Neural Comput & Applic (2012) 21:269–279
DOI 10.1007/s00521-011-0589-3

ORIGINAL ARTICLE

Robust large margin discriminant tangent analysis for face recognition

Nanhai Yang · Ran He · Wei-Shi Zheng · Xiukun Wang

Received: 4 December 2009 / Accepted: 9 April 2011 / Published online: 15 May 2011

Abstract Fisher's Linear Discriminant Analysis (LDA) has been recognized as a powerful technique for face recognition. However, it could be stranded in the non-Gaussian case. Nonparametric discriminant analysis (NDA) is a typical algorithm that extends LDA from Gaussian case to non-Gaussian case. However, NDA suffers from outliers and unbalance problems, which cause a biased estimation of the extra-class scatter information. To address these two problems, we propose a robust large margin discriminant tangent analysis method. A tangent subspace-based algorithm is first proposed to learn a subspace from a set of intra-class and extra-class samples which are distributed in a balanced way on the local manifold patch near each sample point, so that samples from the same class are clustered as close as possible and samples from different classes will be separated far away from the tangent center. Then each subspace is aligned to a global coordinate by tangent alignment. Finally, an outlier detection technique is further proposed to learn a more accurate decision boundary. Extensive experiments on challenging face recognition data set demonstrate the effectiveness and efficiency of the proposed method for face recognition. Compared to other nonparametric methods, the proposed one is more robust to outliers.

Keywords Nonparametric discriminant analysis · Linear discriminant analysis · Tangent distance · Face recognition

N. Yang · X. Wang
School of Software Technology,
Dalian University of Technology,
116620 Dalian, China
e-mail: nanhai@dlut.edu.cn

X. Wang
e-mail: jsjwxk@dlut.edu.cn

R. He (✉)
National Laboratory of Pattern Recognition,
Institute of Automation Chinese Academy of Sciences,
100190 Beijing, China
e-mail: rhe@nlpr.ia.ac.cn

W.-S. Zheng
School of Information Science and Technology,
Sun Yat-sen University, 510275 Guangzhou, China
e-mail: wszheng@ieee.org

1 Introduction

Fisher's Linear Discriminant Analysis (LDA) [1] is one of the most popular feature dimension reduction techniques in computer vision and machine learning [13, 14]. It maximizes the ratio between extra-class variation and intra-class variation. Despite its wide use, LDA assumes that the data distribution of each class is Gaussian. Thus, this is certainly hard to make LDA adapted to data under non-Gaussian distribution. Nonparametric estimation is one of the most popular algorithms for solving this problem. Nonparametric LDA (NDA) [3] has therefore been developed. The main difference between NDA and LDA is that NDA introduces a nonparametric extra-class scatter matrix, which computes the extra-class variation of data along the decision boundary.

Recently, NDA has received more and more attentions and has many variants. The traditional NDA is for the two-class classification problem. Then it was extended to the multi-class case [15]. In [2], the extra-class scatter matrix is constructed by a dataset near the decision boundary that is learned by support vector machine (SVM). In [32], a marginal Fisher analysis (MFA) method is proposed, and it

calculates the intra-class and extra-class scatter matrices in a local manifold patch. In [23], a nonparametric margin maximum criterion (NMMC) is discussed. The intra-class and extra-class scatter matrices are calculated by the furthest intra-class neighbor and the nearest extra-class neighbor respectively. In [17], a method is presented to determine the optimal dimensionality for discriminant analysis. In [18], the intra-class and extra-class scatter matrices are only defined on pair-wise points, which are neighboring. In [9], an adaptive nonparametric discriminant analysis is proposed by defining neighbors of a sample using a maximal intra-class distance. In [40], the intra-class and extra-class scatter matrices are constructed by Laplacian matrices, and a linear Laplacian discriminant algorithm is proposed for feature dimension reduction. In [35, 39], local coordinate alignment technique is used to learn a global embedding for dimension reduction. Then Zhang et al. [36, 37] proposed a unified patch alignment framework for subspace learning.

Although many nonparametric LDA models have been developed, the NDA-based methods always suffer from the outlier problem, which is not well addressed yet to our knowledge. It is always assumed in NDA that a decision boundary between different classes exists, so that samples near the decision boundary can be used to estimate the extra-class information. However, outliers may bias the estimation of data distribution, and such kind of boundary may be estimated inaccurately.

Moreover, NDA would always suffer from the unbalance problem and this is also not explored by existing NDA methods. In this paper, we analyze NDA from an approximate viewpoint of tangent subspace [5]. From the view point of tangent subspace, we find that NDA has a close connection with tangent subspace learning and can be treated as a special case of tangent analysis. In NDA, the number of intra-class samples is larger than that of extra-class samples in a tangent space of a sample point, so that the intra-class information is excessively emphasized. Such unbalance problem as well as outliers would make the NDA-based methods perform worse than LDA.

In this paper, we mainly address these two problems in NDA and then develop a novel discriminant tangent subspace method for robust face recognition. The proposed method makes samples of the same class clustered as close as possible and samples of different classes separated far away from the tangent center. It naturally integrates the tangent subspace techniques due to the close connection between NDA and tangent subspace methods and also benefits from the advantages of tangent approximation and tangent alignment. The intra and extra samples are balanced in a tangent space with respect to any sample point. In addition, we propose a simple yet efficient outlier detection technique. Outliers are filtered out according to the weights computed in the tangent subspace. Integrating this technique into the proposed nonparametric method can improve the robustness of the algorithm to local noise.

The rest of paper is organized as follows. We start our work with a discussion of NDA from an approximate perspective of tangent subspace learning and patch alignment framework in Sect. 2. After that, we present the new method in Sect. 3. In Sect. 4, we systemically evaluate the proposed method on five face recognition datasets. Finally, we conclude the paper in Sect. 5.

2 NDA and its problems

In this section, we first briefly review NDA and its differences from LDA. Then we investigate the outlier problem and unbalance problem by exploring its connection to tangent subspace analysis [11, 25].

2.1 LDA and NDA

Linear Discriminant Analysis has been widely used in computer vision and pattern recognition. It can be viewed as a special case of graph embedding. Given the training data $X = [x_1, \ldots, x_n] \in \Re^{d\times n}$ where $x_1, \ldots, x_n$ are drawn from classes $C_1, \ldots, C_L$, LDA learns a projection matrix for maximization of the ratio between extra-class variance and intra-class variance as follows:

$$U^* = \arg\max_U \frac{tr(U^T S_B U)}{tr(U^T S_W U)} \tag{1}$$

where $U \in \Re^{d\times m}$ $(m<d)$ is an orthonormal matrix and $tr(\cdot)$ is the trace operator. S_B and S_W are extra-class scatter matrix and intra-class scatter matrix respectively, which are specified as follows:

$$\begin{aligned} S_W &= \sum_{k=1}^{L}\sum_{j\in C_k}(x_j-\mu_k)(x_j-\mu_k)^T \\ &= \sum_{k=1}^{L}\sum_{i,j\in C_k} w_{ij}^I (x_j-x_i)(x_j-x_i)^T \end{aligned} \tag{2}$$

$$\begin{aligned} S_B &= \sum_{k=1}^{L} n_k(\mu_k-\mu)(\mu_k-\mu)^T \\ &= \sum_{k=1}^{L}\sum_{k'=1,k'\neq k}^{L}\sum_{i\in C_k}\sum_{j\in C_{k'}} w_{ij}^E (x_j-x_i)(x_j-x_i)^T \end{aligned} \tag{3}$$

where w_{ij}^E and w_{ij}^I are weights (constant), μ_k denotes the sample mean of class C_k, μ denotes the mean of all samples, and n_k denotes the number of samples in class C_k. Therefore, (1) can be written as follows:

$$U^* = \arg\max_U \frac{\sum_{k=1}^{L}\sum_{i\in C_k} tr(U^T(\sum_{k'\neq k}\sum_{j\in C_{k'}} w_{ij}^E(x_j - x_i)(x_j - x_i)^T)U)}{\sum_{k=1}^{L}\sum_{i\in C_k} tr(U^T(\sum_{j\in C_k} w_{ij}^I(x_j - x_i)(x_j - x_i)^T)U)} \tag{4}$$

The main difference between NDA and LDA lies in the definition of the extra-class scatter matrix. In [3], the extra-class scatter matrix of NDA is defined as,

$$S_B^N = \sum_{i=1}^{L}\sum_{j\in C_i} w_{ij}(x_j - \mu(x_j))(x_j - \mu(x_j))^T \tag{5}$$

where $\mu(x_j)$ is the extra-class local k-NN mean defined by

$$\mu(x_j) = \frac{1}{K}\sum_{k=1}^{K} NN_k(x_j, C_{x_j}) \tag{6}$$

where c_{x_j} is the class label of sample x_j, $NN_k(x_j, c_{x_j})$ is the k-th extra-class nearest neighbor to x_j. Using the Euclidean metric, w_{ij} is value of the weighting function and defined as,

$$w_{ij} = \frac{\min\{||x_j - \mu_i||^\beta, ||x_j - \mu(x_j)||^\beta\}}{||x_j - \mu_i||^\beta + ||x_j - \mu(x_j)||^\beta} \tag{7}$$

where β is an integral power [2] and $\|\cdot\|$ is the Euclidean norm. The purpose of the w_{ij} is to de-emphasize the contribution of samples that are far from the decision boundary. Li and Ito [15] extends computation of S_B^N from two-class case to multi-class case.

Since LDA maximizes the mean value of the Kullback–Leibler (KL) divergences between different classes, Tao et al. [27] proposed a geometric mean method for subspace selection, which maximizes the geometric mean of the KL divergences or the normalized KL divergences. The geometric mean method can in principle be faster than several nonparametric methods [27].

2.2 Patch alignment framework

Patch alignment framework [36, 37] is a unified framework for subspace learning. It divides the subspace learning into the patch optimization step and whole alignment step. Most subspace learning methods, such as PCA, LDA, and Laplacian Eigenmaps, can be formulated in this framework.

Zhang et al. [34] developed linear local tangent space alignment approach for face recognition, which uses the tangent space in the neighborhood of a data point to represent the local geometry. Zhao et al. [39, 40] developed Laplacian PCA and linear Laplacian discrimination respectively by constructing a Laplacian matrix in a local patch. Zhang et al. [35] proposed a local coordinates alignment (LCA) technique for manifold learning. LCA obtains local coordinates as representations of local neighborhood relations on a patch. Based on the orthogonal neighborhood-preserving projection, Zhang et al. [38] proposed discriminative orthogonal neighborhood-preserving projections by combining both intra-class and inter-class geometries. Zhang et al. [36] proposed a discriminative locality alignment (DLA) for supervised subspace learning. In DCA, one local patch is built by one given sample and its neighbors, which include the samples from not only a same class but also different classes. PCA is used as a preprocessing step for DCA to reduce noise. Based on DCA and the optimal sparse solution of a manifold learning, Zhou et al. [28] proposed a manifold elastic net approach for sparse dimension reduction. To deal with the undersampled problem in face recognition and human gait recognition, Song and Tao [26] develop a discrminative geometry preserving projections method. Yang [33] studied the problem of aligning overlapping locally scaled patches for dimension reduction.

Although patch alignment-based methods indeed improve the performance of subspace learning in terms of recognition rate, there are few methods to deal with noise and outliers especially in the challenging face recognition problem. PCA is often used to reduce small noise [36], but it fails to deal with large noise and outliers [6, 7, 30].

2.3 The problems

In manifold learning, we often assume that a local manifold patch is isotropic to an Euclidean space. Let $x \in \Re^d$ be a pattern and $t(x, \alpha)$ be the pattern transformed from x by a parameter α. Tangent approximation method, often called the invariant pattern classification method [5, 11, 20, 25], learns an approximate tangent subspace to model the local variation of a manifold. It assumes that $t(x, \alpha)$ has the same class membership as x. The set of transformed patterns, $M_x = \{t(x, \alpha)\}$, now form a manifold in pattern space. The transformed pattern $t(x, \alpha)$ can be approximated by a Taylor expansion at $\alpha = 0$.

$$t(x, \alpha) = x + \alpha T + o(\alpha^2) \tag{8}$$

where T is a tangent vector, namely $T = \partial(x, \alpha)/\partial\alpha$. Tangent vector approximates the manifold M_x by the first-order Taylor expansion. In practice, the tangent vector T is often approximated by x's neighborhood points [5]. From this view point, the tangent approximation on point x can be viewed as the local patch optimization in patch alignment framework.

From the viewpoint of tangent subspace, if we treat $x_j - x_i$ in NDA as the tangent vector of x_i, it is easy to find that NDA utilizes the nearby intra-tangent vectors and extra-tangent vectors to construct intra-class matrix and extra-class matrix respectively. So, NDA can be interpreted as learning an approximate tangent subspace

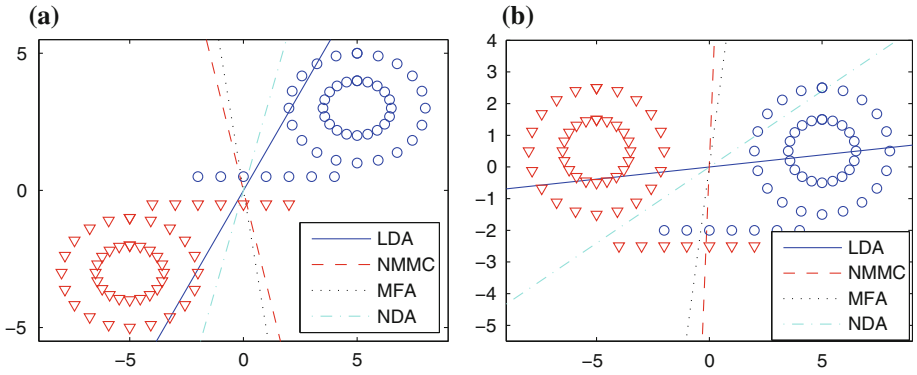

Fig. 1 Toy problems: LDA and NDAs (see text for details). **a** Optimal projections of different methods calculated on a synthetic dataset where two classes are separated. **b** Optimal projections of different methods calculated on a synthetic dataset where two classes are not separated

by a global projection matrix, where intra-class samples are close to tangent center and extra-class samples are far away from it. However, it can be found that there are many intra-class samples and only one extra-class sample (μ (x_j)) in tangent space. This implies that NDA excessively focuses on intra-class variation and results in an unbalance problem.

Another important problem in NDA is caused by outliers, which certainly have negative effect on learning the discriminant projection. In NDA, the decision boundary has to be known first, so that samples near the decision boundary can be used to construct the extra-class scatter matrix. The most common way is to assume that the samples adjacent to different classes are also close to the decision boundary. However, this assumption is not appropriate especially when outliers exist. Since outliers may be close to a class but far away from decision boundary, the decision boundary estimated by them will be bias.

In order to further specify the outlier problem, a two-class toy example is shown in Fig. 1, where data of each class follows non-Gaussian distribution. Four optimal projections with respect to LDA, canonical NDA, MFA, and NMMC are shown respectively. NDA and its other variants such as MFA and NMMC select the same samples to construct the extra-class scatter matrix in Fig. 1a, b. The results in Fig. 1b clearly demonstrate that NDA, MFA, and NMMC fail to find the optimal projection direction. This example illustrates that outliers make NDA and its variations estimate a bias decision boundary. To address the unbalance problem and outlier problem in NDA, we will introduce a novel discriminant technique in the following section.

3 Large margin discriminant tangent subspace

This section proposes to learn a robust large margin discriminant tangent subspace (LMTS) method. In Sect. 3.1, inspired by the invariant pattern classification method [25], we discuss how to select balanced intra-class and extra-class samples to learn a set of discriminant features in a tangent space. In Sect. 3.2, inspired by the patch alignment framework [37], we also align the local tangent subspaces to a global subspace. After that, in Sect. 3.3, we further propose a simple yet efficient outlier detection technique to reduce noise.

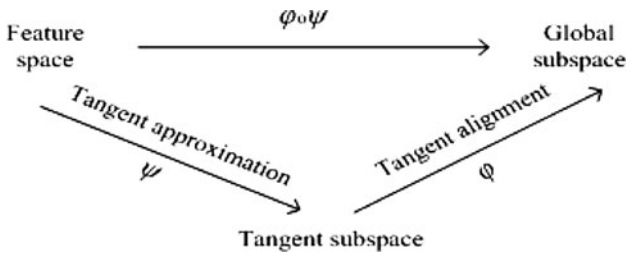

Fig. 2 A basic methodology of robust large margin discriminant tangent subspace. It learns a tangent approximation mapping and a tangent alignment mapping to project the original feature space to a global subspace

As an overview, Fig. 2 illustrates the basic methodology of the proposed method. A tangent subspace is first learned to model the local data variation on manifold and then tangent alignment is performed to get a global subspace. Similar to the patch alignment framework [36], we also harness the whole alignment to learn a global coordinate (or subspace). However, our tangent subspace belongs to the invariant pattern classification method [5, 25] and aims to learn the invariant local structure of a manifold in (8).

3.1 Tangent approximation

The tangent approximation assumes that a local variation of a sample point can be depicted by a linear approximation of its tangent space given by the first-order Taylor expansion. Using the tangent approximation technique

could make the modeling of data variation more accurate, and it also has been shown as a promising technique for research in other fields [19, 25].

Note that data variations can be derived from intra-class samples and extra-class samples [31]. Our goal is to learn a large margin tangent subspace for each sample point on a manifold. We expect intra-class samples cluster as compactly as possible while data of different classes scatter far away from each other. Suppose $x_{k,i}^E$ is the kth nearest extra-class sample of x_i and $x_{k,i}^I$ is the kth nearest intra-class sample of x_i. Let $y_{k,i}^E$ be extra-class tangent vector of $x_{k,i}^E$ and $y_{k,i}^I$ be intra-class vector of $x_{k,i}^I$ in the tangent subspace of sample x_i. We then propose the following optimization problem:

$$\max_{y_{1,i}^E,\ldots,y_{K_1,i}^E,y_{1,i}^I,\ldots,y_{K_2,i}^I} \left(\frac{\sum_{k=1}^{K_1} w_{ik}^E (y_{k,i}^E)^T y_{k,i}^E}{\sum_{k=1}^{K_2} w_{ik}^I (y_{k,i}^I)^T y_{k,i}^I} \right) \tag{9}$$

where w_{ik}^E and w_{ik}^I are the corresponding importance weights, and K_1 and K_2 are the number of intra and extra samples respectively. For simplicity, we set both w_{ik}^E and w_{ik}^I to 1 in the experiment. If there is a linear mapping $U_i^T = \psi$ from the tangent space and its subspace, then for some extra-class sample $x_{k,i}^E$ and intra-class sample $x_{k,i}^I$, we get

$$y_{k,i}^E = U_i^T (x_{k,i}^E - x_i) \quad \text{and} \quad y_{k,i}^I = U_i^T (x_{k,i}^I - x_i) \tag{10}$$

In the tangent space of x_i, we expect to learn a linear projection matrix U_i^T which is the solution of following criterion:

$$\max_{U_i} \frac{tr\left(U_i^T \left(\sum_{k=1}^{K_1} w_{ik}^E \left(x_{k,i}^E - x_i \right) \left(x_{k,i}^E - x_i \right)^T \right) U_i \right)}{tr\left(U_i^T \left(\sum_{k=1}^{K_2} w_{ik}^I \left(x_{k,i}^I - x_i \right) \left(x_{k,i}^I - x_i \right)^T \right) U_i \right)} \tag{11}$$

3.2 Tangent alignment

Aiming to learn a global coordinate system for all the tangent subspaces, we perform the tangent alignment technique to derive a global map $U^T = \phi \circ \psi$. We assume there exists a global coordinate $Y = [y_1, \ldots, y_n]$, a global map U so that

$$y_{k,i}^E = U^T (x_{k,i}^E - x_i) \quad \text{and} \quad y_{k,i}^I = U^T (x_{k,i}^I - x_i) \tag{12}$$

Then we have the following problem

$$\arg\max_{\{y_1,\ldots,y_n\}} \frac{tr\left(\sum_{i=1}^{N} \sum_{k=1}^{K_1} w_{ik}^E y_{k,i}^E \left(y_{k,i}^E \right)^T \right)}{tr\left(\sum_{i=1}^{N} \sum_{k=1}^{K_2} w_{ik}^I y_{k,i}^I \left(y_{k,i}^I \right)^T \right)} \tag{13}$$

It is straightforward to verify that

$$\begin{aligned}
&tr\left(\sum_{i=1}^{N} \sum_{k=1}^{K_1} w_{ik} y_{k,i}^E \left(y_{k,i}^E \right)^T \right) \\
&= tr\left(U^T \left(\sum_{i=1}^{N} \sum_{k=1}^{K_1} w_{ik} \left(x_{k,i}^E - x_i \right) \left(x_{k,i}^E - x_i \right)^T \right) U \right) \\
&= tr\left(U^T \left(\sum_{i=1}^{N} \sum_{j=1}^{K_1} w_{ik} \left(x_{k,i}^E \left(x_{k,i}^E \right)^T - 2 x_{k,i}^E x_i^T + x_i x_i^T \right) \right) U \right) \\
&= tr\left(U^T X \sum_{i=1}^{N} S_i^E W_i^E (S_i^E)^T X^T - 2X \sum_{I=1}^{N} S_i^E W_i^E X^T - X W_i^E X^T U \right) \\
&= tr\left(U^T X \left(\sum_{i=1}^{N} S_i^E W_i^E (S_i^E)^T - \sum_{i=1}^{N} S_i^E W_i^E + W_i^E \right) X^T U \right)
\end{aligned}$$

where S_i^E is the selection projection matrix [40] such that $Y_i^E = Y S_i^E$. Similarly, we can get

$$\begin{aligned}
&tr\left(\sum_{i=1}^{N} \sum_{k=1}^{K_2} w_{ik} y_{k,i}^I \left(y_{k,i}^I \right)^T \right) \\
&\quad = tr\left(U^T X \left(\sum_{i=1}^{N} S_i^I W_i^I (S_i^I)^T - \sum_{i=1}^{N} S_i^I W_i^I + W_i^I \right) X^T U \right)
\end{aligned}$$

where S_i^I is the selection projection matrix $Y_i^I = Y S_i^I$. Thus, we have the following maximization problem:

$$\arg\max_U \frac{tr(U^T X P^E X^T U)}{tr(U^T X P^I X^T U)} \tag{14}$$

where

$$P^E = \sum_{i=1}^{N} S_i^E W_i^E (S_i^E)^T - S_i^E W_i^E + W_i^E \tag{15}$$

$$P^I = \sum_{i=1}^{N} S_i^I W_i^I (S_i^I)^T - S_i^I W_i^I + W_i^I \tag{16}$$

For computation, the above optimization problem in (14) can be solved by using generalized eigen-decomposition as follows:

$$(X P^E X^T) U = (X P^I X^T) U \Lambda \tag{17}$$

where Λ is the diagonal matrix whose diagonal terms are eigen-values. Then U can be the eigenvectors corresponding to the first m largest eigen-values.

When feature dimension is high, $X L^I X^T$ tends to be singular such that discriminant analysis methods suffer the undersampled problem [27]. To deal with the undersampled problem, a regularizer can be further introduced in (17) and we have

$$(X P^E X^T) U = (X P^I X^T + \delta I) U \Lambda \tag{18}$$

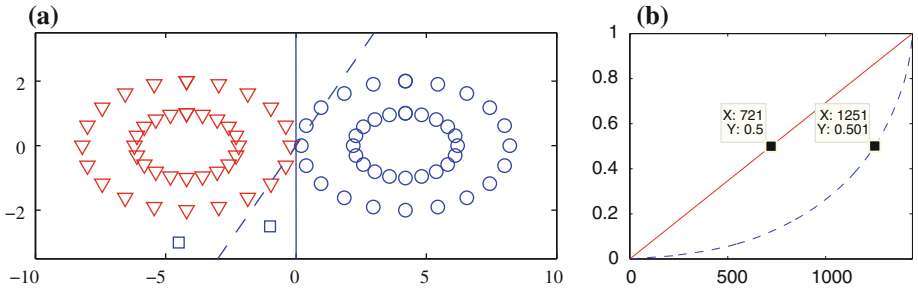

Fig. 3 **a** Example of decision boundaries in a two-class problem (*triangle* class and *ellipse* class). The two rectangle points are outliers belong to the ellipse class. The *solid line* is the decision boundary calculated by NDA regardless of those two outliers; the *dash line* is the decision boundary calculated by NDA when those two outliers are involved. **b** Accumulation curves of two weight distributions. The *solid line* presents an accumulation curve of the distribution of sample weights of ideal data set. The *dash line* presents the accumulation curve of the distribution of sample weight of FRGC data set. The sample weights are calculated by (18)

In the experiment, we empirically set $\delta = 0.01 \times (\frac{1}{N}\sum_{i=1}^{N} \max(x_i(j)))^2$, where $x_i(j)$ is jth entry of feature vector x_i.

3.3 Outlier detection

Lee and Landgrebe [10] introduced the decision boundary feature matrix for a direct computation of the intrinsic discriminant dimensionality. In NDA and its variations, the decision boundary is approximately calculated by intra-class and extra-class matrix. However, the estimation of extra-class matrix becomes bias when there are outliers. The outliers will have large effect on the data distribution and biases the data distribution estimation, which are major problems in robust learning [6–8].

Figure 3a shows a two-class toy problem (triangle class and ellipse class). The two rectangle points (outliers) belong to the ellipse class. It is clear that the outliers are closer to another class than the rest of the same class. If we remove the two outliers, we will learn a decision boundary (solid line in Fig. 3a) which can entirely separate the main parts of two classes. If we include the two outliers, we will obtain a biased decision boundary (dash line in Fig. 3a). It is clear that the outliers make NDA estimate a bias decision boundary. This inaccurate estimation is due to the inaccurate estimation of extra-class matrix.

For nonparametric methods, some samples near the desired decision boundary are selected to calculate the extra-class matrix. To describe the importance of a selected sample, we introduce the weight of a sample x_i as following:

$$\begin{aligned} \text{weight}(x_i) &= \sum_j \sigma(x_i, x_j) \quad \text{and} \\ \sigma(x_i, x_j) &= \begin{cases} w_{ij}^E & x_i \in NN_k(x_j, C_{x_j}) \\ 0 & \text{else} \end{cases} \end{aligned} \tag{19}$$

where $NN_k(x_j, c_{x_j})$ is the kth nearest neighbor of x_j and is not in C_{xj}. A large weight(x_i) shows that x_i is close to the decision boundary. Let us first consider an ideal case that samples in this set have nearly equal weights. The solid line in Fig. 3b shows the accumulation curve of weight in this case. Each point near the boundary receives equally importance. Thus, accumulation curve is nearly linear. However, the distribution of weight will be extremely different when there are outliers. The dash line in Fig. 3b shows an example of the accumulation distribution on FRGC dataset. Therefore, we find that some points whose weights are larger than the others may be outliers. This motivates us to find a strategy to alleviate the effect of outliers.

Inspired by soft-threshold [30] in robust face recognition, we suggest directly removing those x_i whose weights are larger than a threshold from the training set when constructing the extra-class matrix. Note that an outlier always gets a excessive high sample weight such that the estimated decision boundary is biased. Therefore, we assume that weights of samples nearby a decision boundary should be almost equal. To this end, if there is a x_i whose weight is larger than that of the others, it may be an outlier and it would be removed in our algorithm.

4 Experiments

In this section, the proposed method is extensively compared with other related methods on five commonly used face recognition data sets. Since the face recognition problem is an open set problem,[1] we divide the data set into training set, probe set, and gallery set [40]. A probe set contains one or more images from a set of individuals. Each person will have exactly one match in the gallery. The gallery may contain images from other individuals who are not in the probe population [22, 40]. Considering that the dimension of facial images is often high, we discuss standard discriminant analysis and regularized

[1] Open set domains assume that new classes may be encountered. In contrast, closed set domains assume that all classes of a domain have been known and can be used in training.

discriminant analysis in Sects. 4.2 and 4.3 respectively. All algorithms are implemented in MATLAB.

The number of intra-class nearest neighbors K_1 and extra-class nearest neighbors K_2 for all non-parameter methods are set to 4 and 8 respectively. The nearest neighbor classifier is adopted for classification. The number of features, extracted by all methods, ranges from 10 to 350 where the best accuracy is reported. The value of threshold for LMTS is set to the value whose corresponding probability value in accumulation curve is 0.7. Note that the goal of this experiment is to fairly compare the related nonparametric discriminant methods rather than achieve the highest face recognition accuracies on these data sets.

4.1 Datasets

4.1.1 FRGC dataset

The facial images are collected from a subset of FRGC version 2 [21] face database. There are uncontrolled 8,014 images of 466 subjects in the query set for the FRGC experiment 4. These still images contain the variations of illumination, expression, time, and blurring. However, there are only two facial images available for some persons. Thus, a subset is selected in our experiments. We take the first 10 facial images if the number of facial images is not less than 10. Then we get 3160 facial images of 316 subjects. The first 200 subjects are used as the training set. Then we take the first five facial images of each person in the last 116 subjects as the gallery set and the remaining five images as the probe set. Therefore, the set of persons for training is different from that for testing.

4.1.2 CMU PIE dataset

The CMU PIE database [24] contains more than 40,000 facial images of 68 subjects. We select a subset in our experiment which contains five near frontal poses (C27, C05, C29, C09, and C07) and illumination indexed by 03 and 11. So there are 10 images for each subject. The first 38 subjects are used as the training set, and the remaining 30 subjects are set as the gallery set and probe set, where we take the first five facial images of each person in the last 30 subjects as the gallery set and the remaining five images as the probe set. Therefore, the set of persons for training is different from that for probe. Histogram equilibrium was applied as the preprocessing step in PIE dataset.

4.1.3 FERET dataset

The facial recognition technology (FERET) database [22] contains 1,564 sets of images for a total of 14,126 images, including 1,199 individuals and 365 duplicate sets of images. We take the first 10 frontal facial images if the number of frontal facial images is not less than 10. Then we get 1,690 facial images of 169 subjects. The first 100 subjects are used as the training set. Then we take the first five facial images of each person in the last 69 subjects as the gallery set and the remaining five images as the probe set.

4.1.4 ORL dataset

The Cambridge (ORL) database[2] contains 40 distinct subjects, each subject having ten different images, taken at different times, varying the lighting, facial expressions, and facial details (glasses/no glasses). All the images are taken against a dark homogeneous background and the subjects are in upright, frontal position. The first 20 subjects are used as the training set, and the remaining 20 subjects are set as the gallery set and probe set, where we take the first five facial images of each subject in the last 20 subjects as the gallery set and the remaining five images as the probe set.

4.1.5 Extended Yale B dataset

The Extended Yale B database [4] consists of 2,414 frontal face images from 38 subjects under various lighting conditions. The first 20 subjects are used as the training set, and the remaining 18 subjects are set as the gallery set and probe set, where we take the first five facial images of each subject in the last 18 subjects as the gallery set and the remaining five images as the probe set.

For the first four data sets, the grayscale images were resized to resolution 64×64; and for the last Extended Yale B Dataset, the grayscale images were resized to resolution 96×84. Table 1 summarizes the details of the data sets used in the experiments.

4.2 Experiment I: discriminant analysis

To solve the singularity problem of the intra-class matrix of LDA, PCA is often used to reduce the high-dimensional image space. Hence, we compare different methods on PCA subspace. Wang and Tang [29] showed that the dimension of principal subspaces significantly affects the performance of recognition for the PCA plus LDA strategy. Based on their work, the optimal number of principal components is determined manually. We find that PCA performs the best when the dimension of feature vectors are 453 and 150 on FRGC and PIE respectively.

[2] http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html.

Table 1 Description of the data sets used in the experiments

	FRGC dataset			PIE dataset			Feret dataset			ORL dataset			YALE dataset		
	T	G	P	T	G	P	T	G	P	T	G	P	T	G	P
Number of subjects	200	116	116	38	30	30	100	69	69	20	20	20	20	18	18
Number of images	2,000	580	580	380	150	150	1,000	345	345	200	100	100	640	288	288

Character "T" represents training set; character "P" represents probe set; character "G" represents gallery set

Table 2 Face recognition rates (%) on PCA features

	PCA + LDA	PCA + NMMC	PCA + NDA	PCA + MFA	PCA + LMTS
FRGC (*T*6)	70.60	65.17	69.48	70.17	**71.65**
FRGC (*T*10)	77.19	73.45	75.34	77.24	**78.28**
FERET (*T*6)	57.68	50.43	55.94	51.88	**58.55**
FERET (*T*10)	60.87	54.78	59.42	58.55	**62.35**
PIE (*T*6)	71.03	76.05	75.86	74.48	**76.55**
PIE (*T*10)	86.21	88.97	92.41	95.17	**96.55**
ORL (*T*6)	96.00	96.00	**98.00**	96.00	**98.00**
ORL (*T*10)	97.00	98.00	**99.00**	**99.00**	**99.00**
YALE B (*T*20)	91.32	95.14	97.22	94.10	**97.57**
YALE B (*T*32)	98.26	98.96	99.31	98.96	**99.65**

The bold numbers are the highest recognition rates for each configuration

Table 2 tabulates the experimental results on five data sets. The "*Tn*" represents *n* facial images for each person that are randomly selected from the training set on that experiment. On FRGC data set, the recognition accuracy by only using PCA subspace is 51.03%. It is obvious that all compared discriminant methods can learn a discriminant subspace and further improve recognition accuracy. The recognition accuracy of different methods increases as the number of training set increases.

The FRGC and FERET data set seem to be more difficult than the other three data sets. The highest rates on FRGC and FERET are 78.28 and 62.35% respectively. This may be due to the fact that FRGC and FERET data set are the most challenging data sets. The facial images in the two data sets have large variations under uncontrolled environment so that NDA estimates a bias decision boundary. LMTS addresses this problem in NDA by using large margin strategy and robust method. We can see that LMTS obtains a notable improvement over NDA and achieves the best performance on the two data sets.

On PIE, ORL, and extended YALE B data set, all methods obtain high recognition rates. Since there are only 40 subjects in ORL data set and the facial images are nearly frontal images, NDA, MFA, and LMTS all achieve the highest recognition rate 99%. It seems that NDA and MFA can more efficiently use local structure on small data set than LDA so that they perform better than LDA. As expected, LMTS consistently outperforms the other methods.

4.3 Experiment II: regularized discriminant analysis

Although PCA is often used as a preprocessing step for discriminant analysis to reduce the dimension of image feature, regularization methods are often used in local feature (e.g., Gabor features [12]) -based face recognition. Regularized discriminant analysis method is often used to deal with undersampled problem [27] and to further improve recognition accuracy [16]. In this experiment, we evaluate different regularized methods for face recognition. More specifically, we filter each facial image and perform a Gabor filter on every six pixels. Thus, there are 99 Gabor values on an image by one Gabor filter. Four scales and four directions of Gabor are used. So the number of Gabor filters in a Gabor feature vector is $99 \times (4 \times 4) = 1{,}584$. The regularized parameter is set as indicated below (18).

Table 3 shows the comparison results of regularized discriminant methods. As expected, regularized LMTS achieves the best accuracy. Compared with RLDA, the improvement of RNDA is limited because RNDA calculates intra-class matrix in the same way as RLDA. RMFA and RLMTS perform better than RLDA and RNDA because they construct intra-class and extra-class matrix from the information in local area. By using Gabor

Table 3 Face recognition rates (%) of different regularized methods on Gabor features

	RLDA	RNMMC	RNDA	RMFA	RLMTS
FRGC (*T*6)	75.86	70.01	76.90	80.17	**81.16**
FRGC (*T*8)	77.24	70.62	77.45	80.12	**81.55**
FRGC (*T*10)	77.25	70.41	77.59	80.05	**81.55**
PIE (*T*6)	92.41	93.79	93.1	93.79	**94.48**
PIE (*T*8)	94.48	93.10	95.05	93.10	**95.17**
PIE (*T*10)	98.62	93.10	97.93	98.62	**99.31**

The bold numbers are the highest recognition rates for each configuration

features, RLMTS can consistently outperform other methods and further improve recognition rates.

4.4 Parameter selection

The threshold mentioned in Sect. 3.3 is an important parameter to remove noise from the decision boundary. Its value directly affects final recognition accuracy. For demonstration, Fig. 4a shows the face recognition rates when different values of the threshold parameter are set, where the experimental setting is the same as that of FRGC (*T*10) in Sect. 4.2. When the threshold ranges from 10 to

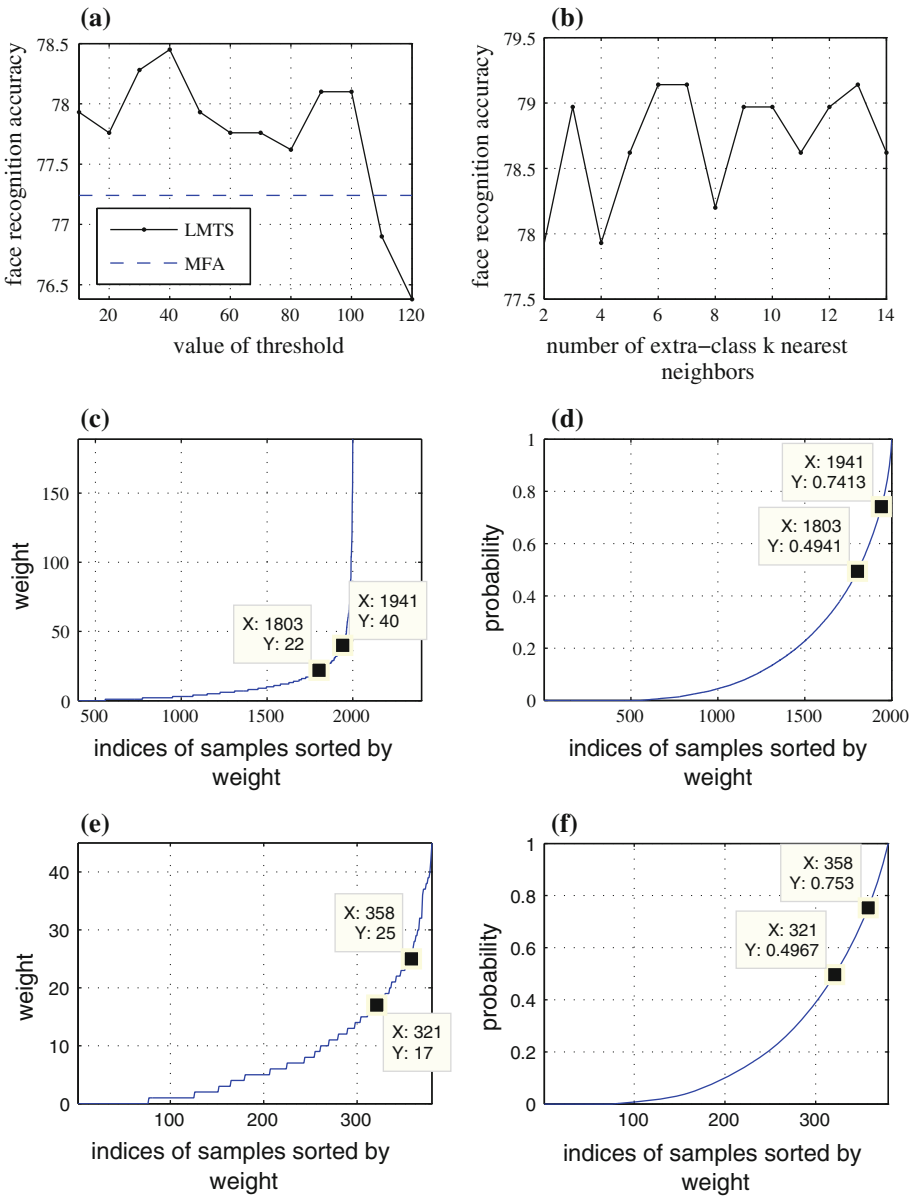

Fig. 4 Face recognition accuracy when different thresholds are set. **a** The accuracies of MFA and LMTS with different values of threshold in Sect. 3.3 on FRGC dataset; **b** the accuracy of LMTS whose number of extra-class k near neighbors ranges from 2 to 14 when number of k near intra-class neighbors equals 4; **c** the weights calculated by (19) on the same dataset of **a**; **d** the accumulation curve calculated by the weights in **c**; **e** the weights on PIE dataset; **f** the accumulation curve calculated by the weights in **e**

100, the accuracy of LMTS is always larger than that of MFA. But different values of the threshold will make LMTS perform differently. When the threshold ranges from 30 to 40 and from 90 to 100, LMTS can obtain higher accuracy, and when the threshold is larger than 100, recognition accuracy decreases because outliers make LMTS estimate a biased extra-class matrix. Thus, it is necessary to determinate an appropriate value of the threshold for particular application.

Figure 4c shows the sorted weights of all samples. We can find that the weights of some samples are significantly larger than the weights of others. When the value of the threshold is set to 40, there are nearly 39 images that should be removed. Looking at the accumulation curve in Fig. 4d, we also observe the same phenomenon. Given a new data set, we suggest to decide the threshold by the weight distribution and accumulation curve. Figure 4e, f further show the weight curve and accumulation curve on PIE database respectively.

The number of intra-class k nearest neighbors and extra-class k nearest neighbors are also important to recognition accuracy. Figure 4b depicts the accuracy of LMTS whose number of extra-class k near neighbor ranges from 2 to 14 and the number of intra-class k near neighbors is fixed to 4. For each extra-class k near neighbors, we select the weight corresponding to the point at 0.65 on the accumulation curve. The accuracy of LMTS varies between 78% and 79%. The average accuracy is 78.71%. The different number of extra-class k near neighbors results in different accuracy.

5 Conclusion

This paper addresses the outlier and unbalance problems in nonparametric discriminant analysis, which are less investigated in existing nonparametric methods. Based on tangent approximation and tangent alignment techniques, we propose a novel large margin discriminant tangent analysis for robust face recognition. An outlier detection technique is further suggested to construct a robust extra-class matrix such that the proposed method can learn a more robust subspace for learning. Extensive experiments on five face recognition data sets show the improvements of the proposed method as compared to related nonparametric discriminant methods and validate its usefulness for face recognition especially on challenging data sets. Our work also shows that there is a potential threshold parameter for LMTS to detect outlier. Future work is to study robust LMTS and its parameter selection from the view point of robust statistics [7, 30].

Acknowledgments This work was supported in part by Research Foundation for the Doctoral Program of the Ministry of Education of China (Grant No. 20100041120009), 985 Project in Sun Yat-sen University (Grant No. 35000-3181305), and NSF-Guangdong (Grant No. U0835005).

References

1. Belhumeur P, Hespanha J, Kriegman D (1997) Eigenfaces vs. fisherfaces: recognition using class specific linear projection. IEEE Trans Pattern Anal Mach Intell 19(1):711–720
2. Fransens R, Prins J, Gool L (2003) Svm-based nonparametric discriminant analysis, an application to face detection. In: IEEE international conference on computer vision
3. Fukunaga K (1990) Statistical pattern recognition. Academic Press, San Diego
4. Georghiades A, Belhumeur P, Kriegman D (2001) From few to many: illumination cone models for face recognition under variable lighting and pose. IEEE Trans Pattern Anal Mach Intell 23(6):643–660
5. He R, Ao M, Xiang S, Li SZ (2008) Nearest feature line: a tangent approximation. In: Chinese conference on pattern recognition
6. He R, Hu BG, Zheng WS, Kong XW (2011a) Robust principal component analysis based on maximum correntropy criterion. IEEE Trans Image Process (preprint)
7. He R, Zheng WS, Hu BG (2011b) Maximum correntropy criterion for robust face recognition. IEEE Trans Pattern Anal Mach Intell (preprint)
8. He R, Zheng WS, Hu BG, Kong XW (2011c) A regularized correntropy framework for robust pattern recognition. Neural Comput (preprint)
9. Huang Y, Liu Q, Metaxas D (2007) A component based deformable model for generalized face alignment. In: International conference on computer vision, pp 1–8
10. Lee C, Landgrebe D (1993) Feature extraction based on decision boundaries. IEEE Trans Pattern Anal Mach Intell 15(4):388–400
11. Lee J, Wang J, Zhang C, Bian Z (2004) Probabilistic tangent subspace: a unified view. In: IEEE conference on computer vision and pattern recognition
12. Li JB, Pan JS, Lu ZM (2009) Face recognition using gabor-based complete kernel fisher discriminant analysis with fractional power polynomial models. Neural Comput Appl 18(6):613–621
13. Li JB, Pan JS, Lu ZM (2009) Kernel optimization-based discriminant analysis for face recognition. Neural Comput Appl 18(6):603–612
14. Li X, Tao D (2010) Subspace learning. Neurocomputing 73(10-12):1539–1540
15. Li Y, Ito W (2005) Shape parameter optimization for adaboosted active shape model. In: International conference on computer vision, pp 251–258
16. Liu C, Wechsler H (2002) Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. IEEE Trans Image Process 11(4):467–476
17. Nie F, Xiang S, Song Y, Zhang C (2007a) Optimal dimensionality discriminant analysis and its application to image recognition. In: IEEE conference on computer vision and pattern recognition
18. Nie F, Xiang S, Zhang C (2007b) Neighborhood minmax projections. In: International joint conference on artificial intelligence, pp 993–998
19. Pang Y, Yuan Y, Li X (2007) Generalized nearest feature line for subspace learning. IEE Electron Lett 43(20):1079–1080
20. Pang Y, Yuan Y, Li X (2009) Iterative subspace analysis based on feature line distance. IEEE Trans Image Process 18:903–907
21. Philips P, Flynn P, Sruggs T, Bowyer K (2005) Overview of the face recognition grand challenge. In: Computer vision and pattern recognition

22. Phillips PJ, Moon H, Rizvi SA, Rauss PJ (2000) The feret evaluation methodology for face-recognition algorithms. IEEE Trans Pattern Anal Mach Intell 22(10):1090–1104
23. Qiu X, Wu L (2005) Face recognition by stepwise nonparametric margin maximum criterion. In: IEEE international conference on computer vision
24. Sim T, Baker S, Bsat M (2005) The cmu pose, illumination, and expression database. IEEE Trans Pattern Anal Mach Intell 25:1615–1618
25. Simard P, LeCun Y, Denker J, Victorri B (2001) Transformation invariance in pattern recognition-tangent distance and tangent propagation. Int J Imaging Syst Technol 11:181–194
26. Song D, Tao D (2009) Discrminative geometry preserving projections. In: Proceedings of the international conference on image processing, pp 2457–2460
27. Tao D, Li X, Wu X, Maybank SJ (2009) Geometric mean for subspace selection. IEEE Trans Pattern Anal Mach Intell 31(2):260–274
28. Zhou T, Tao D, Wu X (2010) Manifold elastic net: A unified framework for sparse dimension reduction. Data Min Knowl Discov 22(3):340–371
29. Wang X, Tang X (2004) A unified framework for subspace face recognition. IEEE Trans Pattern Anal Mach Intell 26(9):1222–1228
30. Wright J, Yang A, Ganesh A, Sastry S, Ma Y (2008) Robust face recognition via sparse representation. IEEE Trans Pattern Anal Mach Intell 31(2):210–227
31. Xiong L, Li J, Zhang C (2007) Discriminant additive tangent spaces for object recognition. In: IEEE conference on computer vision and pattern recognition
32. Yan S, Xu D, Zhang B, Zhang H, Yang Q, Lin S (2007) Graph embedding and extensions: a general framework for dimensionality reduction. IEEE Trans Pattern Anal Mach Intell 29(1):40–51
33. Yang L (2008) Alignment of overlapping locally scaled patches for multidimensional scaling and dimensionality reduction. IEEE Trans Pattern Anal Mach Intell 30(3):438–450
34. Zhang T, Yang J, Zhao D, Ge X (2007) Linear local tangent space alignment and application to face recognition. Neurocomputing 70:1547–1553
35. Zhang T, Li X, Tao D, Yang J (2008) Local coordinates alignment (lca): a novel manifold learning approach. Int J Pattern Recognit Artif Intell 22(4):667–690
36. Zhang T, Tao D, Li X, Yang J (2008c) A unifying framework for spectral analysis based dimensionality reduction. In: Proceedings of the international joint conference on neural networks, pp 1670–1677
37. Zhang T, Tao D, Li X, Yang J (2009) Patch alignment for dimensionality reduction. IEEE Trans Knowl Data Eng 21(9):1299–1313
38. Zhang T, Huang K, Li X, Yang J, Tao D (2010) Discriminative orthogonal neighborhood-preserving projections for classification. IEEE Trans Syst Man Cybern B 40(1):253–263
39. Zhao D, Lin Z, Tang X (2007a) Laplacian pca and its applications. In: International conference on computer vision
40. Zhao D, Lin Z, Xiao R, Tang X (2007b) Linear laplacian discrimination for feature extraction. In: IEEE conference on computer vision and pattern recognition