phi(x_N)](D_w - W_w)[\Phi(x_1), \Phi(x_2), \ldots, \Phi(x_N)]^T \sum_{i=1}^N a_i \Phi(x_i)$$  \hspace{1cm} (20)

Both sides of eq.(20) are multiplied by $[\Phi(x_i)]^T$, then we get:

$$[\Phi(x_1) \cdot \Phi(x_1), \Phi(x_1) \cdot \Phi(x_2), \ldots, \Phi(x_1) \cdot \Phi(x_N)](D_b - W_b) \sum_{i=1}^N a_i \Phi(x_i) =$$

$$[\Phi(x_2) \cdot \Phi(x_1), \Phi(x_2) \cdot \Phi(x_2), \ldots, \Phi(x_N) \cdot \Phi(x_1)]$$
We define a nuclear matrix $K_{N \times N}$ (its elements are described as $K_{ij} = (\Phi(x_i))^T \Phi(x_j)$) and vector $\alpha = [\alpha_1, \alpha_2, ..., \alpha_N]$, then we get eq.(22).

$$K(D_b - W_b)K\alpha = \lambda K(D_w - W_w)K\alpha$$

Now that calculation of KLDPD boils down to a problem of generalized eigenvalue decomposition.

The first $d$ discriminant vectors of KLDPD are eigenvectors corresponding to first $d$ maximum eigenvalue of matrix $(K(D_w - W_w)K)^{-1}(K(D_b - W_b)K)$. Its schmidt orthogonalization group is vector set of OKLDPD. Suppose first $d$ vector sets of OKLDPD are $\beta_1, \beta_2, ..., \beta_d$, the $(d+1)$th vector is

$$\beta_{d+1} = a_{d+1} - \sum_{i=1}^{d} \frac{\beta_i^T a_{d+1}}{\beta_i^T \beta_i} \beta_i$$

4. EXPERIMENTS

In this section, we verify the effectiveness of the proposed algorithm by experiments on 2005 national biometric competition face database (BAC2005). BAC 2005 database includes 2340 images of 117 people (Each person has 20 images). We resize each image to 50 by 50 pixels. Some images for one person in BAC2005 face database are shown in figure 1 below. Easy to see, images in BAC2005 contain rich change of facial expression, illumination and attitude. Obviously, these changes are closer to the truth. Using such a complex database will effectively illustrate the effectiveness of our method.

**Figure 1.** Example face images in BAC2005 face database

Distribution of facial patterns becomes complex and nonconvex when illumination, expression and posture changes drastically. At first step, we make experiments to validate the nonlinear nuclear identification method’s ability to adapt complex distribution of face pattern. Ten images are extracted randomly from each person’s twenty images as training samples, then we get 1170 images altogether. We use LFDP, OLDPD, KPCA, KFDA and OKLDPD methods for training, hence get their respective characteristics. Among the generated subspace, LFDP, OLDPD method get linear ones, KPCA, KFDA and OKLDPD get nonlinear ones. Then we extracted seven people’s images randomly, ten for each person, seventy images altogether, and projected the images to five kinds of subspace got above. For each image, two most important characteristics of each subspace’s projection visualization are as shown in fig.2, in which images belonging to a same person are described as same symbols (such as “*”, “o”). Here we adopt standard gaussian kernel (suppose $\sigma = 10000$).
\[ k(x, y) = \exp\left(-\frac{\|x - y\|_2^2}{2\sigma^2}\right) \] (24)

As can be seen from fig.2(a), only one kind of face samples is linearly separable from other kinds. While in fig.2(b), there are three kinds of face samples are linearly separable. The reason is that comparing with LFDP, OLFDP can retain measure relationship in high dimensional space. Nevertheless, both LFDP and OLFDP are linear classification methods, their performance are not satisfactory. In fig.2(c), KPCA algorithm changes the data distribution by nonlinear mapping; however, the distribution is basically desultorily. This is due to that KPCA algorithm minimizes the reconstruction error of samples without considering the reparability of the samples in feature space. Class information optimization goal is taken into account in KFDA method, so it gets low dimensional representation. The most distinguishing features extracted by KFDA are described in fig.2(d). Compared with the former methods, KFDA has better reparability. Meanwhile KFDA does not take local neighborhood into relationship account, whereas OKLFDP combines with neighborhood and class information. In fig.2(e), all of the 7 kinds of face samples are almost completely separable.

![Figure 2. Distribution of 70 Face Samples of 7 Subjects in Different Subspace](image)

The following experiment will validate the effectiveness of the algorithm of face recognition based on nuclear. In order to balance the influence of different training sample sets and make the correct evaluation for algorithms, ten images of each people are extracted randomly as training set, rest ten images of each people compose test set. Training and test sets both have 1170 samples, and they don’t overlap each other. Due to the choosing of kernel function and the corresponding parameters are closely associated with to solve the problem of, different kernel functions meeting the Mercer’s theorem are used in experiments, including Gaussian kernel function \( k(x, y) = \exp\left(-\frac{\|x - y\|_2^2}{2\sigma^2}\right) \) and Polynomial kernel function \( k(x, y) = (1 + x \cdot y)^d \). To choose the parameters \( \sigma \) of Gaussian kernel function and parameters \( d \) of polynomial kernel function, we firstly use the fixed sample set, then adjust them repeatedly in experiment, at last find the best \( \sigma=10000 \) and the most optimal \( d=0.895 \). Fig.3 shows recognition rate of Gaussian kernel and polynomial kernel on different dimension. It can be seen that our method almost wins the best recognition rate on each dimension. According to the optimal kernel parameters obtained, we repeat ten experiments, and take the average results as estimate of the final recognition rate. Tab.1 shows the optimal results for each method.
5. CONCLUSION

A novel subspace learning method, orthogonal locality Fisher discriminant projections (OLFDP), is proposed. OLFDP method takes local properties of data points and class label information into account. It integrates manifold learning and classification ability into the algorithm. Schmidt orthogonalization is used to get orthogonal subspaces. Experimental results on BAC2005 face image database show that the performance of OKLFDP method is superior to those of LFDP, OLFDP, KPCA, KFDA, and it can achieve the highest recognition rate (98.64%). Experiment results demonstrate the effectiveness of the proposed method. However, for both OKLFDP method proposed here and existing method of KPCA, KFDA, choosing the proper kernel function can greatly improve the effect of recognition result. Yet how to determine the optimal kernel function and its parameters also lack of guiding theory now, more research work is needed.

6. ACKNOWLEDGMENT

This work was supported in part by Grant from the National Natural Science Foundation of China (No.61503173), Science and Technology Program of Henan Province (No.152300410115); Foundation of Henan Educational Committee (No.14A520055).

7. REFERENCES


