PAPER

Dimensionality Reduction by Locally Linear Discriminant Analysis for Handwritten Chinese Character Recognition

Xue GAO^{†a)}, Jinzhi GUO[†], *Nonmembers*, and Lianwen JIN[†], *Member*

SUMMARY Linear Discriminant Analysis (LDA) is one of the most popular dimensionality reduction techniques in existing handwritten Chinese character (HCC) recognition systems. However, when used for unconstrained handwritten Chinese character recognition, the traditional LDA algorithm is prone to two problems, namely, the class separation problem and multimodal sample distributions. To deal with these problems, we propose a new locally linear discriminant analysis (LLDA) method for handwritten Chinese character recognition. Our algorithm operates as follows. (1) Using the clustering algorithm, find clusters for the samples of each class. (2) Find the nearest neighboring clusters from the remaining classes for each cluster of one class. Then, use the corresponding cluster means to compute the between-class scatter matrix in LDA while keeping the within-class scatter matrix unchanged. (3) Finally, apply feature vector normalization to further improve the class separation problem. A series of experiments on both the HCL2000 and CASIA Chinese character handwriting databases show that our method can effectively improve recognition performance, with a reduction in error rate of 28.7% (HCL2000) and 16.7% (CASIA) compared with the traditional LDA method. Our algorithm also outperforms DLA (Discriminative Locality Alignment, one of the representative manifold learning-based dimensionality reduction algorithms proposed recently). Large-set handwritten Chinese character recognition experiments also verified the effectiveness of our proposed approach.

key words: character recognition, dimensionality reduction, linear discriminant analysis, handwritten Chinese character recognition, machine learning

1. Introduction

The recognition of handwritten Chinese characters (HCC) is a very active research area in pattern recognition. Although many studies have been conducted, it still remains a challenging problem since the recognition accuracy of freely handwritten Chinese characters is insufficient [1]. The public recognition contest organized in CCPR2010 [2] indicated that the best system can only achieve 89.99% recognition accuracy for offline handwritten Chinese characters of the GB2312-80 level-1 set. To improve the recognition performance, much effort is needed to develop new algorithms covering all aspects of HCC recognition, for example, pattern description, classification, and so on. In this paper, we focus on refining the pattern description of handwritten Chinese characters; more specially, we study the feature transform and dimensionality reduction to improve the discriminant ability of the pattern representation. Since most feature extraction methods for handwritten Chinese character

recognition have been designed intuitively, dimensionality reduction can be regarded as an optimization process to remove redundant information contained in the extracted raw features. Thus, it generally results in several benefits including less storage, faster recognition, and higher recognition accuracy.

Typical dimensionality reduction algorithms proposed in the field of machine learning can be divided into two groups [3]: conventional linear dimensionality reduction algorithms and manifold learning-based algorithms. The popular conventional linear dimensionality reduction algorithms include principal component analysis (PCA) and linear discriminant analysis (LDA, also known as Fisher discriminant analysis — FDA). Given the pattern representation in high-dimensional space, PCA tries to find a subspace whose basis vectors correspond to the maximum-variance directions in the original space, whereas LDA searches for those vectors in the underlying space that best discriminate among classes. Since supervised information of class labels is used in the training stage, LDA outperforms PCA in most cases, except when the training data set is small [4]. In the field of handwritten Chinese character recognition, LDA has been proven to be very effective and has even become a standard configuration in most handwritten Chinese character recognition systems [2], [5], [6]. LDA was first proposed by Fisher [7] for binary classification and subsequently extended by Rao [8] to multi-class classification, in which pattern features are transformed from a high-dimensional space to a lower-dimensional subspace by a linear projection matrix. However, when LDA is applied in multi-class classification tasks, two typical problems may occur. One is the *class separation* problem [9], [10], that is, classes that are close together in the original feature space, may be merged in a lower-dimensional subspace, thus decreasing the recognition accuracy. The reason for this is that pairs of classes with large distances dominate the eigenvalue decomposition in LDA, as can be seen from the decomposition of the multi-class Fisher criterion (Sect. 2). As a result, the directions obtained by LDA are not optimal for discriminating among neighboring classes. Another problem in traditional LDA is that the sample distributions can be *multimodal*. As pointed out by Fukunaga [11], the Fisher criterion is a good measure when the distributions are unimodal and separated by a scatter of class means. However, if the distributions are multimodal, the Fisher criterion is not a good measure of class separability. Figure 1 shows some examples of multimodal distribution, where M_1 , M_2 are the class means of

Manuscript received March 2, 2012.

Manuscript revised April 27, 2012.

[†]The authors are with the School of Electronic and Information Engineering, South China University of Technology, Guangzhou, 510641, China.

a) E-mail: xuegao@scut.edu.cn

DOI: 10.1587/transinf.E95.D.2533

Fig. 1 Examples of multimodal data distributions [11].

two classes ω_1 and ω_2 , respectively; $M_0 = p_1 M_1 + p_2 M_2$ is the overall mean, where p_1, p_2 are the priori probability of classes ω_1 and ω_2 . If the distributions are multimodal and share the same mean as in Fig. 1 (a), there is no scatter of M_1 and M_2 around M_0 . For this case, we can find clusters and use the cluster means instead of class means to measure the scatter. The same is true for the case of Fig. 1 (b). This case can be better handled by finding three clusters for each class. For unconstrained handwritten Chinese character samples, the different styles of handwriting distortions can actually lead to multimodal sample distributions.

To solve the class separation problem, Tao et al. proposed maximizing the geometric mean of the divergences instead of the arithmetic mean in finding the projection matrix [9]. However, this solution cannot easily be obtained by the eigenvalue decomposition procedure as in LDA. Moreover, the training procedure is rather time-consuming, especially in the case of a large training data set for handwritten Chinese character recognition, since computing the projection matrix requires an iterative optimization procedure instead of the eigenvalue decomposition in LDA. On the other hand, in Tao's method, it is still possible that large divergences dominate the computation of the projection matrix. Loog et al. [10] proposed a weighted pairwise Fisher criterion that weights the contributions of every class pair according to the Euclidian distance of the corresponding class means.

On the other hand, manifold learning-based dimensionality reduction methods have received extensive attention in recent years [3]. Here, we review some related works that offer linear mappings for data points throughout the entire feature space. The linearization extension is commonly more efficient in computation and thus more practical for handwritten Chinese character recognition of large training data sets. A more exhaustive survey of the literature can be found elsewhere [3], [12]. Locality preserving projections (LPP) [13] is a typical linear dimensionality reduction technique that preserves local relationships within the data set and uncovers its essential manifold structure. The projection ensures that neighboring data pairs in the original space are close together in the subspace. However, LPP is an unsupervised learning method that does not consider class label information. By combining the ideas of LDA and LPP, Sugiyama [14] proposed a local Fisher discriminant analysis (LFDA) method, which weights data pairs from identical classes in terms of affinity in the original feature space when

computing the scatter matrices in LDA. LFDA aims to compensate for the drawback of LDA and LPP in the case of multimodal data distributions. Based on graph embedding and extensions, Yan et al. [12] unified most of the existing dimensionality reduction algorithms within a common framework, and proposed a dimensionality reduction algorithm called marginal Fisher analysis (MFA) by using the graph embedding framework to overcome the limitation of LDA. MFA uses an intrinsic graph characterizing the intra-class compactness where each sample is connected to its k_1 nearest neighbors in the same class, and a penalty graph characterizing the interclass separability where marginal data pairs of different classes are connected. Patch alignment is another general framework of dimensionality reduction developed more recently by Zhang et al. [3], which can also be used to explain and understand the existing dimensionality reduction algorithms. Under this patch alignment framework, a dimensionality reduction algorithm called discriminative locality alignment (DLA) was proposed for face recognition to attack the distribution nonlinearity of sample representation. DLA maximizes the distances between arbitrary samples and its k_2 nearest neighbors from different classes and minimizes the distances between each sample and its k_1 neighbors from an identical class simultaneously in the embedding space. The face recognition experiments have shown that DLA is superior to MFA and most other manifold learning-based algorithms [3]. Meanwhile, theoretical analysis and experimental results of the above-mentioned manifold learning-based algorithms have indicated that the neighboring samples and those around class margins are more useful in discovering the nonlinear structure hidden in high-dimensional data. Following this observation, if we reformulate the Fisher criterion and ignore the far apart samples in computing the scatter matrices in LDA, we can expect to improve the recognition performance in the classification tasks, and the class separation problem in LDA can also be alleviated to some extent. However, compared with the use of class means in maximizing the class separability in LDA, the aforementioned algorithms such as MFA and DLA also lead to greater computation in training procedures, especially for large-set classification tasks with large training data sets, since these algorithms need to search for the nearest neighboring samples of each sample in the whole training set. For multimodal sample distributions, we can use the clustering algorithm to find the clusters for each class and treat each cluster as a new class in LDA. Based on the above observations, under the framework of the Fisher criterion, we propose a new locally linear discriminant analysis (LLDA) method for handwritten Chinese character recognition. The algorithm operates as follows. (1) Use the clustering algorithm to find clusters for each class. (2) Find the nearest neighboring clusters for each cluster and use the cluster means to compute the between-class scatter in LDA, while keeping the within-class scatter unchanged. (3) Finally, apply vector normalization to further improve the class separation problem. Our method can effectively utilize the local structure information embedded

in high-dimensional data. It can also be expected to improve the class separation problem and to work with multimodal data in handwritten Chinese character recognition.

The rest of this paper is organized as follows. Section 2 gives a brief description of the LDA algorithm and details of our proposed locally linear discriminant analysis method. Section 3 describes the underlying feature extraction and classification methods. In Sect. 4, we present a series of experimental results on two typical handwriting databases to show the effectiveness of the proposed method. Section 5 concludes the paper.

2. Dimensionality Reduction Algorithms

2.1 Fisher Linear Discriminant Analysis

Fisher linear discriminant analysis (LDA) is one of the most popular linear dimensionality reduction techniques in machine vision and pattern recognition, and has been successfully applied in handwritten Chinese character recognition [5], [6]. The aim of LDA is to search for an embedding subspace such that most of the discriminant information contained in high-dimensional data is preserved. LDA is supervised since the class label information is taken into account. More formally, according to [11], [15], given a set of training measurements from c -class patterns and a linear mapping $y = W^T x$ from the original feature space to the embedding subspace, LDA seeks to find an optimal transformation matrix $W^* \in R^{d \times r}$ ($r < d$) from d to the lower r -dimensional space such that the between-class scatter is maximized, while the within-class scatter is minimized.

For a c -class classification task, let us assume that the training set $X = \{x_{ij}\}$ consists of n samples from c classes. For the i -th class, there are n_i samples x_{ij} ($1 \leq j \leq n_i$) with a class mean of $\mu_i = (1/n_i) \sum_{j=1}^{n_i} x_{ij}$ and $n = \sum_{i=1}^c n_i$. The mean of all training samples is given by

$$\mu = \frac{1}{n} \sum_{i=1}^c \sum_{j=1}^{n_i} x_{ij} \quad (1)$$

LDA defines two measures to compute the transformation matrix W^* ; one is the between-class scatter matrix S_b , and the other is the within-class scatter matrix S_w given by

$$S_b = \frac{1}{n} \sum_{i=1}^c n_i (\mu_i - \mu)(\mu_i - \mu)^T \quad (2)$$

$$S_w = \frac{1}{n} \sum_{i=1}^c \sum_{j=1}^{n_i} (x_{ij} - \mu_i)(x_{ij} - \mu_i)^T \quad (3)$$

The transformation matrix W^* of LDA is defined as follows:

$$W^* = \arg \max_W \text{tr} \left((W^T S_w W)^{-1} (W^T S_b W) \right) \quad (4)$$

In the optimization criterion in (4), it is implicitly assumed that within-class scatter S_w is invertible. This assumption can be generally guaranteed for handwritten Chinese character recognition with enough training samples.

And the column vectors of W^* are composed of eigenvectors of $S_w^{-1} S_b$ corresponding to the largest r eigenvalues.

2.2 Locally Linear Discriminant Analysis

In this section, we reformulate the Fisher optimization criterion and show that it is mathematically equivalent to maximizing the arithmetic average of the squared Mahalanobis distance between all pairs of class mean vectors in the embedding subspace, under the assumption that the Gaussian distributions for the different classes all have the same covariance matrix. Then, we give our proposed locally linear discriminant analysis (LLDA) algorithm by utilizing the local structure information in the data, to attack the class separation problem in LDA and the multimodal data distribution.

For any pair of class mean vectors μ_i and μ_j under the assumption of equal covariance matrices ($\Sigma_i = \Sigma_j = \Sigma$), the squared Mahalanobis distance in the embedding subspace is given as

$$\Delta_{ij} = (W^T \mu_i - W^T \mu_j)^T (W^T \Sigma W)^{-1} (W^T \mu_i - W^T \mu_j) \quad (5)$$

Since $W^T \Sigma W$ is symmetrical and invertible, (5) can be rewritten as follows:

$$\begin{aligned} \Delta_{ij} &= \\ &= \text{tr} \left((W^T \Sigma W)^{-\frac{1}{2}} (W^T \mu_i - W^T \mu_j) (W^T \mu_i - W^T \mu_j)^T (W^T \Sigma W)^{-\frac{1}{2}} \right) \\ &= \text{tr} \left((W^T \Sigma W)^{-\frac{1}{2}} W^T (\mu_i - \mu_j) (\mu_i - \mu_j)^T W (W^T \Sigma W)^{-\frac{1}{2}} \right) \end{aligned} \quad (6)$$

According to the properties of matrix trace, $\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA})$ if $\mathbf{AB}$ is a square matrix, (6) can be reformulated as

$$\begin{aligned} \Delta_{ij} &= \\ &= \text{tr} \left((W^T \Sigma W)^{-\frac{1}{2}} (W^T \Sigma W)^{-\frac{1}{2}} W^T (\mu_i - \mu_j) (\mu_i - \mu_j)^T W \right) \\ &= \text{tr} \left((W^T \Sigma W)^{-1} W^T (\mu_i - \mu_j) (\mu_i - \mu_j)^T W \right) \end{aligned} \quad (7)$$

Then, we have the arithmetic average of the squared Mahalanobis distance over all pairs of class mean vectors in the subspace:

$$\begin{aligned} J(W) &= \sum_{1 \leq i, j \leq c} p_i p_j \Delta_{ij} = \\ &= \text{tr} \left((W^T \Sigma W)^{-1} W^T \left(\sum_{i=1}^c \sum_{j=1}^c p_i p_j (\mu_i - \mu_j) (\mu_i - \mu_j)^T \right) W \right) \end{aligned} \quad (8)$$

where p_i ($1 \leq i \leq c$) is the prior probability of the i -th class (usually we can set $p_i = n_i/n$ or simply set $p_i = 1/c$ as is done in our experiments).

It has been proved [10] that the between-class scatter S_b can be rewritten as

$$S_b = 0.5 \sum_{i=1}^c \sum_{j=1}^c p_i p_j (\mu_i - \mu_j) (\mu_i - \mu_j)^T \quad (9)$$

And the covariance matrix of all samples is given by $\Sigma =$

Fig. 2 Illustration of the importance of different classes in discriminating one class from the remaining classes in LDA.

$S_t = S_b + S_w$ (see [11]). Therefore, we have

$$\begin{aligned} \arg \max_{\mathbf{W}} J(\mathbf{W}) &= \arg \max_{\mathbf{W}} \sum_{1 \leq i, j \leq c} p_i p_j \Delta_{ij} \\ &= \arg \max_{\mathbf{W}} \text{tr} \left((\mathbf{W}^T S_t \mathbf{W})^{-1} \mathbf{W}^T S_b \mathbf{W} \right) \\ &= \arg \max_{\mathbf{W}} \text{tr} \left((\mathbf{W}^T S_w \mathbf{W})^{-1} \mathbf{W}^T S_b \mathbf{W} \right) \end{aligned} \quad (10)$$

It can be seen from (10) that LDA is mathematically equivalent to searching for a linear transformation matrix that maximizes the arithmetic average of the squared Mahalanobis distance over any pairs of class mean vectors assuming that the Gaussian distributions for the different classes all have the same covariance matrix. Obviously, this can lead to the *class separation* problem as described in [9], [10], that is, pairs of classes with large distances will dominate the eigenvalue decomposition in LDA, and classes, which are close in the original feature space, may be merged in the subspace. On the other hand, for a c -class classification problem, neighboring classes are more important than far distant classes in discriminating one class from the remaining classes. Consider the case of discriminating one class (C_1) from the remaining classes in a two-dimensional feature space as shown in Fig. 2, where classes C_2 and C_3 are two neighboring classes of class C_1 , and C_4 is a class far distant from C_1 . We give an intuitive explanation of the basic idea. For simplicity of illustration, we also use $C_1, \dots, C_4$ to represent the class means of the corresponding classes. It can be seen that, if we only use C_1 and its neighboring classes C_2 and C_3 to compute the between-class scatter matrix S_b in LDA, we may get the projection direction L_1 using the optimization criterion (4), to be able to discriminate effectively class C_1 from classes C_2 and C_3 , where the projection direction refers to the direction along which the sample features are projected in LDA based feature mapping $y = W^T x$ from the original feature space to the embedding subspace. And this direction can also be used in discriminating effectively between C_1 and C_4 . Thus, the far distant class C_4 can be ignored in computing the between-class scatter S_b , thereby alleviating the *class separation* problem to a certain extent. On the contrary, if all of C_2, C_3 and C_4 are used in computing S_b , the projection direction L_2 may be found due

to the large distance between pair of classes C_1 and C_4 . As a result, it will lead to the merge of classes C_1 and C_2 after projection, thus the lower discriminative ability between C_1 and remaining classes. It should be noted that, we do not assume that the different classes are linear separable as in the traditional LDA, since only class means are used in computing the between-class scatter S_b and the optimal projection directions found will try to maximize the distances between classes. The number of neighboring classes used can affect the results of found projection directions. In our case, the best number of neighboring classes will be chosen by evaluating the corresponding recognition performance.

Based on the aforementioned observation, we propose using only the neighboring classes in computing the optimization objective function in (10). Mathematically, this is equivalent to modifying the between-class scatter matrix S_b in LDA according to (11-12), with the new algorithm known as LLDA in this paper. Assume Ω_i is the set of neighboring class labels of class i , the modified between-class scatter S'_b can be written as follows:

$$S'_b = \sum_{i=1}^c p_i \delta'_{ij} \quad (11)$$

$$\delta'_{ij} = \sum_{j \in \Omega_i} p_j (\mu_i - \mu_j)(\mu_i - \mu_j)^T \quad (12)$$

A similar idea has been used for face recognition [3] (called Discriminative Locality Alignment, DLA). For each given sample, some neighboring samples are selected from the identical class and from different classes. Then, in a low-dimensional space, the distances between each given sample and its neighboring samples from the identical class are made as small as possible, while the distances between each given sample and its neighboring samples from different classes are made as large as possible. As one representative manifold learning-based dimensionality reduction algorithm proposed more recently, DLA has been proven experimentally to be more effective than most existing algorithms on face databases. In our experiments, we compare the performance of our proposed method with DLA, showing the effectiveness of our method.

In addition, traditional LDA is based on the assumption that the data distribution of each class is unimodal. However, from our observations of unconstrained handwritten Chinese character samples, multimodal cases for some classes do exist in practice, for example, due to writing deformation. Therefore, we propose to group the training samples into different clusters for each class by means of the clustering algorithm and then use the cluster centers instead of the class means to compute the between-class scatter matrix. Suppose that $i(k)$ ($1 \leq k \leq K$) is the k -th cluster label of class i , $\mu_{i(k)}$ is the mean of cluster $i(k)$, and $\Omega_{i(k)}$ is the set of neighboring cluster labels of cluster $i(k)$ found within the remaining classes of the i -th class. Then, the modified between-class scatter S''_b is expressed as

$$S''_b = \sum_{i=1}^c \delta''_{ij} \quad (13)$$

$$\delta''_{ij} = \sum_{1 \leq k \leq K} p_{i(k)} \sum_{j(l) \in \Omega_{i(k)}} p_{j(l)} (\mu_{i(k)} - \mu_{j(l)}) (\mu_{i(k)} - \mu_{j(l)})^T \quad (14)$$

In computing the $\Omega_{i(k)}$, we adopt the following strategy. First, we find all clusters $j(l)$ ($j \neq i$; $1 \leq l \leq K$) satisfying (15) as candidates, and then we search for the q nearest neighboring clusters from the candidates as $\Omega_{i(k)}$.

$$\|\mu_{i(k)} - \mu_{j(l)}\| = \min_t \|\mu_{i(t)} - \mu_{j(l)}\| \quad (15)$$

Finally, to improve further the class separation problem, a vector normalization technique is applied in our method. From (15), we can see that a large distance between the different cluster means may still dominate the computation of D''_{ij} and thus would affect the computation of optimal projection directions. For each pair of cluster means $\mu_{i(k)}$ and $\mu_{j(l)}$, vector normalization is carried out as follows:

$$\mu'_{i(k)} - \mu'_{j(l)} = (\mu_{i(k)} - \mu_{j(l)}) / \|\mu_{i(k)} - \mu_{j(l)}\| \quad (16)$$

Then, we have the modified between-class scatter matrix S'''_b given below, which replaces S_b in LDA for computing the linear transformation matrix, while the within-class scatter matrix S_w remains unchanged.

$$S'''_b = \sum_{i=1}^c \delta'''_{ij} \quad (17)$$

$$\delta'''_{ij} = \sum_{1 \leq k \leq K} p_{i(k)} \sum_{j(l) \in \Omega_{i(k)}} p_{j(l)} (\mu'_{i(k)} - \mu'_{j(l)}) (\mu'_{i(k)} - \mu'_{j(l)})^T \quad (18)$$

3. Feature Extraction and Classification

To evaluate the performance of our proposed dimensionality reduction method using locally linear discriminant analysis for handwritten Chinese character recognition, we used an 8-directional gradient feature [6], [16], one of the most effective and frequently used features, in our experiments. To compute the gradient features, two 3×3 Sobel operators (as shown in Fig. 3) were used to obtain the horizontal and vertical gradient at each pixel of the character image. The character image was then decomposed into a number of regions corresponding to D directions with an equal interval $2\pi/D$ (8 directions are most frequently used in the references; we also used this setting in our experiments), and the gradient vector of each pixel was decomposed into its two nearest directions in terms of a parallelogram, as illustrated in Fig. 3, where the gradient vector $\mathbf{g}$ was decomposed into $\mathbf{g}_0$ and $\mathbf{g}_1$.

Fig. 3 Sobel operators and the gradient vector decomposition.

In this way, D directional sub-pattern images can be derived from each character image.

To extract the gradient feature representation of handwritten Chinese characters, the global elastic meshing technique [17] was applied instead of the nonlinear shape normalization method to deal with the shape variation of handwritten characters. The global elastic meshing technique can be regarded as a combination of the nonlinear shape normalization and the uniformly grid generation, since the grid lines were drawn dynamically in terms of the line density as defined in the line-density based nonlinear normalization method (the line density definition by line interval [20] was used in our experiments). Our previous experimental results had shown that this method was more effective than the line-density based nonlinear normalization method. The operations were as follows. The character images were first divided into 8×8 sub-blocks through the global elastic meshing technique and the sub-blocks partition was applied to the 8 directional sub-patterns. Then, in one sub-block of each directional sub-pattern, one feature was computed by summation, thereby forming a $8 \times 8 \times 8 = 512$ dimensional gradient feature vector for each character image. Finally, our proposed locally linear discriminant analysis method was used to reduce the feature dimension for classification.

For the classification test, two different classification strategies were adopted to evaluate our LLDA method. One is the feature match technique based on the minimum Euclidean distance (hereafter referred to as **MED**). Another is the modified quadratic discriminant function (MQDF) [5], [18]. For a c -class classification task, let y represent the r -dimensional sample feature after the dimensionality reduction, $\tilde{\mu}_i$ be the mean vector of class ω_i ($i = 1, \dots, c$) in the lower r -dimensional subspace. The MED classifier used the sum-of-squared errors as the distance definition, as given in (19). For the MQDF classifier, the discriminant function MQDF2 [18] as given by (21) was used. The decision rules for both the classifiers are shown in (20) and (22), respectively.

$$\varepsilon(y, \omega_i) = \|y - \tilde{\mu}_i\|^2 \quad (19)$$

$$y \in \omega_j \text{ if } \varepsilon(y, \omega_j) = \min_{i=1, \dots, c} \varepsilon(y, \omega_i) \quad (20)$$

$$g(y, \omega_i) = \frac{1}{h} \left\{ \|y - \tilde{\mu}_i\|^2 - \sum_{j=1}^m \left(1 - \frac{\lambda_{ij}}{\lambda_{ij}}\right) [\varphi_{ij}^T (y - \tilde{\mu}_i)]^2 \right\} + \sum_{j=1}^m \log \lambda_{ij} + (r - m) \log h \quad (21)$$

$$y \in \omega_j \text{ if } g(y, \omega_j) = \min_{i=1, \dots, c} g(y, \omega_i) \quad (22)$$

where h is a class independent constant, m denotes the number of dominant eigenvectors, and λ_{ij} , φ_{ij} ($i = 1, \dots, c$; $j = 1, \dots, m$) are the j -th eigenvalue (ordered in decreasing order) and the corresponding eigenvector of the covariance matrix of class ω_i in r -dimensional subspace after feature dimensionality reduction. We empirically chose $m = 9$ (in the case of 100 Chinese characters) and $m = 12$ (in the case of 3755 Chinese characters) for our recognition experiments.

4. Experiments and Results

4.1 Experimental Setup

To evaluate the performance of our proposed LLDA for handwritten Chinese character recognition, a series of experiments were carried out for comparison with the traditional Fisher LDA and DLA. LDA is one of the most popular and efficient dimensionality reduction algorithms used in handwritten Chinese character recognition. DLA is one of the manifold-learning based dimensionality reduction algorithms proposed more recently and which has been shown to outperform most of the representative algorithms in face recognition. Two offline handwriting databases were used in our experiments. One is the China 863 National Handwriting Database HCL2000 released by Beijing University of Posts and Telecommunications [19], which is a popular database in handwritten Chinese character recognition. HCL2000 contains 1000 sets of isolated handwritten Chinese character samples written by different writers for each of 3755 Chinese characters, which were scanned with a resolution of 300 DPI and normalized into 64×64 binary images by linear normalization. HCL2000 is relatively less challenging since high accuracies can be obtained as indicated elsewhere [2], [6]. To keep our performance evaluations more reasonable, we removed some regular handwritten character samples from HCL2000 and selected a subset of 400 sets of character samples with the lowest recognition rates when using simpler distance classifier without reducing dimension of features. Thus, we can form a more challenging database for the recognition test. In our experiments, we used 300 sets as the training data and 100 sets as the test data.

The other offline handwriting database used is the CASIA-HWDB1.1 (also referred to as CASIA), which was released more recently by the Institute of Automation at the Chinese Academy of Sciences [2]. CASIA-HWDB1.1 contains 300 sets of isolated handwritten Chinese character samples written by different writers for each of 3755 Chinese characters, which were scanned with a resolution of 300 DPI and converted to gray character images of different sizes. Since the numbers of valid samples are less than 300 for several characters in CASIA database we have (CASIA-HWDB1.1), to keep the sample numbers for each character balanced, we selected 250 sets of samples as our evaluation database. In our experiments, we randomly selected 200 sets as the training data and 50 sets as the test data. Before feature extraction, the operations of binarization and linear normalization were applied, to convert the gray samples to 64×64 binary images. Figure 4 illustrates some samples from both Chinese character handwriting databases.

Before the training and recognition process, each handwritten Chinese character was represented by a 512-dimensional 8-directional gradient feature as described in Sect. 3. Furthermore, the k -means clustering algorithm was adopted to group the samples into different clusters for each

Fig. 4 Some samples from Chinese character handwriting databases: (a) HCL2000, (b) CASIA.

character. The best cluster number was determined experimentally. Then, the features with dimensionality reduction were used to train the classifiers. MED and MQDF2 classifiers were used in the performance comparison of different dimensionality reduction techniques.

4.2 Performance Comparison

In this section, having selected 100 categories of Chinese characters for extensive evaluation, we give the recognition performance of our proposed LLDA algorithm with respect to numbers of neighbors, numbers of clusters, and vector normalization as discussed in Sect. 2. Thereafter, we compare the best recognition rates obtained using different dimensionality reduction algorithms, i.e., our proposed LLDA, LDA, and DLA (we used DLA1 in our experiments; see [3] for details). The character feature dimension was reduced to 99 (set as $c - 1$, an upper bound on the dimension of the subspace in LDA) in all the experiments.

Two important parameters in our LLDA algorithm are the numbers of nearest neighbors and clusters. For DLA, the parameters include k_1 (the number of nearest neighbors from an identical class) and k_2 (the numbers of nearest neighbors from different classes). How to initialize these parameters is still an open problem as explained by Yan et al. [12]. In our experiments, we set these parameters empirically. For our proposed LLDA algorithm, we give the experimental results of the selected numbers of nearest neighbors from 1 to 10 at sampled intervals of 1, and from 15 to 20 at sampled intervals of 5, and the numbers of clusters from 1 to 5 at sampled intervals of 1, and from 10 to 25 at sampled intervals of 5. We also tested the recognition performance of our LLDA algorithm with the number of clusters equal to the number of samples for each character, i.e., each training sample is treated as a cluster center. For the DLA algorithm, k_1 was chosen from 1 to 10 and k_2 from 1 to 10 at

Fig. 5 Recognition performance of LLDA on HCL2000: the best recognition rates among different numbers of nearest neighbors for varying numbers of clusters of each class with and without vector normalization. (The asterisk * on the X-axis indicates that the cluster number is set equal to the number of training samples and thus each sample in the training set is treated as one cluster center. It is the same in Fig. 6.

Fig. 6 Recognition performance of LLDA on CASIA: the best recognition rates among different numbers of nearest neighbors for varying numbers of clusters of each class with and without vector normalization.

sampled intervals of 1 and from 15 to 20 at sampled intervals of 5. As our experiments and the results given by Zhang et al. [3] have shown, k_1 and k_2 are usually small to represent the local structure information, since larger values can lead to performance degradation after the recognition rate reaches its peak. Thus, this parameter setting is reasonable and time-saving. From the experiments conducted on both HCL2000 and CASIA, Figs. 5 and 6 show the best recognition rates for LLDA among the different numbers of nearest neighbors for varying numbers of clusters of each class with and without vector normalization.

From Figs. 5 and 6, it can be seen that the best recognition rate generally increases as the number of clusters increases. However, if the number of clusters is equal to the number of training samples and we set each training sample as a cluster, the recognition rate decreases. Meanwhile, whether or not vector normalization is applied, LLDA performs better than the traditional LDA in recognition performance.

By fixing the number of clusters to an arbitrary value, we can obtain recognition rate curves with respect to the number of neighbors. Figures 7–8 and 9–10 show the ex-

perimental results when fixing the number of clusters to 1 or 10 on HCL2000 and CASIA, respectively.

From Figs. 7–10, it can be seen that, for a fixed number of clusters, the different numbers of neighbors affect the classification results. Generally speaking, when the number of neighbors is large, the recognition performance deteriorates and there is a peak corresponding to a small number of neighbors.

In addition, from Figs. 5–10, we can see that the recognition rates of our proposed LLDA, when vector normalization is applied, are generally better than those without vector normalization.

In Table 1, we summarize the best recognition rates for LDA, DLA, and our proposed LLDA algorithms on both the HCL2000 and CASIA handwriting databases. We also show the recognition rates without reducing dimension (hereafter referred to as ORGFEA) on the same databases for comparison. In the experiments of ORGFEA, the original feature with $512 (= 8 \times 8 \times 8)$ dimension is used.

From Table 1, it can be seen that our proposed LLDA algorithm outperforms the traditional LDA in all cases and that the Error Rate reduction reaches 28.7% (HCL2000,

Fig. 7 Recognition rate versus number of neighbors for LLDA by fixing the cluster number to 1 on HCL2000.

Fig. 8 Recognition rate versus number of neighbors for LLDA by fixing the cluster number to 10 on HCL2000.

Fig. 9 Recognition rate versus number of neighbors for LLDA by fixing the cluster number to 1 on CASIA.

Table 1 Best recognition rates (%) for LDA, LLDA, and DLA. (For LLDA, the first and second numbers in the parentheses are the numbers of clusters and neighbors, respectively. For DLA, the first and second numbers in the parentheses are the numbers of neighbor samples from an identical class and different classes, respectively.)

	HCL2000		CASIA	
	MED	MQDF2	MED	MQDF2
ORGFEA	97.22	99.14	91.74	94.86
LDA	98.13	98.92	94.32	95.58
DLA	97.34 (5,4)	99.24 (8,6)	92.70 (9,9)	94.72 (9,9)
LLDA	98.54 (15,6)	99.23 (15,6)	95.08 (5,6)	96.32 (2,8)

MQDF2) and 16.7% (CASIA, MQDF2) compared with the traditional LDA. On the other hand, the DLA algorithm achieves a lower recognition performance, except for the case (HCL2000, MQDF2), and it performs even worse than the traditional LDA in some cases. It should be noted that CASIA is a more challenging database of unconstrained Chinese handwriting. For this database, our algorithm shows significant improvement compared with DLA. The experimental results verify the effectiveness of our proposed LLDA method for handwritten Chinese character recogni-

Fig. 10 Recognition rate versus number of neighbors for LLDA by fixing the cluster number to 10 on CASIA.

Fig. 11 Recognition rate versus reduced dimension on HCL2000.

Fig. 12 Recognition rate versus reduced dimension on CASIA.

tion.

In Figs. 11–12, we show the experimental results of recognition rates versus reduced dimensions on HCL2000 and CASIA, when using the parameter settings as in Table 1. Generally speaking, the lower the reduced dimension is, the lower recognition accuracy we may have. When the reduced dimension is up to a certain value, increasing the dimension can no longer improve recognition rate significantly. It can be seen that our proposed method outperforms the other algorithms.

Finally, based on the parameter settings of our algorithm as given in Table 1, we compared the recognition per-

formance of LDA and LLDA on 3755 categories of handwritten Chinese characters. Since DLA algorithm needs to build one local patch for each measurement in data set to achieve the part optimization and the whole alignment, for large-set classification tasks, the patch building process is too time-consuming to be practical. Therefore, we do not test the performance of DLA on 3755 categories of handwritten Chinese characters. In these experiments, the feature dimension was reduced to 196. In addition, we also tested the recognition rates without reducing dimension (referred to as ORGFEA) on the same database for comparison. The recognition results are given in Table 2.

Table 2 Recognition rates (%) for 3755 Chinese characters by LDA and LLDA on both HCL2000 and CASIA handwriting databases.

	HCL2000		CASIA	
	MED	MQDF2	MED	MQDF2
ORGFEA	83.68	93.54	69.51	80.04
LDA	88.43	94.92	77.37	84.47
LLDA	89.03	95.32	78.72	85.26

From Table 2, it can be seen that our proposed LLDA algorithm also performs better than the traditional LDA in large-set handwritten Chinese character recognition.

4.3 Discussion

Based on two representative Chinese character handwriting databases, we have reported a series of experimental results, as shown in Figs. 5–10 and Table 1. From these results, we make the following observations:

1. In contrast to the global property of traditional LDA, our LLDA algorithm utilizes the locally discriminative information in searching for the optimal subspace. This can better represent the nonlinear structure hidden in high-dimensional data and improve the class separation problem in LDA. Thus, LLDA outperforms LDA in terms of recognition accuracy. When the number of nearest neighbors used in LLDA reaches a certain value, the recognition rates decrease gradually. At the one extreme, when all neighbor centers from different classes are used, LLDA is similar to LDA, and recognition accuracy is obviously worse than the best results for LLDA as shown in Figs. 7 and 9.
2. By grouping the training samples of each class into different clusters and using the cluster centers for the subspace selection, LLDA can be used to attack multimodal distributions of data. On the other hand, since DLA works in terms of the nearest neighboring samples, some statistical information in the data distribution can be lost, especially in classification tasks with large training data sets. The experimental results in Table 1 confirm that DLA is not as good as LLDA in classification accuracy on either handwriting database. Moreover, it is even worse than LDA, although DLA has been shown to be superior to LDA in face recognition with small training data sets. Similarly, if the number of clusters is equal to the number of training samples, the recognition performance of LLDA also decreases greatly.
3. LLDA with feature vector normalization performs better than without normalization, as shown in Figs. 5–10, and the best recognition rates are obtained when conducting vector normalization. Vector normalization performs better since it effectively balances the weight factors of cluster pairs with different distances in learning the discriminative information. Thus, it can alleviate the class separation problem in traditional LDA to some extent and improve the recognition performance.

It should be noted that our proposed algorithm may behave slightly differently on the two separate handwriting databases HCL2000 and CASIA, due to their different writing styles of character samples, and the Error Rate reduction on CASIA is comparatively less than that of HCL2000. The samples in HCL2000 are relatively more regular as a result of the constraint in collecting that each character needs to be written in a square box. While the CASIA was collected by using Anoto pen (digital pen) and the samples is more freely-written. This may lead to the data distribution of character samples more complex in feature space. As a result, the separability of different classes may not be well characterized with the assumption of Gaussian distributions for each class in linear discriminant analysis algorithms. In the future, we will exploit the application of nonlinear techniques, for example, kernel-based nonlinear mapping, to further improve our algorithm.

5. Conclusions

To deal with the *class separation* problem and *multimodal* sample distributions, both significant issues in the traditional LDA method, we have proposed a new locally linear discriminant analysis method for handwritten Chinese character recognition in this paper. The main contributions of the proposed method are that we introduce cluster means instead of class means into the LDA, use the nearest neighboring cluster means to compute the between-class scatter matrix, and finally apply vector normalization to further improve the class separation problem. A series of experiments on both the HCL2000 and CASIA handwriting databases confirmed that our method can effectively improve the recognition performance, with a reduction in error rate of 28.7% (HCL2000) and 16.7% (CASIA) compared with the traditional LDA method. Our algorithm also outperforms DLA, except for the case (HCL2000, MQDF2); for the CASIA handwriting database, in particular, the improvement is more significant. Large-set handwritten Chinese character recognition experiments also verified the effectiveness of our algorithm. However, our LLDA algorithm is still based on the assumption that the Gaussian distributions for the different classes have the same covariance matrix. In practice, the covariances of different classes can be different. In the future, we will investigate how to optimize the projection matrix when considering the covariance of each class.

Acknowledgments

This work is supported in part by NSFC (no. U0735004, 60772116, 61075021, 60902087) and GDSFC (no. S2011020000541, 2010B090400397, 2011B090400146).

References

- [1] C.L. Liu and H. Fujisawa, "Classification and learning for character recognition: comparison of methods and remaining problems," in

Machine Learning in Document Analysis and Recognition, ed. S. Marinai, H. Fujisawa, pp.139–161, Springer-Verlag, Berlin, 2008.

- [2] C.L. Liu, F. Yin, D.H. Wang, and Q.F. Wang, “Chinese handwriting recognition contest 2010,” Proc. 2010 Chinese Conference on Pattern Recognition, pp.1–5, Beijing, China, Oct. 2010.
- [3] T.H. Zhang, D.C. Tao, X.L. Li, and J. Yang, “Patch alignment for dimensionality reduction,” IEEE Trans. Knowl. Data Eng., vol.21, no.9, pp.1299–1313, Sept. 2009.
- [4] A.M. Martinez and A.C. Kak, “PCA versus LDA,” IEEE Trans. Pattern Anal. Mach. Intell., vol.23, no.2, pp.228–233, Feb. 2001.
- [5] T. Long and L.W. Jin, “Building compact MQDF classifier for large character set recognition by subspace distribution sharing,” Pattern Recognit., vol.41, no.9, pp.2916–2925, Sept. 2008.
- [6] H. Liu and X. Ding, “Handwritten character recognition using gradient feature and quadratic classifier with multiple discrimination schemes,” Proc. 8th Int’l Conf. Document Analysis and Recognition, pp.19–23, Seoul, Korea, Sept. 2005.
- [7] R.A. Fisher, “The use of multiple measurements in taxonomic problems,” Annals of Eugenics, vol.7, pp.179–188, 1936.
- [8] C.R. Rao, “The Utilization of multiple measurements in problems of biological classification,” J. Royal Statistical Soc. B: Methodological, vol.10, pp.159–203, Oct. 1948.
- [9] D.C. Tao, X.L. Li, X.D. Wu, and S.J. Maybank, “Geometric mean for subspace selection,” IEEE Trans. Pattern Anal. Mach. Intell., vol.31, no.2, pp.260–274, Feb. 2009.
- [10] M. Loog, R.P.W. Duin, and R. Haeb-Umbach, “Multiclass linear dimension reduction by weighted pairwise Fisher criteria,” IEEE Trans. Pattern Anal. Mach. Intell., vol.23, no.7, pp.762–766, July 2001.
- [11] K. Fukunaga, Introduction to Statistical Pattern Recognition, 2nd ed. Academic Press, 1990.
- [12] S.C. Yan, D. Xu, B.Y. Zhang, H.J. Zhang, and Q. Yang, “Graph embedding and extensions: a general framework for dimensionality reduction,” IEEE Trans. Pattern Anal. Mach. Intell., vol.29, no.1, pp.40–51, Jan. 2007.
- [13] X. He and P. Niyogi, “Locality preserving projections,” Advances in Neural Information Processing System, vol.16, pp.153–160, 2004.
- [14] M. Sugiyama, “Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis,” J. Machine Learning Research, vol.8, no.5, pp.1027–1061, 2007.
- [15] R.O. Duda, P.E. Hart, and D.G. Stork, Pattern Classification, 2nd ed. John Wiley & Sons, 2001.
- [16] C.L. Liu, K. Nakashima, H. Sako, and H. Fujisawa, “Handwritten digit recognition: investigation of normalization and feature extraction techniques,” Pattern Recognit., vol.37, no.2, pp.265–279, 2004.
- [17] L.W. Jin and G. Wei, “Handwritten Chinese character recognition with directional decomposition cellular features,” J. Circuits Syst. Comput., vol.8, no.4, pp.517–524, Aug. 1999.
- [18] F. Kimura, K. Takashina, S. Tsuruoka, and Y. Miyake, “Modified quadratic discriminant functions and the application to Chinese character recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol.9, no.1, pp.149–153, Jan. 1987.
- [19] H. Zhang, J. Guo, G. Chen, and C. Li, “HCL2000—A large-scale handwritten Chinese character database for handwritten character recognition,” Proc. 10th Int’l Conf. Document Analysis and Recognition, Paris, France, pp.286–290, July 2009.
- [20] S.W. Lee and J.S. Park, “Nonlinear shape normalization methods for the recognition of large set handwritten character,” Pattern Recognit., vol.27, no.7, pp.895–902, July 1994.

Xue Gao received the Ph.D. degree in Electronic Science and Technology from South China University of Technology in 2003. From 2004 to 2005, he worked as a postdoctoral researcher at Ecole Polytechnique de l’Université de Nantes, France. He is now an associate professor at School of Electronic and Information Engineering, South China University of Technology. His current research interests include document analysis and recognition, machine learning.

Jinzhi Guo received his B.S. degree in 2009, and pursuing his M.S. degree, in Communication and Information System at South China University of Technology, Guangzhou, China. His research interests include computer vision and signal processing.

Lianwen Jin obtained B. Engineering from University of Science and Technology of China and Ph.D. in Communication and Information System from South China University of Technology in 1991 and 1996, respectively. He visited Motorola China Research Center in 2000, the University of Hongkong in 2002 and 2006 as research fellow, respectively. He is now a professor at School of Electronic and Information Engineering, South China University of Technology. His current research fields include hand-

written character recognition, pattern analysis and recognition, image processing, machine learning and intelligent system. He is a member of IEEE.