Hindawi
Mathematical Problems in Engineering
Volume 2019, Article ID 3145973, 10 pages
https://doi.org/10.1155/2019/3145973

Research Article

A Similar Distribution Discriminant Analysis with Orthogonal and Nearly Statistically Uncorrelated Characteristics

Zhibo Guo and Ying Zhang

College of Information Engineering, Yangzhou University, Yangzhou 225009, China

Correspondence should be addressed to Zhibo Guo; zbguo@yzu.edu.cn

Received 28 April 2019; Revised 29 June 2019; Accepted 3 September 2019; Published 20 October 2019

Academic Editor: David González

It is very difficult to process and analyze high-dimensional data directly. Therefore, it is necessary to learn a potential subspace of high-dimensional data through excellent dimensionality reduction algorithms to preserve the intrinsic structure of high-dimensional data and abandon the less useful information. Principal component analysis (PCA) and linear discriminant analysis (LDA) are two popular dimensionality reduction methods for high-dimensional sensor data preprocessing. LDA contains two basic methods, namely, classic linear discriminant analysis and FS linear discriminant analysis. In this paper, a new method, called similar distribution discriminant analysis (SDDA), is proposed based on the similarity of samples' distribution. Furthermore, the method of solving the optimal discriminant vector is given. These discriminant vectors are orthogonal and nearly statistically uncorrelated. The disadvantages of PCA and LDA are overcome, and the extracted features are more effective by using SDDA. The recognition performance of SDDA exceeds PCA and LDA largely. Some experiments on the Yale face database, FERET face database, and UCI multiple features dataset demonstrate that the proposed method is effective. The results reveal that SDDA obtains better performance than comparison dimensionality reduction methods.

1. Introduction

The data collected by various sensors (such as visual sensors and sound sensors) are mostly high-dimensional, which brings inconvenience to the later processing and analysis of data. In order to effectively utilize these high-dimensional data, it is necessary to adopt effective dimensionality reduction algorithms. In fact, dimensionality reduction is an effective data preprocessing method. It reduces the size of data while retaining the valid data, which brings great convenience to the later analysis and calculation of data. In pattern recognition, data dimensionality reduction has a wide range of applications. KL-based principal component analysis (PCA) [1, 2] and linear discriminant analysis (LDA) [3–6] are the two most widely used dimensionality reduction methods. PCA and LDA have been widely used in the analysis and processing of various types of data. They can be used in data compression, data preprocessing, data mining, data retrieval, data classification, and so on. Independent component analysis (ICA) is a data processing method developed from solving blind source separation, which decomposes the original data to obtain independent components. ICA is helpful to find the maximal independent projection direction as the dimensions of data are reduced. But in ICA, there is a preprocessing process for data, that is, PCA and whitening. In the pattern recognition field, some researchers have proven that the overall performance of ICA is not better than that of PCA by conducting experimental comparisons between the two methods [7, 8]. At present, PCA and LDA have a lot of applications in image processing, voice processing, communication, network, and others. Many researchers [9–27] have done extensibility research based on LDA and PCA methods and have made some progress. But there are some shortcomings in the use of PCA and LDA. The disadvantage of PCA is that the data after dimensionality reduction have no clustering characteristics. The classification accuracy is uncertain by using the features after dimensionality reduction. The disadvantage of LDA is that there is a phenomenon of overfitting to training samples. The classification accuracy is closely related to the

characteristics of training samples. PCA is a dimension reduction analysis method that maintains the maximum dispersion of samples. However, category information is not introduced into the dimensionality reduction process of PCA, with the result that the accuracy of using minimum distance measurement is usually lower than the accuracy of using nearest neighbor measurement. In contrast, LDA can obtain the best identifying projection information for classification. Therefore, the accuracy of LDA by the minimum distance method is close to the nearest neighbor method. LDA takes category information into account, so its test accuracy is better than that of PCA, but on the other side, the LDA method may overfit the training set, which worsens the generalization ability. In particular, when there is quite a difference between the training set and the test set, it is likely that the test result of LDA will be not ideal [7, 28, 29]. Pattern recognition for images has high application and research value, so it has become a hot area of research in the field of pattern recognition and machine vision, especially for face recognition [30–42]. Meanwhile, face recognition is also an important way to verify the effectiveness of pattern recognition methods. The LDA-based methods are widely used in face recognition [23–27, 30–42]. The research on LDA can be traced back to a classic paper [3] by Fisher in 1936. The basic idea is to choose the vector that makes the Fisher criterion function as max as optimal projection vector, so that the sample can achieve the maximum between-class scatter and the smallest within-class scatter after being projected in this direction. Based on the Fisherface method, Wilks and Duda proposed classic linear discriminant analysis (CLDA), respectively [4, 5]. Foley and Sammon proposed a method called FS linear discriminant analysis (FSLDA) [6], in which a set of optimal discriminant vectors satisfying the orthogonal condition is used for dimensionality reduction. The specific algorithm for solving the optimal discrimination vectors of two-class cases is presented by Foley, and the solution of the optimal discriminant vectors in multiclass cases is given by Duchene and Leclercq [43]. Jin et al. proposed the concept of uncorrelated linear discriminant analysis (ULDA) [44, 45] for optimal discriminant vectors from the perspective of statistical irrelevance. A simple algorithm for solving the optimal set of uncorrelated discriminant vectors is presented in the literature [45], and it is pointed out that ULDA is equal to CLDA under the condition that the eigenvalues of the generalized characteristic equation corresponding to the Fisher criterion function are not equal.

Although the discriminant vectors of CLDA have statistically uncorrelated characteristics, they are not orthogonal. In contrast, the discriminant vectors of FSLDA are orthogonal and statistically correlated. Some researchers have argued the performance of orthogonal discriminant vectors is better than that of the statistically uncorrelated vectors [46, 47], and some researchers hold a contrary opinion [45, 48]. Actually, both of these characteristics have certain pertinence. If only one of them is considered, it is deficient. Firstly, nonorthogonal discriminant vectors are unfavorable factors for extracting useful features, which weakens the generalization ability of test samples. Especially when the number of training samples is small and the distance between samples is small, the test performance of CLDA is inferior to that of FSLDA. Secondly, the discriminant vectors of FSLDA are composed of orthogonal normalized vectors. However, in the case of fewer categories, more samples per class, and larger intraclass dispersion, the redundancy between the discriminant features obtained by each discriminant vector is very large, that is to say, the statistical uncorrelation characteristics of FSLDA is very poor. For example, in terms of character recognition, the performance of FSLDA is significantly worse than that of CLDA.

In general, statistical uncorrelation is only strictly statistically uncorrelated for training samples but only nearly statistically uncorrelated for test samples. Therefore, only nearly statistical uncorrelation is required for the optimal discriminant vectors. On the other hand, orthogonality is a strict restriction, which reflects the perpendicular relation of each axis in Euclidean space and enhances the generalization ability of test samples. So we conclude that the discriminant vectors of the most effective discriminant method should be orthogonal and nearly statistically uncorrelated.

To solve the above problems, the paper presents a similar distribution discriminant analysis (SDDA) method from the similarity of samples' distribution. The advantage of the SDDA is that the projection vector has orthogonal characteristics, and the data after dimensionality reduction have nearly statisically uncorrelated characteristics, and the distribution of the projection vector approximates the distribution of principal components in the center of the sample class. The proposed method uses the statistically uncorrelated characteristics of PCA to combine PCA with the class labels of samples and gives the solution of the optimal discriminant vectors. These discriminant vectors have orthogonal characteristics and nearly statisically uncorrelated characteristics. The SDDA algorithm requires that the data distribution after the dimensionality reduction of samples is similar to the distribution of the principal component of the original samples. That is to say, in the process of dimensionality reduction, the principal component characteristics of the original sample are better preserved. The SDDA algorithm requires that the data distribution after the dimensionality reduction of the sample is similar to the distribution of the principal component of the original sample. The principal component property can suppress overfitting well, which solves the problem of overfitting of LDA. The proposed SDDA method overcomes the disadvantages of the two basic methods of LDA and concentrates the advantages of the basic methods of LDA together, so the extracted distinguishing features are more effective, which improves the recognition performance and adaptability. Finally, the effectiveness is validated through some experiments on the Yale face database, FERET face database, and UCI multiple features dataset. The results of these experiments indicate that the recognition accuracy of SDDA is superior to the two basic methods of PCA and LDA.

2. Related Work

The N samples in the training set $\{\mathbf{x}_i^{(j)}\}$ come from c categories: $\omega_1, \omega_2, \ldots, \omega_c$, where $i = 1, 2, \ldots, c$ and $j = 1, 2, \ldots, n_i$. n_i is

the number of samples in the ith class, and $\mathbf{x}_i^{(j)}$ is the jth sample that comes from the ith class. All samples are m dimension column vectors. Thus, the within-class scatter matrix, the between-class scatter matrix, and the total scatter matrix are defined as $\mathbf{S}_w$, $\mathbf{S}_b$, and $\mathbf{S}_t$, respectively, with the following expressions:

$$\mathbf{S}_w = \sum_{i=1}^{c} P(\omega_i) \frac{1}{n_i} \sum_{j=1}^{n_i} \left(\mathbf{x}_i^{(j)} - \overline{\mathbf{x}}_i\right)\left(\mathbf{x}_i^{(j)} - \overline{\mathbf{x}}_i\right)^{\mathrm{T}}, \tag{1}$$

$$\mathbf{S}_b = \sum_{i=1}^{c} P(\omega_i) \left(\overline{\mathbf{x}}_i - \overline{\mathbf{x}}\right)\left(\overline{\mathbf{x}}_i - \overline{\mathbf{x}}\right)^{\mathrm{T}}, \tag{2}$$

$$\mathbf{S}_t = \mathbf{S}_w + \mathbf{S}_b = \sum_{i=1}^{c} P(\omega_i) \frac{1}{n_i} \sum_{j=1}^{n_i} \left(\mathbf{x}_i^{(j)} - \overline{\mathbf{x}}\right)\left(\mathbf{x}_i^{(j)} - \overline{\mathbf{x}}\right)^{\mathrm{T}}, \tag{3}$$

where $\overline{\mathbf{x}}_i = E(\mathbf{x} \mid \omega_i) = (1/n_i)\sum_{j=1}^{n_i} \mathbf{x}_i^{(j)}$ denotes the mean vector of all samples in the ith class and $\overline{\mathbf{x}} = E(\mathbf{x}) = \sum_{i=1}^{c} P(\omega_i)\overline{\mathbf{x}}_i$ is the expected mean vector of all samples. $P(\omega_i)$ is a prior probability of the samples in the ith class, which is generally taken as $P(\omega_i) = n_i/N$. Then, the mean vector of all samples can be represented as $\overline{\mathbf{x}} = (1/N)\sum_{i=1}^{c}\sum_{j=1}^{n_i} \mathbf{x}_i^{(j)}$.

2.1. Classical Principal Component Analysis. The criterion function of PCA is defined as (4). The vectors in the optimal projection vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_d$ make (4) reach the maximum, and they are a group of normal orthogonal vectors. Its physical significance can be interpreted as maximizing the total dispersion of the projected features:

$$J_p(\mathbf{a}) = \mathbf{a}^{\mathrm{T}} \mathbf{S}_t \mathbf{a}. \tag{4}$$

Actually, vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_d$ in the optimal projection vectors are normal orthogonal eigenvectors corresponding to d largest eigenvalues. The criterion function of PCA can be also represented as follows:

$$J_p(\mathbf{A}) = \mathrm{tr}\left(\mathbf{A}^{\mathrm{T}} \mathbf{S}_t \mathbf{A}\right). \tag{5}$$

And the best projecting matrix is $\mathbf{A}_{\mathrm{opt}} = \underset{\mathbf{A}}{\mathrm{argmax}}\, J_p(\mathbf{A}) = (\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_d)$.

2.2. Linear Discriminant Analysis (LDA). LDA was proposed by Fisher firstly. The basic idea is to select the vector $\boldsymbol{\varphi}_{\mathrm{opt}}$ that maximizes the Fisher criterion function and take $\boldsymbol{\varphi}_{\mathrm{opt}}$ as the optimal projection direction, which is also called optimal discriminant vector. Then, the ratio of the interclass dispersion to the intraclass dispersion reaches the maximum after the samples projected in this direction. Fisher discrimination criterion function is defined as

$$J_f(\boldsymbol{\varphi}) = \frac{\boldsymbol{\varphi}^{\mathrm{T}} \mathbf{S}_b \boldsymbol{\varphi}}{\boldsymbol{\varphi}^{\mathrm{T}} \mathbf{S}_w \boldsymbol{\varphi}}, \tag{6}$$

where $\mathbf{S}_w$ is the within-class scatter matrix, $\mathbf{S}_b$ is the between-class scatter matrix, and $\boldsymbol{\varphi}$ is a nonzero column vector of any number of dimensions.

The Fisher criterion function combines the between-class and within-class dispersion of samples skillfully and provides a perfect criterion for determining the optimal projection direction.

2.2.1. Classic Linear Discriminant Analysis. Inspired by LDA, Wilks and Duda extended the two-class classification problem of finding one optimal projection direction to the multiclass classification problem of finding multiple optimal projection directions. Their idea is called the classic Fisher linear discriminant analysis, and the classical Fisher discriminant criterion function is (7) or (8):

$$J_c(\mathbf{A}) = \frac{\left|\mathbf{A}^{\mathrm{T}} \mathbf{S}_b \mathbf{A}\right|}{\left|\mathbf{A}^{\mathrm{T}} \mathbf{S}_w \mathbf{A}\right|}, \tag{7}$$

$$J_c(\mathbf{A}) = \mathrm{tr}\left[\left(\mathbf{A}^{\mathrm{T}} \mathbf{S}_w \mathbf{A}\right)^{-1}\left(\mathbf{A}^{\mathrm{T}} \mathbf{S}_b \mathbf{A}\right)\right]. \tag{8}$$

In fact, the column vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_d$ in the optimal projection matrix $\mathbf{A}_{\mathrm{opt}}$ of classic Fisher linear discriminant analysis are taken from the eigenvectors corresponding to d largest eigenvalues of the generalized characteristic equation $\mathbf{S}_b\mathbf{A} = \lambda \mathbf{S}_w \mathbf{A}$.

2.2.2. FS Linear Discriminant Analysis. FSLDA aims at finding a set of optimal discriminating vectors $\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_d$. They maximize the Fisher criterion function and satisfy the following orthogonal condition:

$$\boldsymbol{\varphi}_i^{\mathrm{T}} \boldsymbol{\varphi}_j = 0, \quad \forall i \neq j. \tag{9}$$

The first vector of the FS optimal discriminant vectors is the Fisher optimal discriminating direction, that is, the unit eigenvector $\boldsymbol{\varphi}_1$ corresponding to the maximum eigenvalue of the generalized characteristic equation $\mathbf{S}_b\mathbf{U} = \lambda \mathbf{S}_w \mathbf{U}$. After the first k discriminant vectors $\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_k$ are found, the $k+1$th discriminant vector $\boldsymbol{\varphi}_{k+1}$ is obtained by solving the following optimization problem:

$$\begin{cases} \max\left(J_f(\boldsymbol{\varphi})\right), \\ \boldsymbol{\varphi}_j^{\mathrm{T}} \boldsymbol{\varphi} = 0, \quad j = 1, \ldots, k, \\ \boldsymbol{\varphi} \in R^n. \end{cases} \tag{10}$$

In fact, $\boldsymbol{\varphi}_{k+1}$ is the eigenvector corresponding to the maximum eigenvalue of the generalized characteristic equation:

$$\mathbf{B}_k \mathbf{S}_b \boldsymbol{\varphi} = \lambda \mathbf{S}_w \boldsymbol{\varphi}, \tag{11}$$

where $\mathbf{B}_k = \mathbf{I} - \mathbf{D}_k^{\mathrm{T}}(\mathbf{D}_k \mathbf{S}_w^{-1} \mathbf{D}_k^{\mathrm{T}})^{-1} \mathbf{D}_k \mathbf{S}_w^{-1}$ and $\mathbf{D}_k = (\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2, \ldots, \boldsymbol{\varphi}_k)^{\mathrm{T}}$.

3. Similar Distribution Discriminant Analysis (SDDA)

By reducing the dimensions, the proposed SDDA method in this paper makes the total distribution of extracted features closest to the principal component distribution and the extracted features satisfy the minimization of within-class dispersion. In other words, the extracted features not only have a good performance in discrimination, but also retain

the principal component characteristics. At the same time, the optimal discriminant vectors are composed of orthogonal and nearly statistically uncorrelated vectors, which makes the extracted discriminant features more effective and improves the performance of classification and recognition.

3.1. Theoretical Framework of SDDA. Supposing there are two points $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in m-dimensional space, which represent the vector $(\alpha_1, \alpha_2, \ldots, \alpha_m)$ and the vector $(\beta_1, \beta_2, \ldots, \beta_m)$, respectively. For the similarity between $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, a similarity measurement is usually adopted, whose formula is as follows:

$$s = \frac{\widehat{\boldsymbol{\alpha}}\widehat{\boldsymbol{\beta}}^{\mathrm{T}}}{\|\widehat{\boldsymbol{\alpha}}\|_2 \cdot \|\widehat{\boldsymbol{\beta}}\|_2} = \frac{\widehat{\boldsymbol{\alpha}}\widehat{\boldsymbol{\beta}}^{\mathrm{T}}}{\sqrt{\left(\widehat{\boldsymbol{\alpha}}\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\right)} \times \sqrt{\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\beta}}^{\mathrm{T}}}}, \tag{12}$$

where vector $\widehat{\boldsymbol{\alpha}}$ is $\widehat{\boldsymbol{\alpha}} = (\alpha_1 - \overline{\alpha}, \alpha_2 - \overline{\alpha}, \ldots, \alpha_m - \overline{\alpha})$ and vector $\widehat{\boldsymbol{\beta}}$ is $\widehat{\boldsymbol{\beta}} = (\beta_1 - \overline{\beta}, \beta_2 - \overline{\beta}, \ldots, \beta_m - \overline{\beta})$, in which $\overline{\boldsymbol{\alpha}}$ and $\overline{\boldsymbol{\beta}}$ represent the mean of all elements in $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, respectively. The larger the value of s is, the more similar the two vectors are, and $s = 1$ means that the two vectors are completely similar.

Extend the similarity measurement from two vectors to two sets of vectors. Supposing one set of vectors is $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]$ and the other set of vectors is $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n]$, where $\mathbf{x}_j$ and $\mathbf{y}_j$ are both m-dimensional column vectors. Set $\widehat{\mathbf{X}} = (\widehat{\mathbf{x}}_1, \widehat{\mathbf{x}}_2, \ldots, \widehat{\mathbf{x}}_n)$ and $\widehat{\mathbf{Y}} = (\widehat{\mathbf{y}}_1, \widehat{\mathbf{y}}_2, \ldots, \widehat{\mathbf{y}}_n)$, the columns of which are represented as $\widehat{\mathbf{x}}_j = \mathbf{x}_j - \overline{\mathbf{x}}$ and $\widehat{\mathbf{y}}_j = \mathbf{y}_j - \overline{\mathbf{y}}$, and where vectors $\overline{\mathbf{x}}$ and $\overline{\mathbf{y}}$ represent the mean of all column vectors in $\mathbf{X}$ and $\mathbf{Y}$, respectively. Let $\widehat{\mathbf{x}}^i$ and $\widehat{\mathbf{y}}^i$ be row vectors of $\widehat{\mathbf{X}}$ and $\widehat{\mathbf{Y}}$, then the similarity measurement formula of $\mathbf{X}$ and $\mathbf{Y}$ can be defined as (13) or (14):

$$S = \frac{1}{m}\sum_{i=1}^{m} \frac{\widehat{\mathbf{x}}^i\left(\widehat{\mathbf{y}}^i\right)^{\mathrm{T}}}{\sqrt{\widehat{\mathbf{x}}^i\left(\widehat{\mathbf{x}}^i\right)^{\mathrm{T}}}\sqrt{\widehat{\mathbf{y}}^i\left(\widehat{\mathbf{y}}^i\right)^{\mathrm{T}}}}, \tag{13}$$

$$S = \frac{1}{m}\sum_{i=1}^{m} \frac{\widehat{\mathbf{x}}^i\left(\widehat{\mathbf{y}}^i\right)^{\mathrm{T}}\widehat{\mathbf{y}}^i\left(\widehat{\mathbf{x}}^i\right)^{\mathrm{T}}}{\widehat{\mathbf{x}}^i\left(\widehat{\mathbf{x}}^i\right)^{\mathrm{T}}\widehat{\mathbf{y}}^i\left(\widehat{\mathbf{y}}^i\right)^{\mathrm{T}}}, \tag{14}$$

where $\widehat{\mathbf{x}}^i(\widehat{\mathbf{x}}^i)^{\mathrm{T}} = \widehat{\mathbf{y}}^i(\widehat{\mathbf{y}}^i)^{\mathrm{T}}$. Equation (14) is easier to analyze, so it is adopted in this paper. $S = 1$ indicates that the distribution of the two sets of vectors is completely consistent, which is called distribution equivalence.

For a given matrix $\widehat{\mathbf{Y}}$, in other words, $\widehat{\mathbf{Y}}$ is a matrix with certain distribution, and the dimension of $\widehat{\mathbf{x}}_j$ is larger than that of $\widehat{\mathbf{y}}_j$ (the dimension of $\widehat{\mathbf{y}}_j$ is m). If the discriminative feature of the samples can be extracted by dimension reduction, then the distribution of the overall discrimination features is closest to the expected distribution. In other words, we need to find an optimal projection matrix $\mathbf{A}_{\mathrm{opt}}$ to satisfy the condition $\mathbf{A}_{\mathrm{opt}} = \arg\max_{\mathbf{A}}(J(\mathbf{A}))$:

$$\begin{cases} J(\mathbf{A}) = \dfrac{1}{m}\displaystyle\sum_{i=1}^{m} \dfrac{\mathbf{a}_i^{\mathrm{T}}\widehat{\mathbf{X}}\left(\widehat{\mathbf{y}}^i\right)^{\mathrm{T}}\widehat{\mathbf{y}}^i\widehat{\mathbf{X}}^{\mathrm{T}}\mathbf{a}_i}{\left(\mathbf{a}_i^{\mathrm{T}}\widehat{\mathbf{X}}\widehat{\mathbf{X}}^{\mathrm{T}}\mathbf{a}_i\right)\widehat{\mathbf{y}}^i\left(\widehat{\mathbf{y}}^i\right)^{\mathrm{T}}}, \\ \mathbf{a}_i^{\mathrm{T}}\widehat{\mathbf{X}}\widehat{\mathbf{X}}^{\mathrm{T}}\mathbf{a}_i = \widehat{\mathbf{y}}^i\left(\widehat{\mathbf{y}}^i\right)^{\mathrm{T}}, \end{cases} \tag{15}$$

where $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m$ are column vectors of $\mathbf{A}$.

Given a set of n samples $\mathbf{X} = \{\mathbf{x}_j\}$ from c class, where $\mathbf{x}_j$ is an l-dimensional column vector. Set $\widehat{\mathbf{X}} = \{\widehat{\mathbf{x}}_j\}$, in which $\widehat{\mathbf{x}}_j = (\mathbf{x}_j - \overline{\mathbf{x}})$ and $\overline{\mathbf{x}}$ is the mean vector of all the samples.

Because the principal components of samples are statistically uncorrelated, using the principal components to construct the expected matrix $\widehat{\mathbf{Y}}$, and then solving the projection matrix with orthogonal characteristics, the discriminant vectors can have both orthogonal and nearly statistically correlated characteristics. In addition, the obtained discriminant vectors should be helpful for classification, that is to say, the expected matrix should have the characteristics of the smallest distance within the class, so the expected matrix is established by using the principal components of the sample class mean, and the expected vectors belonging to each class are the same.

Let $\mathbf{Z} = \{\mathbf{z}_k\}$ be the set of the principal components of the sample class mean in the total samples $\mathbf{X}$, where $\mathbf{z}_k = \mathbf{P}^{\mathrm{T}}(\overline{\mathbf{x}}_k - \overline{\mathbf{x}})$. $\overline{\mathbf{x}}_k\,(k = 1, \ldots, c)$ is the mean vector for each class. $\mathbf{P}$ is the projection matrix of the principal component of the class mean, which consists of $c - 1$ standard orthogonal column vectors $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_{c-1}$ for nonzero eigenvalues.

The principal component extension matrix can be defined as

$$\widehat{\mathbf{Y}} = \begin{bmatrix} \widehat{\mathbf{y}}^1 \\ \vdots \\ \widehat{\mathbf{y}}^{c-1} \end{bmatrix} = \left[\widehat{\mathbf{y}}_1, \widehat{\mathbf{y}}_2, \ldots, \widehat{\mathbf{y}}_n\right] = \left[\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_c\right], \tag{16}$$

where $\mathbf{Z}_k = \overbrace{[\mathbf{z}_k, \ldots, \mathbf{z}_k]}^{d_k}$ in which d_k is the number of samples for the kth class.

The set of the class mean principal components $\mathbf{Z}$ has statistically uncorrelated characteristics, which means $\mathbf{Z}\mathbf{Z}^{\mathrm{T}} = \sum_{i=1}^{c}\mathbf{z}_k\mathbf{z}_k^{\mathrm{T}} = \Lambda$ and matrix Λ is diagonal. For the same number of samples per class $d_k = d$, we get $\widehat{\mathbf{Y}}\widehat{\mathbf{Y}}^{\mathrm{T}} = \sum_{k=1}^{c}(d_k\mathbf{z}_k\mathbf{z}_k^{\mathrm{T}}) = d\sum_{i=1}^{c}\mathbf{z}_k\mathbf{z}_k^{\mathrm{T}}$. Therefore, the principal component extension matrix also has the property of statistical uncorrelation.

3.2. Solution to the Projection Matrix $\mathbf{A}_{opt}$. Due to the statistically uncorrelated characteristics of the principal component extension matrix of the projection matrix, $\mathbf{A}_{\mathrm{opt}} = \arg\max_{\mathbf{A}}(J(\mathbf{A}))$ makes $\mathbf{A}_{\mathrm{opt}}\mathbf{X}$ statistically uncorrelated to some extent, so the discriminant vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m$ to be solved only need to be mutually orthogonal. That is to say, after the first k discriminant vectors are solved, the $k + 1$th discriminant vector $\mathbf{a}_{k+1}$ is obtained by solving the following optimization problem:

$$\begin{cases} \max\left(J\left(\mathbf{a}_{k+1}\right)\right), \\ \mathbf{a}_i^{\mathrm{T}}\mathbf{a}_{k+1} = 0, \quad i = 1, \ldots, k, \\ \mathbf{a}_{k+1} \in R^n. \end{cases} \tag{17}$$

In order to obtain the $k + 1$th discriminant vector $\mathbf{a}_{k+1}$, we define the following function:

$$f(\mathbf{a}_{k+1}) = \frac{\mathbf{a}_{k+1}^{\mathrm{T}}\widehat{\mathbf{X}}\left(\widehat{\mathbf{y}}^{k+1}\right)^{\mathrm{T}}\widehat{\mathbf{y}}^{k+1}\widehat{\mathbf{X}}^{\mathrm{T}}\mathbf{a}_{k+1}}{\left(\mathbf{a}_{k+1}^{\mathrm{T}}\widehat{\mathbf{X}}\widehat{\mathbf{X}}^{\mathrm{T}}\mathbf{a}_{k+1}\right)\widehat{\mathbf{y}}^{k+1}\left(\widehat{\mathbf{y}}^{k+1}\right)^{\mathrm{T}}}, \tag{18}$$

where $\widehat{\mathbf{y}}^{k+1}$ is a given vector, so $\widehat{\mathbf{y}}^{k+1}(\widehat{\mathbf{y}}^{k+1})^{\mathrm{T}}$ has no effect on the solution of $\mathbf{a}_i$, and thus $f(\mathbf{a}_{k+1})$ is rewritten as

$$f(\mathbf{a}_{k+1}) = \frac{\mathbf{a}_{k+1}^{\mathrm{T}}\widehat{\mathbf{X}}\left(\widehat{\mathbf{y}}^{k+1}\right)^{\mathrm{T}}\widehat{\mathbf{y}}^{k+1}\widehat{\mathbf{X}}^{\mathrm{T}}\mathbf{a}_{k+1}}{\left(\mathbf{a}_{k+1}^{\mathrm{T}}\widehat{\mathbf{X}}\widehat{\mathbf{X}}^{\mathrm{T}}\mathbf{a}_{k+1}\right)\widehat{\mathbf{y}}^{k+1}\left(\widehat{\mathbf{y}}^{k+1}\right)^{\mathrm{T}}} = \frac{\mathbf{a}_{k+1}^{\mathrm{T}}\mathbf{S}_{k+1}\mathbf{a}_{k+1}}{\mathbf{a}_{k+1}^{\mathrm{T}}\mathbf{S}_{\mathrm{t}}\mathbf{a}_{k+1}}. \tag{19}$$

According to Lagrange multipliers, $\mathbf{a}_{k+1}$ makes (20) achieve the maximum value:

$$L(\mathbf{a}_{k+1}) = \mathbf{a}_{k+1}^{\mathrm{T}}\mathbf{S}_{k+1}\mathbf{a}_{k+1} - \lambda\left[\mathbf{a}_{k+1}^{\mathrm{T}}\mathbf{S}_{\mathrm{t}}\mathbf{a}_{k+1} - \theta\right] - \sum_{i=1}^{k}\boldsymbol{\eta}_i\mathbf{a}_{k+1}^{\mathrm{T}}\mathbf{a}_i. \tag{20}$$

Taking the derivative of $\mathbf{a}_{k+1}$ and setting it to zero, then we have

$$2\mathbf{S}_{k+1}\mathbf{a}_{k+1} - 2\lambda\mathbf{S}_{\mathrm{t}}\mathbf{a}_{k+1} - \sum_{i=1}^{k}\boldsymbol{\eta}_i\mathbf{a}_i = 0. \tag{21}$$

We define $\mathbf{A}_k = [\mathbf{a}_1, \ldots, \mathbf{a}_k]$ and $\boldsymbol{\eta} = [\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_k]^{\mathrm{T}}$, thus

$$2\mathbf{S}_{k+1}\mathbf{a}_{k+1} - 2\lambda\mathbf{S}_{\mathrm{t}}\mathbf{a}_{k+1} - \mathbf{A}_k^{\mathrm{T}}\boldsymbol{\eta} = 0. \tag{22}$$

Multiply $\mathbf{a}_{k+1}^{\mathrm{T}}$ on both sides of (21), and according to (17), the third term is zero. Thus, we obtain

$$\lambda = f(\mathbf{a}_{k+1}). \tag{23}$$

The solution to the problem is to maximize the value of λ.

Multiply $\mathbf{a}_j^{\mathrm{T}}\mathbf{S}_{\mathrm{t}}^{-1}$ $(j = 1, \ldots, k)$ on both sides of (21), and according to (17), the second term is zero. Then, we get

$$2\mathbf{a}_j^{\mathrm{T}}\mathbf{S}_{\mathrm{t}}^{-1}\mathbf{S}_{k+1}\mathbf{a}_{k+1} - \sum_{i=1}^{k}\boldsymbol{\eta}_i\mathbf{a}_j^{\mathrm{T}}\mathbf{S}_{\mathrm{t}}^{-1}\mathbf{a}_i = 0, \tag{24}$$

which can be rewritten as

$$2\begin{bmatrix}\mathbf{a}_1^{\mathrm{T}}\\ \vdots\\ \mathbf{a}_k^{\mathrm{T}}\end{bmatrix}\mathbf{S}_{\mathrm{t}}^{-1}\mathbf{S}_{k+1}\mathbf{a}_{k+1} - \begin{bmatrix}\mathbf{a}_1^{\mathrm{T}}\\ \vdots\\ \mathbf{a}_k^{\mathrm{T}}\end{bmatrix}\mathbf{S}_{\mathrm{t}}^{-1}\begin{bmatrix}\mathbf{a}_1^{\mathrm{T}}\\ \vdots\\ \mathbf{a}_k^{\mathrm{T}}\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\boldsymbol{\eta}_1\\ \vdots\\ \boldsymbol{\eta}_k\end{bmatrix} = 0. \tag{25}$$

And (22) is rewritten as

$$2\mathbf{A}_k^{\mathrm{T}}\mathbf{S}_{\mathrm{t}}^{-1}\mathbf{S}_{k+1}\mathbf{a}_{k+1} = \mathbf{A}_k^{\mathrm{T}}\mathbf{S}_{\mathrm{t}}^{-1}\mathbf{A}_k\boldsymbol{\eta}. \tag{26}$$

So, the updating rule of $\boldsymbol{\eta}$ is presented as

$$\boldsymbol{\eta} = 2\left(\mathbf{A}_k^{\mathrm{T}}\mathbf{S}_{\mathrm{t}}^{-1}\mathbf{A}_k\right)^{-1}\mathbf{A}_k^{\mathrm{T}}\mathbf{S}_{\mathrm{t}}^{-1}\mathbf{S}_{k+1}\mathbf{a}_{k+1}. \tag{27}$$

By combining formulas (27) and (22), we have

$$2\mathbf{S}_{k+1}\mathbf{a}_{k+1} - 2\lambda\mathbf{S}_{\mathrm{t}}\mathbf{a}_{k+1} - 2\mathbf{A}_k^{\mathrm{T}}\left(\mathbf{A}_k^{\mathrm{T}}\mathbf{S}_{\mathrm{t}}^{-1}\mathbf{A}_k\right)^{-1}\mathbf{A}_k^{\mathrm{T}}\mathbf{S}_{\mathrm{t}}^{-1}\mathbf{S}_{k+1}\mathbf{a}_{k+1} = 0. \tag{28}$$

After some rearrangement, we obtain

$$\left(\mathbf{S}_{k+1} - \mathbf{A}_k^{\mathrm{T}}\left(\mathbf{A}_k^{\mathrm{T}}\mathbf{S}_{\mathrm{t}}^{-1}\mathbf{A}_k\right)^{-1}\mathbf{A}_k^{\mathrm{T}}\mathbf{S}_{\mathrm{t}}^{-1}\mathbf{S}_{k+1}\right)\mathbf{a}_{k+1} = \lambda\mathbf{S}_{\mathrm{t}}\mathbf{a}_{k+1}, \tag{29}$$

where $\mathbf{a}_1$ is the eigenvector corresponding to the largest eigenvalue of $\mathbf{S}_1\mathbf{a}_{k+1} = \lambda\mathbf{S}_{\mathrm{t}}\mathbf{a}_{k+1}$. $\mathbf{a}_{k+1}$ is the eigenvector corresponding to the largest eigenvalue of generalized characteristic of (29). In order to satisfy $\mathbf{a}_i^{\mathrm{T}}\widehat{\mathbf{X}}\widehat{\mathbf{X}}^{\mathrm{T}}\mathbf{a}_i = \widehat{\mathbf{y}}^i(\widehat{\mathbf{y}}^i)^{\mathrm{T}}$, $\mathbf{a}_i$ needs to be adjusted to $\omega_i\mathbf{a}_i$ after being calculated, where $\omega_i = \sqrt{(\widehat{\mathbf{y}}^i(\widehat{\mathbf{y}}^i)^{\mathrm{T}})/(\mathbf{a}_i^{\mathrm{T}}\widehat{\mathbf{X}}\widehat{\mathbf{X}}^{\mathrm{T}}\mathbf{a}_i)}$.

Remarkably, all the samples $\widehat{\mathbf{x}}_j$ have to be compressed by K-L transform to reduce the original samples from high-dimensional to low-dimensional ones if the matrix $\mathbf{S}_{\mathrm{t}}$ is not invertible, so it can be ensured that $\mathbf{S}_{\mathrm{t}}$ is reversible after dimensionality reduction.

The SDDA method proposed in this manuscript mainly solves the adaptability problem in various applications of two basic methods of LDA. The architecture of the proposed SDDA starts from the classic PCA and the classic LDA, so SDDA itself is also a basic method, which is in the same level as the comparison methods and can be used as a supplement to the classic PCA and the classic LDA. So in this manuscript, SDDA is only used as a basic method and compared with the existing classical methods. Actually, some techniques for improving PCA and improving LDA can also be used in the proposed SDDA method. For example, we can learn from the construction process of KPCA, KFSLDA, and KFDA to use nuclear techniques to construct KSDDA.

4. Experiment Results and Analysis

We conduct some experiments on the Yale face database, FERET face database, and UCI multiple features dataset to demonstrate the adaptability and effectiveness of the proposed algorithm to different objects. The proposed algorithm is compared with SDDA, PCA, and two basic methods of LDA (CLDA and FSLDA), and we analyze the comparison results.

4.1. Experiment on the Yale Face Database. Yale face database [49] is taken from 15 volunteers with each one having 11 images. Different images of each person are quite different in expression changes and light changes. Figure 1 is 11 images of one person in the Yale face database.

Since the discriminant vectors of SDDA are obtained by the orthonormal constraint, it is not necessary to carry out the experiment on its orthogonal characteristics. We have only done the experiment to verify the statistically uncorrelated characteristic and show it intuitively with the statistical uncorrelation diagrams. The elements in the diagrams of statistical uncorrelation are $p_{i,j} = \mathbf{a}_i^{\mathrm{T}}\mathbf{S}_{\mathrm{t}}\mathbf{a}_j$, where $p_{i,j}$ means the element that comes from the ith row and the jth column. As shown in Figure 2, in addition to the diagonal element values, the closer the element values of other locations are to 0 (black), the better the statistically uncorrelated characteristics between the discriminant vectors are. Comparing the statistical uncorrelation diagrams of SDDA, CLDA, FSLDA, and PCA, it is showed that the discriminant vectors obtained by SDDA are almost completely statistically uncorrelated while FSLDA has poorly statistically uncorrelated characteristics.

In order to evaluate the performance of the proposed SDDA, we conduct two sets of experiments on the Yale face

Figure 1: Eleven images of one person in the Yale face database.

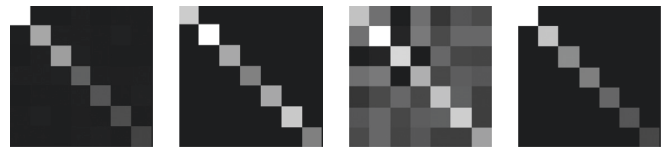

Figure 2: Statistical uncorrelation diagrams: (a) SDDA, (b) CLDA, (c) FSLDA, and (d) PCA.

database. One set selects the odd number (6 samples) of each person as the training set and the even number (5 samples) of each person as the test set, and the other set of experiments selects the even number (5 samples) of each person as the training set and the odd number (6 samples) of each person as the test set. The final results are averaged from the results of the two sets of experiments. Minimum distance and nearest neighbor are adopted as the measurement method in this paper.

Because of the small number of samples in the Yale face database, the lower dimension is taken in the experiment, that is to say, only a few projection vectors are used to extract features. In this experiment, the number of features is the dimension of sample reduction. The dimension of samples is reduced by various algorithms. Table 1, aimed at the number of features from 4 to 11, shows the experimental results of various algorithms. With the increase of the number of features, the experimental accuracy of each algorithm is improved. When the dimension of samples is reduced to nine, the test accuracy of the SDDA algorithm is higher than the maximum accuracy of other algorithms. Under the condition of the same number of features, the experimental results of SDDA are better than LDA and PCA in both minimum distance and nearest neighbor measurements. FSLDA is similar to CLDA, while the PCA method has the worst performance because it ignores category information. The results of our experiment demonstrate that the discriminant vectors of SDDA not only have principal component characteristics, but also have orthogonal and nearly statisically uncorrelated characteristics, so its performance is the best.

4.2. Experiment on the FERET Database. To validate the effect of SDDA on a dataset with large categories, we choose the FERET face database [50]. The FERET face database contains 1400 images of 200 persons. For each person, there are 7 images and whose file names contain the identification string "ba," "bj," "bk," "be," "bf," "bd," and "bg" to indicate the change of each image. Changes in posture ($\pm 15^{\circ}$ and $\pm 25^{\circ}$), illumination, and expression are all contained in the samples. In the experiment, the face in each original image is acquired according to the position of the eyes and then adjusted to 80×80 and preprocessed with histogram equalization. Figure 3 shows 7 images of one person in the database.

We also conducted two sets of experiments on the FERET human face database: images with file names contain the identification string ba, bd, be, and bf for training and the rest images for testing, and ba, be, bg, and bk for training and the rest for testing. We calculate the mean of the two results as the final results. Minimum distance and nearest neighbor measurement methods are both used.

Because the number of faces in the FERET database is large and the number of classes is 200, we use various algorithms to reduce the dimension of samples from 9 to 99. In order to reflect the trend of each algorithm when the dimension changes, we add line charts to reflect the accuracy change of each algorithm in different dimensions. Experimental results are shown in Figures 4 and 5 and Table 2. From the experimental results, we can see that the SDDA algorithm is much better than other algorithms when the dimension is the lowest. With the increase of the dimension, the test accuracy of the SDDA algorithm reaches the maximum effect quickly. When the dimension is 29, it has exceeded the maximum accuracy of all algorithms. When the dimension is 59, the test accuracy of the SDDA algorithm has reached its maximum. Due to the large difference between the training samples and the test samples, the training sample space cannot contain the test sample space well. Experimental results are shown in Figures 4 and 5. The CLDA method is statistically irrelevant, which fits the data too tight, resulting in the worst test results. The PCA method maintains certain test accuracy although the category

Table 1: The comparison of experimental results on the Yale face database.

Number of features	Minimum distance measurement				Nearest neighbor measurement			
	SDDA	CLDA	FSLDA	PCA	SDDA	CLDA	FSLDA	PCA
4	88.44	85.44	81.78	64.33	86.00	85.44	81.78	74.22
5	92.44	87.89	86.00	71.67	91.67	87.89	86.00	77.33
6	93.56	92.22	90.89	76.11	93.44	92.22	90.33	77.78
7	93.44	92.89	92.33	82.78	93.44	92.89	91.78	80.22
8	94.67	92.11	92.22	82.89	93.44	92.11	92.22	81.56
9	95.89	94.56	93.89	84.67	95.22	94.56	93.33	83.78
10	95.89	93.89	94.00	84.67	95.33	94.00	92.89	84.56
11	96.56	94.44	94.56	86.44	96.00	95.00	94.56	85.78

Figure 3: Seven images of one person in the FERET human face database.

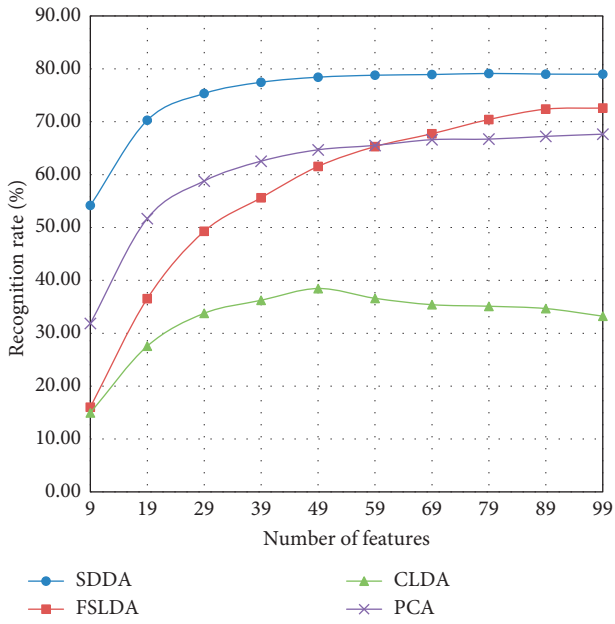

Figure 4: Experimental results on the FERET face database using minimum distance measurement.

information is not considered. The test result of FSLDA is second, which has orthogonal characteristics and good generalization ability. The proposed SDDA method is orthogonal and nearly statistically uncorrelated. The test result of SDDA is the best, and the recognition accuracy is significantly improved compared with the other three methods.

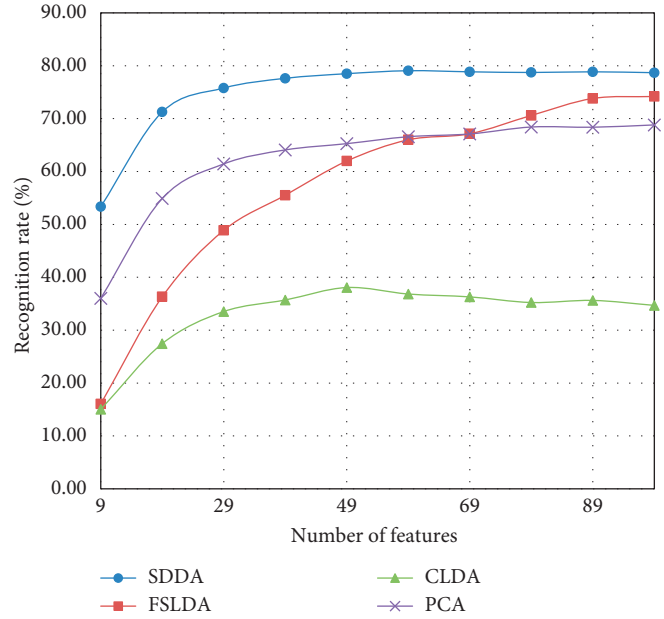

Figure 5: Experiment results on the FERET face database using nearest neighbor measurement.

4.3. Experiment on the UCI Multiple Features Dataset. In the previous two experiments, the number of samples per person is small. To further evaluate the performance of SDDA, the UCI multiple features dataset [51] is used for experiments. The UCI multiple features dataset contains six feature structures of handwritten numbers 0 to 9. Each feature structure is divided into 10 categories, each of which has 200 samples and a total of 2000 samples.

The 240-dimensional pixel average feature of the sample data is selected to reflect the dimensionality reduction effect of these algorithms. The minimum distance measurement method is used to verify the effectiveness of the feature. We also conduct two experiments.

The first experiment randomly selects 100 samples as the training set, and the remaining 100 samples are used for testing. The experiments are repeated 10 times, and the average results are given in Table 3. SDDA algorithm still has the best experimental effect. Because of the large number of samples selected for each class, CLDA algorithm's superiority is reflected. The effect has been greatly improved, ranking second, and FSLDA performance has been greatly reduced, ranking fourth. We can see that the results of

Table 2: The comparison of experimental results on the FERET face database.

Number of features	Minimum distance measurement				Nearest neighbor measurement			
	SDDA	CLDA	FSLDA	PCA	SDDA	CLDA	FSLDA	PCA
9	54.17	16.02	14.96	31.85	53.35	16.04	15.02	36.02
19	70.27	36.54	27.56	51.69	71.27	36.31	27.46	54.90
29	75.33	49.27	33.77	58.79	75.79	48.88	33.52	61.44
39	77.46	55.63	36.23	62.54	77.63	55.52	35.71	64.08
49	78.42	61.54	38.46	64.69	78.50	61.98	38.06	65.27
59	78.79	65.29	36.60	65.52	79.06	66.00	36.83	66.58
69	78.92	67.73	35.40	66.60	78.85	67.15	36.29	67.10
79	79.13	70.42	35.10	66.71	78.73	70.60	35.23	68.42
89	79.00	72.38	34.67	67.23	78.85	73.81	35.63	68.38
99	78.98	72.58	33.23	67.65	78.69	74.19	34.69	68.81

Table 3: Experimental results on the UCI multiple features dataset.

Number of features	SDDA	CLDA	FSLDA	PCA
1	45.58	41.75	42.38	39.10
2	73.43	69.7	47.65	57.01
3	80.61	82.52	50.28	68.44
4	87.42	89.49	57.80	75.27
5	91.28	92.16	67.47	77.39
6	93.59	93.40	73.94	82.18
7	94.38	94.21	79.35	85.89
8	94.64	94.75	82.44	87.93
9	94.81	94.79	84.56	88.54

SDDA and CLDA are similar, SDDA is significantly better than PCA and FSLDA, and the performance of PCA is better than that of FSLDA.

In the second experiment, only 20 samples were selected from each class as the training set. The experiment was repeated 10 times, and the average results are obtained. Figure 6 is the line chart of the results with the dimension of the extracted feature on the horizontal and the percentage of test accuracy on the vertical axis. As the line chart shows, the results of CLDA are the worst and the performance declined greatly. The experimental results of PCA and FSLDA are similar, and the performance of SDDA is still the best.

The experiment on the UCI multiple features dataset shows that the dispersion of the samples is large when the number of training samples of each class is large, so that the accuracy of the identification method with statistical uncorrelation characteristics is obviously superior to that of the orthogonal one. That is to say, in the case of larger samples, the performance of CLDA is much better than that of FSLDA. The performance of FSLDA is significantly better than that of CLDA when the number of training samples of each class is small. With the orthogonal characteristics and nearly statisically uncorrelated characteristics, the discriminant vectors of SDDA maintain the best performance regardless of the number of training samples. As can be seen from Figure 6, the accuracy of SDDA is higher than the comparison algorithms under the same number of features, and the curve shows a smooth upward trend. The increasing gradient of the accuracy is decreasing, which means the increase rate is large at first, while as the number of features increases, the increase rate gradually decreases. The proposed method can obtain superior performance under a small number of features, and the curve is rising smoothly, which means the overall performance of SDDA is stable and reliable.

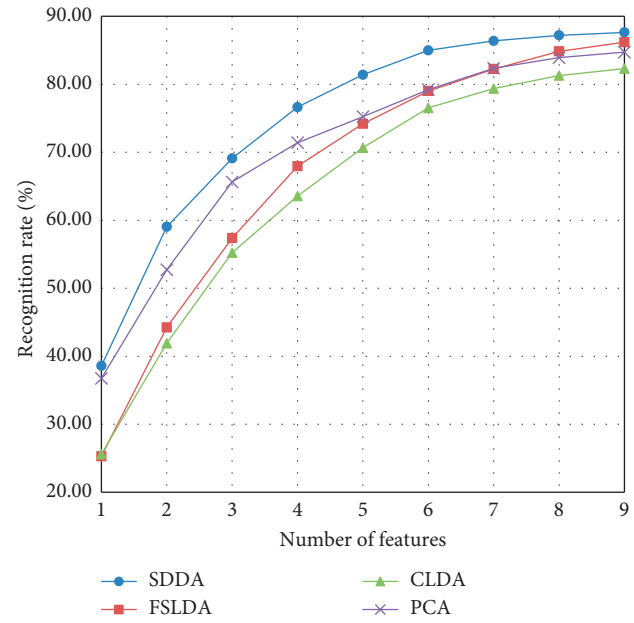

Figure 6: Experimental results on the UCI multiple features dataset.

Analyzing the experimental results on different databases synthetically, we can conclude that the SDDA method has the best performance in all databases, while the two basic methods of LDA have unstable performance. In contrast, the SDDA method has stronger adaptability than CLDA and

FSLDA, which not only retains the principal component characteristics, but also overcomes the shortcomings of CLDA and FSLDA, and can extract more outstanding identification features. In addition, under the same sample conditions, CLDA has the fastest training speed, while FSLDA and SDDA have similar training speed. The training speed of CLDA is about 1.5 times that of SDDA and FSLDA. When testing, all algorithms have the same test speed because they use the same dimension projection vector.

5. Conclusions

For a large number of collected high-dimensional data, the proposed method can effectively reduce information overload and improve data transmission and processing, so the importance of this study is more prominent. In view of the shortcomings of the two basic methods of LDA, the paper proposes a similar distribution discriminant analysis method (SDDA) and presents the solutions of the optimal discriminant vectors. The optimal discriminant vectors are mutually orthogonal and nearly statistically uncorrelated. The proposed SDDA method in this manuscript mainly aims at two basic methods of LDA. One is FSLDA whose projection vectors are orthogonal, and the other is CLDA which is statistically uncorrelated after dimensionality reduction. The performance of the two algorithms is different under different data. The FSLDA with orthogonal characteristics has stronger generalization ability, but with the increase of training sample size, the performance of statistically unrelated CLDA is improved. Taking both the two characteristics into consideration, SDDA performs well regardless of sample size. SDDA actually takes advantage of both PCA and LDA methods to maintain optimal performance and better adaptability in each experiment. A large number of experiments on the Yale face database, FERET face database, and UCI handwritten digits multiple features database confirm that SDDA is a more effective and adaptable dimensionality reduction method, which can extract better identification feature than CLDA, FSLDA, and PCA. Many theories and applications based on LDA can also be extended on the basis of the method proposed in this paper.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was funded by the National Natural Science Foundation of China (grant no. 61503329) and Prospective Joint Research Project of Jiangsu Province (grant no. BY201506-01).

References

[1] K. Fukunaga and W. L. G. Koontz, "Representation of random processes using the finite Karhunen-Loève expansion," *Information and Control*, vol. 16, no. 1, pp. 85–101, 1970.

[2] T. Y. Young, "The reliability of linear feature extractors," *IEEE Transactions on Computers*, vol. C-20, no. 9, pp. 967–971, 1971.

[3] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, no. 2, pp. 178–188, 1936.

[4] S. S. Wilks, *Mathematical Statistics*, Wiley, New York, NY, USA, 1962.

[5] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification and Scene Analysis*, Wiley, New York, NY, USA, 1973.

[6] D. H. Foley and J. W. Sammon, "An optimal set of discriminant vectors," *IEEE Transactions on Computers*, vol. 24, no. 3, pp. 281–289, 1975.

[7] K. Delac, M. Grgic, and S. Grgic, "Independent comparative study of PCA, ICA, and LDA on the FERET data set," *International Journal of Imaging Systems and Technology*, vol. 15, no. 5, pp. 252–260, 2005.

[8] A. B. Musa, "A comparison of ℓ1-regularizion, PCA, KPCA and ICA for dimensionality reduction in logistic regression," *International Journal of Machine Learning and Cybernetics*, vol. 5, no. 6, pp. 861–873, 2014.

[9] S. Bernhard, S. Alexander, and M. Klaus-Robert, "Nonlinear component analysis as a Kernel Eigenvalue problem," *Neural Computation*, vol. 10, no. 5, pp. 1299–1319, 1998.

[10] M. Klaus-Rober, M. Sebastian, R. Gunnar, T. Koji, and S. Bernhard, "An introduction to Kernel-based learning algorithms," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 12, no. 2, pp. 181–201, 2001.

[11] G. Baudat and F. Anouar, "Generalized discriminant analysis using a Kernel approach," *Neural Computation*, vol. 12, no. 10, pp. 2385–2404, 2000.

[12] J. Yang, A. F. Frangi, J. Y. Yang, D. Zhang, and Z. Jin, "KPCA plus LDA: a complete Kernel Fisher discriminant framework for feature extraction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 2, pp. 230–244, 2005.

[13] A. Papaioannou and S. Zafeiriou, "Principal component analysis with complex Kernel: the widely linear model," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 9, pp. 1719–1726, 2014.

[14] Z. Lai, Y. Xu, Q. Chen, J. Yang, and D. Zhang, "Multilinear sparse principal component analysis," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 10, pp. 1942–1950, 2014.

[15] R. He, B. G. Hu, W. S. Zheng, and X.-W. Kong, "Robust principal component analysis based on maximum correntropy criterion," *IEEE Transactions on Image Processing*, vol. 20, no. 6, pp. 1485–1494, 2011.

[16] X. Han, "Nonnegative principal component analysis for cancer molecular pattern discovery," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 7, no. 3, pp. 537–549, 2010.

[17] M. Asteris, D. S. Papailiopoulos, and G. N. Karystinos, "The sparse principal component of a constant-rank matrix," *IEEE Transactions on Information Theory*, vol. 60, no. 4, pp. 2281–2290, 2014.

[18] R. Wang, F. Nie, X. Yang, F. Gao, and M. Yao, "Robust 2DPCA with non-greedy L1-Norm maximization for image Analysis," *IEEE Transactions on Cybernetics*, vol. 45, no. 5, pp. 1108–1112, 2015.

[19] F. Han and H. Liu, "High dimensional semiparametric scale-invariant principal component analysis," *IEEE Transactions*

on Pattern Analysis and Machine Intelligence, vol. 36, no. 10, pp. 2016–2032, 2014.

[20] N. Kwak, "Principal component analysis based on L1-Norm maximization," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 30, no. 9, pp. 1672–1680, 2008.

[21] Y. Yan, E. Ricci, R. Subramanian, G. Liu, and N. Sebe, "Multitask linear discriminant analysis for view invariant action recognition," *IEEE Transactions on Image Processing*, vol. 23, no. 12, pp. 5599–5611, 2014.

[22] H. Wang, X. Lu, Z. Hu, and W. Zheng, "Fisher discriminant analysis with L1-Norm," *IEEE Transactions on Cybernetics*, vol. 44, no. 6, pp. 828–842, 2014.

[23] R. Saeidi, R. F. Astudillo, and D. Kolossa, "Uncertain LDA: including observation uncertainties in discriminative transforms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 7, pp. 1479–1488, 2016.

[24] J. Wang, "Generalized 2-D principal component analysis by Lp-Norm for image analysis," *IEEE Transactions on Cybernetics*, vol. 46, no. 3, pp. 792–803, 2016.

[25] W. Zheng, Z. Lin, and H. Wang, "L1-Norm Kernel discriminant analysis via Bayes error bound optimization for robust feature extraction," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 4, pp. 793–805, 2014.

[26] X. Zhang, D. Chu, and R. C. E. Tan, "Sparse uncorrelated linear discriminant analysis for undersampled problems," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 7, pp. 1469–1485, 2017.

[27] D. Chu, L. Z. Liao, M. K. P. Ng, and X. Wang, "Incremental linear discriminant analysis: a fast algorithm and comparisons," *IEEE Transactions on Neural Networks & Learning Systems*, vol. 26, no. 11, pp. 2716–2735, 2015.

[28] A. M. Martinez and A. C. Kak, "PCA versus LDA," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 228–233, 2001.

[29] B. A. Draper, J. R. Beveridge, G. H. Givens, and K. A. She, "A nonparametric statistical comparison of principal component and linear discriminant subspaces for face recognition," in *Proceedings of the CVPR 2001 (CVPR) -2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 535–542, Kauai, HI, USA, December 2001.

[30] P. Samal and P. A. Iyengar, "Automatic recognition and analysis of human faces and facial expressions: a survey," *Pattern Recognition*, vol. 25, no. 1, pp. 65–77, 1992.

[31] D. L. Swets and J. Weng, "Using discriminant eigenfeatures for image retrieval," *IEEE Trans Pattern Analysis and Machine Intelligence*, vol. 18, no. 8, pp. 831–836, 1996.

[32] W. Zhao, R. Chelllappa, A. Rosenfeld, and P. J. Phillips, "Face recognition: a literature survey," *ACM Computing Surveys*, vol. 35, no. 24, pp. 399–459, 2003.

[33] J. Yang, D. Zhang, J. Y. Yang, and A. F. Frangi, "Two-dimensional PCA: a new approach to appearance-based face representation and recognitionl," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 1, pp. 131–137, 2004.

[34] W. X. Liu and N. N. Zheng, "Non-negative matrix factorization based methods for object recognition," *Pattern Recognition Letters*, vol. 26, no. 14, pp. 893–897, 2004.

[35] X. F. He, S. Yan, Y. Hu, P. Niyogi, and H. J. Zhang, "Face recognition using Laplacianfaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 328–340, 2005.

[36] J. Yang, L. Luo, J. J. Qian, Y. Tai, F. L. Zhang, and Y. Xu, "Nuclear norm based matrix regression with applications to face recognition with occlusion and illumination changes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 1, pp. 156–171, 2017.

[37] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriengman, "Eigenfaces vs. Fisherfaces: recognition using class special linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.

[38] L. F. Chen, H. Y. M. Liao, J. C. Ko, J.-C. Lin, and G.-J. Yu, "A new LDA-based face recognition system which can solve the small sample size problem," *Pattern Recognition*, vol. 33, no. 10, pp. 1713–1726, 2000.

[39] J. P. Ye and Q. Li, "A two-stage linear discriminant analysis via QR-decomposition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 6, pp. 929–941, 2005.

[40] H. T. Zhao and C. P. Yuen, "Incremental linear discriminant analysis for face recognition," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 38, no. 1, pp. 210–221, 2008.

[41] J. Zhao, L. Shi, and J. Zhu, "Two-stage regularized linear discriminant analysis for 2-D data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 8, pp. 1669–1681, 2015.

[42] Y. Liu, Q. Gao, S. Miao, X. Gao, F. Nie, and Y. Li, "A non-greedy algorithm for L1-norm LDA," *IEEE Transactions on Image Processing*, vol. 26, no. 2, pp. 684–695, 2017.

[43] J. Duchene and S. Leclercq, "An optimal Transformation for discriminant and principal component analysis," *IEEE Trans Pattern Analysis and Machine Intelligence*, vol. 10, no. 6, pp. 978–983, 1988.

[44] Z. Jin, J. Y. Yang, Z. S. Hu, and Z. Lou, "Face Recognition based on uncorrelated discriminant transformation," *Pattern Recognition*, vol. 34, no. 7, pp. 1405–1416, 2001.

[45] Z. Jin, J. Y. Yang, Z. M. Tang, and Z. S. Hu, "A theorem on uncorrelated optimal discriminant vectors," *Pattern Recognition*, vol. 34, no. 10, pp. 2041–2047, 2001.

[46] T. Okada and S. Tomita, "An optimal orthogonal system for discriminant analysis," *Pattern Recognition*, vol. 18, no. 2, pp. 139–144, 1985.

[47] Y. Hamamoto, Y. Matsuura, T. Kanaoka, and S. Tomita, "A note on the orthonormal discriminant vector method for feature extraction," *Pattern Recognition*, vol. 24, no. 7, pp. 681–684, 1991.

[48] J. Yang, J. Y. Yang, and D. Zhang, "What's wrong with Fisher criterion," *Pattern Recognition*, vol. 35, no. 11, pp. 2665–2668, 2002.

[49] The Yale Face Database, http://cvc.cs.yale.edu/cvc/projects/yalefaces/yalefaces.html.

[50] The FERET Face Database, https://www.nist.gov/programs-projects/face-recognition-technology-feret.

[51] UCI Multiple Features Data Set, http://archive.ics.uci.edu/ml/datasets/Multiple+Features.