2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

A Flexible and Efficient Algorithm for Regularized Marginal Fisher Analysis

Jinrong He, Lixin Ding, Lei Jiang and Li Huang

***Abstract*—Marginal Fisher analysis (MFA) is a well-known linear dimensionality reduction method. However, MFA does not utilize the local diversity information of the training data, which will degrade its performance. In order to enhance the discriminant power of MFA, this paper considers introducing local variation quantity to enlarge the distances between local neighborhood embeddings and proposes a flexible and efficient implementation of MFA (F-MFA) within the regularization framework. Therefore, the discriminant structure and diversity of data are preserved in low-dimensional subspace. Computationally, F-MFA is formulated as a trace differential optimization problem which can completely avoids the singularity problem as it exists in MFA. Further, an efficient algorithm is developed for implementing F-MFA via QR-decomposition. Experimental results on four face data sets demonstrate the effectiveness of our approach.**

I. INTRODUCTION

In past decade, accompanying the advancement of sciences and technologies, scientific data has the tendency of growing in both size and complexity, such that extracting useful knowledge from it is often much harder. The techniques of dimensionality reduction have received broad attention in areas such as data mining, machine learning, and computer vision. It is an important data preparation step which transforms the original high-dimensional data into a lower-dimensional space with limited loss of information, leading us to better models for data analysis.

The popular dimensionality reduction algorithms can be divided into two groups: linear and nonlinear. Currently, nonlinear methods can mainly be divided into two classes: manifold learning-based methods and kernel-based methods. The former aims to preserve the local structure of data points. The representative approaches are Locally Linear Embedding (LLE) [1], Laplacian Eigenmaps (LE) [22], Hessian Eigenmaps (HE) [3], Isometric Mapping (ISOMAP) [4], Maximum Variance Unfolding (MVU) [5], Manifold charting (MC) [6], Local Tangent Space Alignment (LTSA) [7], and others. Kernel-based methods aim to map the input data points into a much higher feature space via a nonlinear mapping and then carry out a linear method using the mapped samples. Kernel principal component analysis (KPCA) [8] and Generalized Discriminant Analysis (GDA) [9] are the representative approaches, which are effectively applied to face recognition. With the help of kernel trick, non-linear problems can be solved in linear case. Though many of nonlinear methods have been validated to be effective, these methods are typically associated with high computational overhead, and the nonlinear projection defined only on the training data space cannot extended to testing data directly , which is also called out-of-sample problem, making them difficult to be applied on real-world data analysis problems.

In recent years, linear dimensionality reduction techniques are of particular interest for researchers since they are simple to calculate and analytically analyze. Two traditional linear dimensionality reduction approaches are Principle Component Analysis (PCA)[10][11] and Linear Discriminant Analysis (LDA)[12][13]. PCA seeks a subspace, in which the projected global variance reaches maximization. LDA seeks a subspace projected onto which the data points of different classes are far away while the data points of the same class are close to each other. These global methods fail to discover the local structure of underlying manifold. In many real world applications such as face recognition, there may not be sufficient training samples. In this case, it may not be able to accurately estimate the global structure, thus the local structure becomes more important.

Recently, Yan et al. proposed a general dimensionality reduction framework called graph embedding [14] which has been shown to be effective in discovering the local geometrical structure of data points. The typical graph based algorithms includes Locality Preserving Projections (LPP) [15], Unsupervised Discriminant Projection (UDP) [16], Local Discriminant Embedding (LDE) [17], Marginal Fisher Analysis (MFA)[14], etc. However, locality characterization of the data is not originally and essentially designed for pattern discrimination purpose. For example, if the patterns lied on multimanifolds and there may be several subclasses in on class, then the locality-preserving algorithms may result in overlapped embeddings belonging to different classes, which may impair the local topology of data, leading to unstable intrinsic structure representation and bad discrimination

Jinrong He is with the School of Computer, State Key Laboratory of Software Engineering, Wuhan University, Wuhan, 430072 China (corresponding author to provide phone: 086-15392822848; e-mail: hejinrong@whu.edu.cn).

Lixin Ding is with State Key Laboratory of Software Engineering, Wuhan University, Wuhan, 430072 China (e-mail: lxding@whu.edu.cn).

Lei Jiang is with Key Laboratory of Knowledge Processing and Networked Manufacture, Hunan University of Science and Technology, Xiangtan, 411201 China (e-mail: jlefe@126.com).

Li Huang is currently a Lecturer with the School of Computer & Information Technology, Shanxi University, Taiyuan, Shanxi, 030006 China. (e-mail:l_huang@126.com)

This work was supported in part by the Fundamental Research Funds for the Central Universities (No. 2012211020209), Special Project on the Integration of Industry, Education and Research of Ministry of Education and Guangdong Province (2011B090400477), Special Project on the Integration of Industry, Education and Research of Zhuhai City (2011A050101005, 2012D0501990016), Zhuhai Key Laboratory Program for Science and Technique (2012D0501990026).

978-1-4799-1484-5/14/$31.00 ©2014 IEEE

performance [18]. Therefore, this can make the algorithm apt to overfit the training data and sensitive to the outliers. In other words, for classification problem, the locality quantity itself is not sufficient.

Recently, following the basic idea of MFA, many variants of dimensionality reduction algorithms have been developed. Quanxue Gao et al.[19] demonstrated that the local variation among the same class characterizes the most important modes of variability of patterns, which will help to improve the stableness of the algorithm. They proposed an algorithm named Stable Orthogonal Local Discriminant Embedding (SOLDE). However, the formulation of SOLDE reduces to the solution of a generalized eigenproblem that requires the scatter matrices in the denominator to be nonsingular. This can become problematic when the dimensionality is larger than the number of samples, which is also called small sample size (SSS)[20] problem. Besides, these algorithms implicitly consider that the inter-class and intra-class relations are equally important. This reduces the flexibility of the algorithm.

To remedy these deficiencies, we propose a flexible and efficient algorithm for Marginal Fisher Analysis (F-MFA), to perform linear supervised dimensionality reduction. Motivated by the idea of SOLDE, F-MFA takes local discriminant information and local variation into account simultaneously in the modeling of manifold, and presents an objective function that seeks to maximize the difference, rather than the ratio, between the regularized local inter-class scatter and local intra-class scatter. Thus, F-MFA can obtain the mutually orthogonal projection directions efficiently. Experimental results on several face databases demonstrate the effectiveness of the proposed algorithm.

The paper is organized as follows. Section 2 analyzes the problem. In Section 3 we propose regularized extension of MFA. An efficient algorithm is presented in Section 4. We conduct the empirical comparisons in Section 5 and conclude in Section 6.

II. Problem Statements

A. Problem Formulation and Notation

Before starting, it is useful to define general terms in this paper we used. Let $X=\{x_1,x_2,\cdots,x_n\}\in R^{d\times n}$ be a data matrix whose columns are training samples and rows are features. The label of a sample denoted by $label(x_i)\in C$, here C is class label set $C=\{c_1,c_2,\cdots,c_m\}$. For linear dimensionality reduction methods, we particularly assume that each data point x_i is mapped to a lower dimensional point y_i through a linear transformation matrix V, which can be written as

$$Y=V^TX, \text{ where } V\in R^{d\times r} \tag{1}$$

here $r<d$. As it is known that orthogonality is of utmost importance to discriminant analysis, since redundant features can be combined back to the same number of variables through orthogonal transformation of the measurement space.

Graph based dimensionality reduction methods represent the data samples as nodes and quantify the similarity among pairs of samples as edges. For completeness, the process of constructing the affinity graph is summarized here, and the details can be found in [1] and [2].

B. Graph Preserving Criterion

Generally speaking, dimensionality reduction is conducted based on a well-defined criterion. Yan et al. [14] claimed that most dimensionality reduction algorithms can be unified into a general framework, namely graph embedding which is described as follows.

Let $G=\{X,W\}$ be an undirected weighted graph with vertex set X and weight matrix $W\in R^{n\times n}$ which is the distance (or similarity) measure between data vertex. The Laplacian matrix L of the graph G are defined as[21]

$$L=D-W \tag{2}$$

here D is a diagonal matrix and $D_{ii}=\sum_{j=1}^{n}W_{ij}$. Clearly, L is symmetric and positive semi-definite. The weight could be realized by the heat kernel (Gaussian kernel) [2] that is defined as

$$W_{ij}=\exp\left(-\frac{\left\|x_i-x_j\right\|^2}{t}\right) \tag{3}$$

where $t>0$ is called local scaling regulator and controls the overall scale or the smoothing of the space.

In dimensionality reduction, there is an assumption that nearby data points are likely to have the low-dimensional representation with similar property or structure. Thus, a natural local information preserving criterion can be defined as

$$\min\sum_{i,j=1}^{n}W_{ij}\left\|y_i-y_j\right\|^2 \tag{4}$$

For any data point x_i, the objective function (4) computes the weighted sum of all squared pairwise Euclidean distances between the data points x_i and x_j that are within the k-nearest neighborhood of x_i. By simple algebra formulation, the objective function (4) can be reduced to following concise matrix form.

$$\begin{aligned}\sum_{i,j=1}^{n}W_{ij}\left\|y_i-y_j\right\|_2^2&=\sum_{i,j=1}^{n}W_{ij}tr\left((y_i-y_j)(y_i-y_j)^T\right)\\&=2tr\left(\sum_{i,j=1}^{n}W_{ij}\cdot y_iy_i^T-\sum_{i,j=1}^{n}W_{ij}\cdot y_iy_j^T\right)\\&=2tr\left(YLY^T\right)\end{aligned} \tag{5}$$

where $S_Y=YLY^T$ is the scatter matrix in the low-dimensional space. Thus, the trace of scatter matrix measures the distance between some specific data points.

For a specific dimensionality reduction algorithm, there may exist two graphs, the intrinsic graph $G=\{X,W\}$ and the penalty graph $G^P=\{X,W^P\}$ with $L^P=D^P-W^P$ and $D_{ii}^P=\sum_{j\neq i}W_{ij}^P$. The intrinsic graph characterizes data properties that the algorithms aim to favor and the penalty graph describes properties that the algorithms aim to avoid. A

linear graph preserving criterion is imposed for these two objectives.

$$\arg\max_{V} \frac{\sum_{i\neq j}\left\|V^T x_i - V^T x_j\right\|^2 W_{ij}^p}{\sum_{i\neq j}\left\|V^T x_i - V^T x_j\right\|^2 W_{ij}} \tag{6}$$

which can be further formulated in trace ratio form:

$$\arg\max_{V} \frac{tr(V^T X L^P X^T V)}{tr(V^T X L X^T V)} \tag{7}$$

This trace ratio form has been successfully used as a general criterion for dimensionality reduction previously. In some case, we can consider the difference-form formulation

$$\max tr\left(W^T X\left(L^P - L\right)X^T W\right) \tag{8}$$

This trace different criterion has been successfully in many algorithms, such as Maximum Margin Criterion (MMC)[22] and Locality Sensitive Discriminant Analysis (LSDA) [23].

C. Limitations of MFA

It can be seen that the objective function of MFA well preserves the similarity of local intra-class data points and discriminant of local inter-class data points. However, the objective function (7) results in the following problems.

(1) **Distort the local intrinsic geometry of data**. MFA emphasizes the data pairs with large distance pairs, which may result in that points with small distance are not embedded nearby in the embedding space. Thus, it may impair the local topology. Moreover, it ignores the variation, which characterizes the different geometrical properties, i.e. diversity of data, resulting in unstable intrinsic structure representation and making it cannot unfold the manifold structure of data.

(2) **It suffers from the small sample size problem**. A difficulty in using the MFA method for image recognition is the high-dimensional nature of the image space, in such a space, the XDXT matrix is always singular, which makes the direct implementation of the MFA algorithm impossible.

(3) **The projection vectors obtained by MFA are not orthogonal**. This makes it difficult to reconstruct the data. The advantage of employing orthogonal transformation is that the correlations among candidate features are decomposed so that the significance of individual features can independently be evaluated [24].

III. Proposed Method

A. Motivations

As above mentioned, the diversity among nearby data is very important for intrinsic geometry preserving and local manifold structure unfolding.

Similar with graph embedding framework, the characterization of local variation in F-MFA is based on the **diversity graph** that incorporates the neighborhood information of the data points, and by contrast, the characterization of discriminant information is based on a **similarity graph** and **dissimilarity graph** that embodies the neighborhood information of the data points belongs to same and different classes, respectively. In this way, the geometrical and discriminant structure of the data manifold can be accurately characterized by these three graphs. The learning procedure is illustrated in Fig. 1. The detail will be described in following subsections.

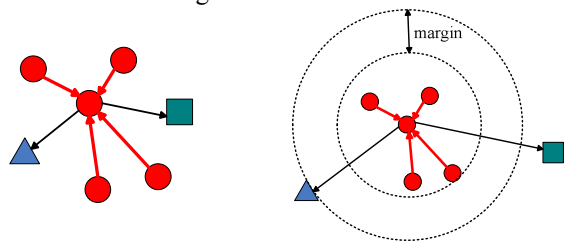


Fig.1. (a) The center point has six neighbors. The points with the same color and shape belong to the same class. (b) After projection, the margin between different classes is maximized and the local variation is preserved.

B. Similarity Preserving Model

In order to preserving the similarity of data, the nearby points belonging to same class in the observed data space should be mapped as close together as possible in the embedding space.

In the similarity graph, the weight matrix W is defined as:

$$W_{ij}^{(S)} = \begin{cases} 1 & i \in N_{k_1}^+(j) \vee j \in N_{k_1}^+(i) \\ 0 & otherwise \end{cases} \tag{9}$$

Here $N_{k_1}^+(i)$ indicates the index set of the k_1 nearest neighbors of the sample x_i in the same class. In the projected low-dimensional space, the intra-class compactness is characterized by the following objective functions:

$$\min \sum_{i,j}\left\|V^T x_i - V^T x_j\right\|^2 W_{ij}^{(S)} \tag{10}$$

Equivalently, it can be rewritten as

$$\min tr(V^T X L^{(S)} X^T V) \tag{11}$$

where $L^{(S)} = D^{(S)} - W^{(S)}$ is the Laplacian matrix of similarity graph. The objective function (10) on the similarity graph incurs a heavy penalty if neighboring points x_i and x_j are mapped far apart while they are actually in the same class. Therefore, minimizing (10) is an attempt to ensure that if x_i and x_j are close and sharing the same label, then their corresponding low-dimensional representations are close as well.

C. Dissimilarity Preserving Model

In the dissimilarity graph, the weight matrix W is defined as:

$$W_{ij}^{(D)} = \begin{cases} 1 & i \in N_{k_2}^-(j) \vee j \in N_{k_2}^-(i) \\ 0 & otherwise \end{cases} \tag{12}$$

here $N_{k_2}^-(i)$ is the index set of the k_2 nearest neighbors of the sample x_i in the different class. According to graph preserving criterion, it follows that

$$\max \sum_{i,j}\left\|V^T x_i - V^T x_j\right\|^2 W_{ij}^{(D)} \tag{13}$$

Equivalently, it can be rewritten as

$$\max tr(V^T X L^{(D)} X^T V) \tag{14}$$

where $L^{(D)} = D^{(D)} - W^{(D)}$ is the Laplacian matrix of dissimilarity graph. The objective function (13) on dissimilarity graph incurs a heavy penalty if neighboring points x_i and x_j are mapped close together while they actually belong to different classes. Therefore, maximizing (13) is an attempt to ensure that if x_i and x_j are close but have different label, then their corresponding low-dimensional representations should be mapped far apart.

D. Diversity Preserving Model

In real-world applications, data points in the neighborhood may come from different classes, and the variation of data points from the same class may reflects the diversity of data points, while the variation of data points from different classes characterizes the discriminating information. Then the weight matrix of local variation can be defined as[25]

$$W_{ij}^{LV} = \begin{cases} \exp\left(-\dfrac{t}{\left\|x_i - x_j\right\|^2}\right) & x_i \in N_k(x_j) \vee x_j \in N_k(x_i) \\ 0 \quad otherwise \end{cases} \tag{15}$$

Here $N_k(x_i)$ indicates the index set of the k nearest neighbors of the sample x_i. In the projected low-dimensional space, the total local variation [24] is preserved by the following minimization problem:

$$\max \quad \sum_{i,j} \left\|V^T x_i - V^T x_j\right\|^2 W_{ij}^{(LV)} \tag{16}$$

Equivalently, it can be rewritten as

$$\max \quad tr(V^T X L^{(LV)} X^T V) \tag{17}$$

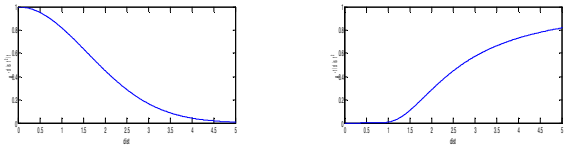
where $L^{(LV)} = D^{(\mathrm{LV})} - W^{(\mathrm{LV})}$ is the Laplacian matrix of diversity graph. By maximizing local variation, we obtain a low-dimensional space that well preserves the intrinsic geometrical structure that characterizes the diversity and discriminating information of data.

As for the diversity preserving model, its properties and the corresponding advantages can be summarized as follows:

Property 1: For data points in any local region, if the variation among nearby data points in the original data space is large, then the variation among the corresponding low-dimensional representations should be large. This gives a certain chance to the points in the same class to be "less similar", i.e. to have a certain value of diversity. This is suitable for multimodal data classification tasks.

Property 2: W_{ij}^{LV} is monotonously increasing with respect to the pairwise Euclidean distance between x_i and x_j. The comparison between weight function (3) and (15) is showed in Fig. 2. With the decreasing of the Euclidean distance, the local variation weight decreases toward 0. It means close points should have a smaller value of diversity. On the other hand, a heavy weight is put between mutually distant samples, which is useful in emphasizing atypical samples and, therefore, makes F-MFA robust to outliers.

Property 3: LPP aims to produce a subspace that preserves the local structure of the data set. However, LPP cannot necessarily guarantee to project mutually distant data points into distant embeddings. Thus, we introduce objective (16) to serve as this purpose, by ensuring that two mutually distant sample are projected as apart as possible. This prevents the neighborhood relationship from being forcefully distorted and the main geometric structure of the data set can be largely preserved. Therefore, it endows the F-MFA with the ability of topology preserving.



(a) (b)

Fig.2. Weight function showing for Gaussian weight (a) and local variation weight (b).

As a consequence of these properties, the diversity graph preserves the intrinsic structure of each neighborhood. Thus, local diversity and intrinsic structure can be preserved by maximizing the variance of data in the local neighborhoods.

E. Formulation of F-MFA

Given above three individual optimization objectives (11), (14) and (17), F-MFA combines the aforementioned insight into the unified objective function which can be formulated as trace difference form:

$$\arg\max_V \left\{tr(V^T X L^{(\mathrm{P})} X^T V) - tr(V^T X L^{(S)} X^T V)\right\} \tag{18}$$

where $L^{(\mathrm{P})} = \alpha L^{(D)} + (1-\alpha) L^{(\mathrm{LV})}$ and α ($0 \le \alpha \le 1$) is a regularization parameter which controls the tradeoff between (14) and (17). The larger the α is, the more favorable the inter-class separability is to win. Maximizing (18) is to find projections such that the intra-class data points are attracted closer (minimizing the (11)) while inter-class data points and mutually distant data points are simultaneously pulled farther away (maximizing the (14) and (17)).

In order to obtain a flexible model, we write

$$L = \beta L^{(P)} + (1-\beta) L^{(S)} \tag{19}$$

The additional regularization parameter β allows one to tune the balance between both terms. With β close to one, local inter-class separability and local variation are dominated for dimensionality reduction, but local intra-class compactness can possibly be neglected. In contrast, with β close to zero, the local intra-class similarity is better preserved, often at the price of some errors in local variation and local inter-class dissimilarity. The criterion in (18) can be maximized by solving

$$\arg\max_V tr(V^T X L X^T V) \tag{20}$$

We can simply enforce the mapping to be orthogonal, and then we obtain the following optimization problem:

$$\begin{cases} \arg\max\limits_V \quad tr(V^T X L X^T V) \\ s.t. \quad V^T V = I \end{cases} \tag{21}$$

The projection matrix V that maximizes the objective function (21) is given by the maximum eigenvalue solution to the standard eigenvalue problem:

$$XLX^T V = \lambda V \tag{22}$$

In this case, the projection matrix V is the set constituted by

the eigenvectors associated with the largest eigenvalues of the matrix XLX^T. Note that the matrix XLX^T is symmetric, thus the matrix V has the orthogonal columns. Thus, the low-dimensional representations are as follows:

$$y_i = V^T x_i \quad i = 1, 2, \cdots, n. \tag{23}$$

It is worthwhile to highlight some properties of F-MFA from a number of perspectives.

(1) F-MFA reflects the **intrinsic geometry of data points**. By considering the variation of data, which characterizes the different geometrical properties, i.e. diversity of data, the locality based dimensionality reduction algorithms can be enhanced to unfold the manifold structure of data. Based on this investigation, it can be seen that the variation among nearby data points characterizes the intrinsic geometry of data points and helps to improve the generalization capability.

(2) Similar to MFA, F-MFA is linear and defined on both the training and the testing data sets, thus it can **avoid out-of-sample problem**. However, MFA only considers the local inter-class scatter and local intra-class scatter while F-MFA takes local variation as an additional regularization term which really reflects the intrinsic geometry of the data set.

(3) F-MFA can avoid the **singularity** problem. When the number of the samples is much smaller than the dimension of the sample space, there will be singularity problem in trace ratio optimization. Although both SOLDE and F-MFA considers the local variation of data, F-MFA formulates the objectives in trace difference criterion which can effectively avoid the inverse matrix operation and the small sample size problem and makes the implementation much easier.

(4) F-MFA produces **orthogonal** projection matrix. This makes it easy to reconstruct the data and preserving the global geometry. Although SOLDE also produces orthogonal projection matrix, it is computationally complex.

IV. Efficient Algorithm for F-MFA

In real-world applications, such as data classification of image, gene expression, and web document, the dimension d of the vector samples is usually large, so the eigen-decomposition of $d \times d$ matrix XLX^T is still computational intensive. To reduce the computational demand, we present an efficient algorithm for performing F-MFA via QR-decomposition [26][27].

Let $X=QR$ be the QR-decomposition of X, where $Q \in R^{d \times t}$ has orthonormal columns, i.e. $Q^T Q = I$, $R \in R^{t \times n}$ is an upper triangular matrix, and $t = rank(X)$ is the rank of X. Obviously, $t < d$.

Let U be a matrix whose columns are eigenvectors of $RLR^T \in R^{t \times t}$ and Λ is the diagonal matrix such that Λ_{ii} is the eigenvalue associated to column i of U. Then we will have

$$RLR^T U = U\Lambda \tag{24}$$

Note that RLR^T is a real symmetric matrix, then U is an orthogonal matrix, i.e., $U^T U = I$.

Let $X=QR$ be the QR-decomposition of X, and $Q^T Q = I$. Thus,

$$RLR^T (Q^T Q) U = U\Lambda \tag{25}$$

Left multiply Q on both sides of the equation (25)

$$QRLR^T (Q^T Q) U = QU\Lambda \tag{26}$$

Equation (26) can be rewritten as

$$(QR) L (QR)^T (QU) = (QU)\Lambda \tag{27}$$

Substituting $X=QR$ in (27), then we have

$$XLX^T (QU) = (QU)\Lambda \tag{28}$$

Therefore, $V = QU$ is the matrix whose columns are eigenvectors of XLX^T and Λ is the diagonal matrix such that Λ_{ii} is the eigenvalue associated to column i of V. If U is composed of the r eigenvectors corresponding to the largest r eigenvalues of RLR^T, Then the optimal V can be computed as $V = QU$. Since $t < d$, the eigen-decomposition of RLR^T is more efficient than that of XLX^T.

As could be seen, the implementation of F-MFA does not involve an inverse matrix, and thus completely avoid the singularity problem. Therefore, it can be applied in small sample size problem directly and efficiently. Now, the algorithmic procedure of F-MFA is formally summarized in Algorithm 1.

Algorithm 1 F-MFA For Dimensionality Reduction

Step 1. Compute the matrix W. Given training samples, compute W according to (9), (12) and (15).

Step 2. Compute the matrix L in (19).

Step 3. QR-decomposition. Decompose the data matrix X as $X=QR$.

Step 4. Eigenvalue decomposition of RLR^T. Let $u_1, u_2, \cdots, u_r$ is eigenvectors of RLR^T associated with the largest eigenvalue and denote $U = [u_1, u_2, \cdots, u_r]$.

Step 5. Compute projection matrix. The optimal projection matrix is given by $V=QU$.

Step 6. Obtain low-dimensional embeddings according to (23).

The computational cost of F-MFA is analyzed as follows. The first part of F-MFA consists of constructing the weight matrices. This scales as $O(n^2)$. Its second part requires the OR-decomposition for X whose time complexity is $O(t^2 n)$. At last, time complexity of the eigen-decomposition is $O(t^3)$. As a result, the total time complexity of the fast algorithm is $O(t^3+t^2 n+n^2)$.

V. Experiments

In this section, we will evaluate the performance of the proposed F-MFA method on four benchmark face data sets: Yale, YaleB, FERET and GeorgiaTech.

A. Data Set Descriptions

We summarize the five data sets that we will use in our experiments in Table I. Some sample images of one individual after preprocessing of the four databases are shown in Fig. 3, Fig. 4, Fig. 5 and Fig. 6. The detail is briefly summarized as follows.

TABLE I
STATISTICS OF DATA SETS

Dataset	*d*	*n*	*c*
Yale	32×32	165	15
YaleB	32×32	2414	38
FERET	32×32	1400	200
Geogia Tech	32×32	750	50

Yale [1] data set is constructed at the Yale Center for Computational Vision and Control. It contains 165 images of 15 individuals (each person providing 11 different images) under variation in facial expressions, lighting conditions, and with/without glasses. In our experiments, each image is manually cropped and resized to 32×32 pixels.

YaleB[2] face data set has 38 individuals and around 64 near frontal images under different illuminations per individual. The images of the cropped version contain illumination variations and facial expression variations. The size of each cropped image in all the experiments is 32×32 pixels, with 256 gray levels per pixel.

FERET[3] face data set consists of 14051 gray scale images of human heads with views ranging from frontal to left and right profiles. It contains more than 1000 subjects. We select a subset of FERET database, which includes 1400 images of 200 distinct subjects; each subject has seven images. The subset involves variations in facial expression, illumination, and pose. In our experiment, the facial portion of each original image is cropped automatically based on the location of the eyes and resized to 40×40 pixels.

Georgia Tech [4] face data set contains images of 50 individuals taken in two or three sessions at different times. Each individual in the database is represented by 15 color JPEG images. The pictures show frontal and/or tilted faces with different facial expressions, lighting conditions and scale. Each image was manually grayed, cropped and resized to 50×36 pixels.

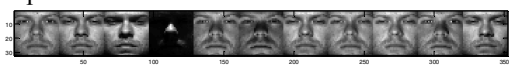


Fig.3. Sample face images from the Yale database.

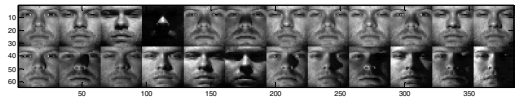


Fig.4. Sample face images from the YaleB database.

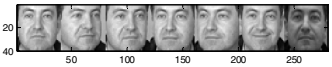


Fig.5. Sample face images from the FERET database.

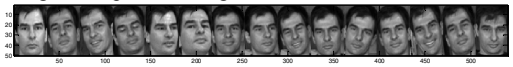


Fig.6. Sample face images from the Georgia Tech database.

B. Experimental Setup

We will compare the classification performance of our method (F-MFA) with other state-of-the-art methods, including Principle Component Analysis (PCA), Linear Discriminant Analysis (LDA), Locality Preserving Projection (LPP), Unsupervised Discriminant Projection (UDP), Marginal Fisher Analysis (MFA) and Stable Orthogonal Local Discriminant Embedding (SOLDE). Note that, all of these approaches involve PCA as a preprocessing step when tackle with singularity problem. In the following experiments, we keep 95% energy of data samples.

Before we do any classification, for each data set, we normalize their features first, making all the values in the range [-1,1]. For each dataset, l (l =3, 4, 5) images of each person are randomly selected as training samples, while the corresponding remain ones of each class are used for testing. For a given l, twenty random partitions were obtained for each data set, and average classification accuracy rate and standard deviation were reported.

Usually, the parameters can be empirically selected in all experiments. To be specific, we sampled several values of parameters and chose the values with the best performance for all approaches. In our experiments, we set the neighborhood parameter k=l-1, k_1=2, k_2=10 and regularization parameters α, β in Table II.

TABLE II
REGULARIZATION PARAMETERS SETTING

Dataset	α	β
Yale	0.001	0.003
YaleB	0.001	0.005
FERET	0.5	0.009
Geogia Tech	0.9	0.007

With respect to pattern discrimination, it is quite reasonable to suppose that the different samples have different contributions to classification. The greater the contributions of the samples are, the more significance for classification they have. We take into account the local scaling regulator of a data to dynamically adjust adjacent weights between pairs of neighbors, so as to well represent the classification contribution of each sample. In our experiments, the parameter t sets as follows [28]

$$t = \frac{1}{k^2}\sum_{j=1}^{k}\left\|x_i - x_j\right\|^2$$

In short, the recognition process has three steps. First, the linear projection matrix V is calculated from the training set; then the new testing face image vectors and all training vectors are projected into r-dimensional subspace; finally, the labels of testing image vectors are identified by nearest neighbor classifier.

C. Classification Results Comparisons

In general, the recognition rate varies with the dimensionality of the face subspace. The best average performance (best rate in the average curve) obtained by the seven dimensionality reduction algorithms as well as the corresponding standard variation and optimal dimensionality (at which the maximum average recognition rate has been reported) on the Yale, YaleB, FERET and Georgia Tech face

[1] http://see.xidian.edu.cn/vipsl/database_Face.html
[2] http://www.cad.zju.edu.cn/home/dengcai/Data/FaceData.html
[3] http://www.itl.nist.gov/iad/humanid/feret/
[4] http://www.anefian.com/research/face_reco.htm

data sets is summarized in Table III – VI respectively.

From both tables and figures, we can see that our proposed methods consistently beat the other six methods on all the data sets. These experiments reveal a number of interesting points.

(1) The top recognition accuracy of F-MFA approach outperforms SOLDE. This is probably because that SOLDE applies PCA preprocessing to tackle with singularity problem. However, some discriminant information will be eliminated in such preprocessing stage. F-MFA formulates the objectives in trace difference criterion which can effectively avoid singularity problem.

(2) As can be seen, F-MFA outperforms PCA, LDA, LPP, UDP and MFA. This is probably because that these algorithms capture only the similarity and ignore the diversity of faces in local neighborhoods, which will lead to unstable intra-class compact representation and degrade the generalization abilities of these algorithms. Differently, F-MFA preserves both the similarity and diversity among nearby faces.

(3) F-MFA and SOLDE performs better than other algorithms. Results on a variety of data sets have shown that F-MFA and SOLDE more stable than others. It indicates that orthogonality of projection and the local geometry is very important in characterizing the intrinsic geometry of faces and improving the discriminant and generalization power of the algorithm.

TABLE III
BEST AVERAGE RECOGNITION ACCURACY CORRESPONDING STANDARD DEVIATION (IN PERCENT) COMPARISON ON YALE FACE SET. THE NUMBER APPEARING IN PARENTHESIS CORRESPONDING TO THE OPTIMAL DIMENSIONALITY OF THE PROJECTED SUBSPACE

Method	3 Train	4 Train	5 Train
PCA	55.33±4.21(44)	55.76±3.11(28)	59.50±2.46(30)
LDA	62.17±5.00(14)	71.76±5.17(14)	76.50±3.89(14)
LPP	40.33±5.40(10)	48.62±5.14(13)	51.11±5.60(17)
UDP	49.63±4.27(39)	47.52±4.90(50)	41.44±8.16(49)
MFA	66.04±4.00(16)	74.24±3.77(17)	77.06±3.25(17)
SOLDE	59.00±4.80(15)	67.00±3.94(14)	70.83±3.08(15)
F-MFA	**69.21±4.08(20)**	**76.52±3.06(23)**	**80.72±3.52(18)**

TABLE IV
BEST AVERAGE RECOGNITION ACCURACY CORRESPONDING STANDARD DEVIATION (IN PERCENT) COMPARISON ON YALEB FACE SET. THE NUMBER APPEARING IN PARENTHESIS CORRESPONDING TO THE OPTIMAL DIMENSIONALITY OF THE PROJECTED SUBSPACE.

Method	3 Train	4 Train	5 Train
PCA	22.67±1.50(50)	26.21±1.32(50)	29.28±1.38(50)
LDA	53.37±2.04(37)	59.93±2.43(37)	65.46±1.62(37)
LPP	33.44±2.63(32)	42.87±1.92(42)	49.94±1.57(50)
UDP	45.67±1.94(50)	49.79±2.49(50)	50.88±1.36(50)
MFA	54.81±2.27(46)	60.40±2.20(49)	66.35±1.96(46)
SOLDE	45.24±3.11(50)	55.66±2.02(50)	62.65±1.21(50)
F-MFA	**58.26±2.09(49)**	**64.65±2.65(50)**	**70.23±1.67(42)**

TABLE V
BEST AVERAGE RECOGNITION ACCURACY CORRESPONDING STANDARD DEVIATION (IN PERCENT) COMPARISON ON FERET FACE SET. THE NUMBER APPEARING IN PARENTHESIS CORRESPONDING TO THE OPTIMAL DIMENSIONALITY OF THE PROJECTED SUBSPACE.

Method	3 Train	4 Train	5 Train
PCA	31.87±1.20(50)	36.27±1.59(50)	40.86±2.14(50)
LDA	37.35±1.78(50)	34.91±1.59(14)	31.51±1.44(50)
LPP	26.36±1.67(50)	30.03±2.01(50)	33.18±2.57(50)
UDP	7.09±0.98(50)	8.34±1.08(50)	8.98±1.42(50)
MFA	43.64±1.33(49)	48.16±2.00(50)	61.10±1.78(44)
SOLDE	67.44±1.77(22)	73.97±1.67(23)	78.21±1.52(32)
F-MFA	**86.75±1.24(28)**	**89.46±0.73(40)**	**90.09±0.75(41)**

TABLE VI
BEST AVERAGE RECOGNITION ACCURACY CORRESPONDING STANDARD DEVIATION (IN PERCENT) COMPARISON ON GEORGIA TECH FACE SET. THE NUMBER APPEARING IN PARENTHESIS CORRESPONDING TO THE OPTIMAL DIMENSIONALITY OF THE PROJECTED SUBSPACE.

Method	3 Train	4 Train	5 Train
PCA	63.14±1.70(33)	67.98±1.72(38)	71.56±1.94(50)
LDA	48.63±2.44(47)	53.76±2.74(49)	54.90±1.68(49)
LPP	45.16±8.93(15)	58.31±4.86(15)	63.02±4.49(17)
UDP	25.26±3.12(45)	28.63±2.30(50)	30.84±1.92(45)
MFA	59.17±2.70(50)	60.71±2.06(50)	58.77±1.91(50)
SOLDE	66.96±2.87(19)	71.25±2.21(22)	73.22±2.07(22)
F-MFA	72.18±2.53(49)	78.96±1.45(37)	83.48±1.60(36)

(4) It would be interesting to note that, The LPP and UDP method performs the worst in almost every case. This because that both of them are unsupervised methods which does not well encode the discriminating information of data and when there are only little training samples for each subject, the manifold structure cannot be characterized by local neighborhoods correctly.

VI. CONCLUSION

In this paper, we have exploited the local variation of data to characterize intrinsic geometric structure of data. In order to obtain an orthogonal projection matrix, we formulated the linear dimensionality reduction problem as a regularized difference criterion which can be solved by standard eigenvalue decomposition, and it effectively circumvents the singularity problem. For high-dimensional data set, the computational costs can be further alleviated by QR decomposition of data matrix X. In fact, F-MFA gives a flexible and efficient method to enhance the performance of MFA. Extensive experiments results of face recognition demonstrate the effectiveness of our method.

One possible extension of our work is to perform F-MFA in the reproducing kernel Hilbert space induced by a nonlinear function. The performance of kernel-based F-MFA needs to be further investigated. Another question is how to choose the regularization parameter theoretically. These works will be discussed in further research.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their constructive comments and suggestions.

REFERENCES

[1] S. Roweis and L. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no.5500, pp. 2323–2326, 2000.

[2] M. Belkin and P. Niyogi. "Laplacian eigenmaps and spectral techniques for embedding and clustering," In *Advances in Neural Information Processing Systems* 14, MIT Press, Cambridge, MA. pp. 585–591, 2001.

[3] D.L. Donoho and C. Grimes. Hessian Eigenmaps, "New Locally Linear Embedding Techniques for High-dimensional Data," *Proc. Nat'l Academy of Sciences* USA, vol. 100, no. 10, pp. 5591-5596, 2003.

[4] J. Tenenbaum, V. de Silva, and J. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol.290, no. 5500, pp. 2319–2323, 2000.

[5] Weinberger, K. Q., and Saul, L. K., "Unsupervised learning of image manifolds by semidefinite programming," In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pp. 988–995, 2004.

[6] M. Brand, "Charting a manifold," In *Advances in Neural Information Processing Systems* 15. Cambridge, MA: MIT Press, pp. 961–968, 2003.

[7] Z. Zhang and H. Zha,"Principal manifolds and nonlinear dimensionality reduction via tangent space alignment," *SIAM J. Sci. Comput*., vol. 26, no. 1, pp. 313–338, 2004.

[8] B. Schölkopf, A. Smola, and K. R.Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comput.*, vol. 10, pp. 1299–1319, 1998.

[9] G. Baudat and F. Anouar, "Generalized discriminant analysis using a kernel approach," *Neural Comput.*, vol. 12, pp. 2385–2404, 2000.

[10] I. T. Jolliffe, *Principal Component Analysis*. New York: Springer -Verlag. 1986.

[11] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol.3, no.1, pp. 71–86, 1991.

[12] R. A. Fisher, "The statistical utilization of multiple measurements," *Ann.Eugenics*, vol. 8, pp. 376–386, 1938.

[13] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs.Fisherfaces: Recognition using class specific linear projection," IEEE *Trans. Pattern Anal. Mach. Intell*., vol. 19, no. 7, pp. 711–720, 1997.

[14] Shuicheng Yan, Dong Xu, Benyu Zhang, and Hong-Jiang Zhang, "Graph embedding: A general framework for dimensionality reduction," In *Proc. Internal Conference on Computer Vision and Pattern Recognition*, 2005.

[15] Xiaofei He and Partha Niyogi, "Locality preserving projections," In *Advances in Neural Information Processing Systems* 16. MIT Press, Cambridge,MA, 2003.

[16] Yang J, Zhang D, Yang J, et al, "Globally maximizing, locally minimizing: unsupervised discriminant projection with applications to face and palm biometrics," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol.29, no.4, pp. 650-664, 2007.

[17] Hwann-Tzong Chen, Huang-Wei Chang,and Tyng-Luh Liu, "Local discriminant embedding and its variants," In Proc. *Internal Conference on Computer Vision and Pattern Recognition*, 2005.

[18] C. Hou, C. Zhang, Y. Wu, and Y. Jiao, "Stable local dimensionality reduction approaches"，*Pattern Recognition*, vol. 42, pp. 2054-2066, 2009.

[19] Q Gao, J Ma, H Zhang, X Gao, Y Liu, "Stable Orthogonal Local Discriminant Embedding for Linear Dimensionality Reduction," *IEEE transactions on image processing*. Vol. 22, no.7, pp. 2521-2531, 2013.

[20] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. New York: Academic, pp. 31, 34, 39–40, 220-221. 1990.

[21] Fan R. K. Chung, "Spectral Graph Theory," volume 92 of *Regional Conference Series in Mathematics*.AMS, 1997.

[22] H. Li, T. Jiang, and K. Zhang, "Efficient and robust feature extraction by maximum margin criterion," *IEEE Trans. Neural Netw.*, vol. 17, no. 1, pp. 157–165, 2006.

[23] D. Cai, X. He, K. Zhou, J. Han, and H. Bao, "Locality sensitive discriminant analysis," In *Proc. Int. Joint Conf. Artif. Intell.*, pp. 708–713, 2007.

[24] K. Z. Mao, "Fast orthogonal forward selection algorithm for feature subset selection," *IEEE Trans. Neural Netw.*, vol. 13, no. 5, pp. 1218–1224, 2002.

[25] Gao Q, Gao F, Zhang H, et al, "Two-dimensional Maximum Local Variation based on Image Euclidean Distance for Face Recognition," *IEEE Trans Image Process*. Vol. 22, no.10, pp. 3807-3817, 2013. doi: 10.1109/TIP.2013.2262286.

[26] Jieping Ye, Qi Li, Hui Xiong, Haesun Park, RaviJanardan, and Vipin Kumar, "Idr/qr: an incremental dimension reduction algorithm via qr decomposition," in *KDD*, pp. 364-373, 2004.

[27] Wang H, Chen S, Hu Z, et al, "Locality-preserved maximum information projection," *Neural Networks, IEEE Transactions on*, vol.19, no.4, pp. 571-585, 2008.

[28] Gou J, Yi Z., "Locality-Based Discriminant Neighborhood Embedding," *The Computer Journal*. published online: September 7, 2012. doi:10.1093/comjnl/bxs113.