

Facial Expression Recognition Based on Local Binary Patterns and Local Fisher Discriminant Analysis

SHIQING ZHANG¹, XIAOMING ZHAO², BICHENG LEI¹

¹School of Physics and Electronic Engineering

Taizhou University

Taizhou 318000

CHINA

tzcqsq@163.com, leibicheng@163.com

²Department of Computer Science

Taizhou University

Taizhou 318000

CHINA

tzxyzxm@163.com

Abstract: - Automatic facial expression recognition is an interesting and challenging subject in signal processing, pattern recognition, artificial intelligence, etc. In this paper, a new method of facial expression recognition based on local binary patterns (LBP) and local Fisher discriminant analysis (LFDA) is presented. The LBP features are firstly extracted from the original facial expression images. Then LFDA is used to produce the low dimensional discriminative embedded data representations from the extracted high dimensional LBP features with striking performance improvement on facial expression recognition tasks. Finally, support vector machines (SVM) classifier is used for facial expression classification. The experimental results on the popular JAFFE facial expression database demonstrate that the presented facial expression recognition method based on LBP and LFDA obtains the best recognition accuracy of 90.7% with 11 reduced features, outperforming the other used methods such as principal component analysis (PCA), linear discriminant analysis (LDA), locality preserving projection (LPP).

Key-Words: - Facial expression recognition, local binary patterns, local Fisher discriminant analysis, support vector machines, principal component analysis, linear discriminant analysis, locality preserving projection

1 Introduction

Facial Expression is one of the most powerful, nature, and immediate means for human beings to communicate their emotions and intentions. Automatic facial expression recognition has increasingly attracted much attention due to its important applications to natural human-computer interaction, data driven animation, video indexing, etc.

An automatic facial expression recognition system involves two crucial parts: facial feature representation and classifier design. Facial feature representation is to extract a set of appropriate features from original face images for describing faces. Mainly two types of approaches to extract facial features are found: geometry-based methods and appearance-based methods [1]. In the geometric

feature extraction system, the shape and location of various face components are considered. The geometry-based methods require accurate and reliable facial feature detection, which is difficult to achieve in real time applications. In contrast, the appearance-based methods, image filters are applied to either the whole face image known as holistic feature or some specific region of the face image known as local feature to extract the appearance change in the face image. So far, principal component analysis (PCA) [2], linear discriminant analysis (LDA) [3], and Gabor wavelet analysis [4] have been applied to either the whole-face or specific face regions to extract the facial appearance changes. Nevertheless, it is computationally expensive to convolve the face images with a set of Gabor filters to extract multi-scale and multi-orientation coefficients. It is thus inefficient in both time and

memory for high redundancy of Gabor wavelet features.

Local binary patterns (LBP) [5], originally proposed for texture analysis [6] and a non-parametric method efficiently summarizing the local structures of an image, have received increasing interest for facial image representation. The most important property of LBP features is their tolerance against illumination changes and their computational simplicity. In recent years, LBP has been successfully applied as a local feature extraction method in facial expression recognition [7-11]. When using the extracted LBP features represented by a set of high dimensional data sets to train and test a classifier, the so-called curse of dimensionality emerges, and thus removing irrelevant feature data, as a preprocessing step to a classifier, is needed. To solve this problem, one usually feasible way is to perform dimensionality reduction for the sake of generating few new features containing most of the valuable facial expression information. The two widely used dimensionality reduction methods are PCA and LDA. However, these two methods, i.e., PCA and LDA, still have their respective inherent drawbacks, resulting in decreasing their performance on facial expression recognition tasks to some extent. In detail, PCA, as an unsupervised learning method, fails to extract the discriminative embedded information from high dimensional data. In contrast, LDA is a supervised learning method, but still has an essential limitation. That is, the maximum of embedded features by LDA must be less than the number of data classes due to the rank deficiency of the between-class scatter matrix [3].

In recent years, a new dimensionality reduction method called local Fisher discriminant analysis (LFDA) [12] has been proposed to overcome the limitation of LDA. LFDA effectively combines the ideas of LDA and locality preserving projection (LPP) [13], that is, LFDA maximizes between-class separability and preserves within-class local structure at the same time. LFDA is thus capable of extracting the low dimensional discriminative embedded data representations. Motivated by the deficiency of studies on LFDA for facial expression recognition, in this work we explore the performance of LFDA on facial expression recognition tasks. We firstly use LFDA to extract the low dimensional discriminative embedded data representations from the original extracted high dimensional LBP features. Then the popular support vector machines (SVM) is adopted for facial expression classification. To verify the effectiveness of LFDA we compare LFDA with PCA, LDA and LPP for facial expression recognition. We

conduct facial expression recognition experiments on the popular Japanese female facial expression (JAFFE) [14] database.

The remainder of this paper is organized as follows. Local Binary Patterns (LBP) is given in Section 2. In Section 3, PCA, LDA and LPP are reviewed. In Section 4, LFDA is described. SVM is introduced in Section 5. The popular JAFFE facial expression database is introduced in Section 6. Section 7 shows the experiment results and analysis. Finally, the conclusions are given in Section 8.

2 Local Binary Patterns

The original local binary patterns (LBP) [5] operator takes a local neighborhood around each pixel, thresholds the pixels of the neighborhood at the value of the central pixel and uses the resulting binary-valued image patch as a local image descriptor. It was originally defined for 3×3 neighborhoods, giving 8 bit codes based on the 8 pixels around the central one. The operator labels the pixels of an image by thresholding a 3×3 neighborhood of each pixel with the center value and considering the results as a binary number, and the 256-bin histogram of the LBP labels computed over a region is used as a texture descriptor. Fig.1 gives an example of the basic LBP operator.

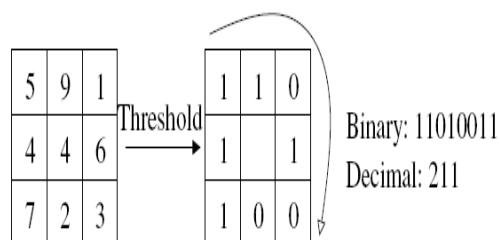


Fig.1 An example of basic LBP operator

The limitation of the basic LBP operator is that its small 3×3 neighborhood cannot capture the dominant features with large scale structures. As a result, to deal with the texture at different scales, the operator was later extended to use neighborhoods of different sizes [5]. Fig.2 gives an example of the extended LBP operator, where the notation (P, R) denotes a neighborhood of P equally spaced sampling points on a circle of radius of R that form a circularly symmetric neighbor set. The second defined the so-called uniform patterns: an LBP is 'uniform' if it contains at most one 0-1 and one 1-0 transition when viewed as a circular bit string. For instance, 00000000, 001110000 and 11100001 are uniform patterns. It is observed that uniform patterns account

for nearly 90% of all patterns in the (8, 1) neighborhood and for about 70% in the (16, 2) neighborhood in texture images. Accumulating the patterns which have more than 2 transitions into a single bin yields an LBP operator, $LBP_{P,R}^{u2}$, with less than 2^p bins. Here, the superscript u2 in $LBP_{P,R}^{u2}$ indicates using only uniform patterns and labeling all remaining patterns with a single label.

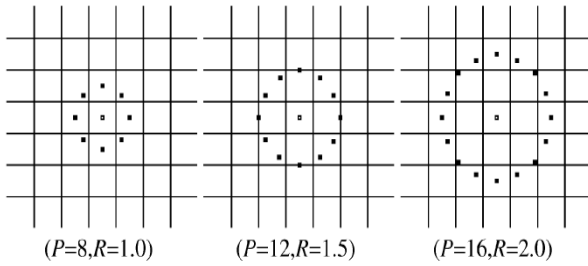


Fig.2 An example of the extended LBP with different (P, R)

After labeling an image with the LBP operator, a histogram of the labeled image $f_i(x, y)$ can be defined as

$$H_i = \sum_{x,y} I(f_i(x, y) = i), \quad i = 0, \dots, n-1 \quad (1)$$

where n is the number of different labels produced by the LBP operator and

$$I(A) = \begin{cases} 1, & A \text{ is true} \\ 0, & A \text{ is false} \end{cases} \quad (2)$$

This LBP histogram contains information about the distribution of the local micro-patterns, such as edges, spots and flat areas, over the whole image, so can be used to statistically describe image characteristics. For efficient face representation, face images were equally divided into m small regions R_1, R_2, \dots, R_m . Once the m small regions R_1, R_2, \dots, R_m are determined, a histogram is computed independently within each of the m small regions. The resulting m histograms are concatenated into a single, spatially enhanced histogram which encodes both the appearance and the spatial relations of facial regions. In this spatially enhanced histogram, we effectively have a description of the face image on three different levels of locality: the labels for the histogram contain information about the patterns on a pixel-level, the labels are summed over a small region to produce information on a regional level and the regional histograms are concatenated to build a global description of the face image.

3 Review of PCA, LDA and LPP

The general dimensionality reduction problem is as follows. Given n data points $\{x_1, x_2, \dots, x_n\}$ with dimension D , dimensionality reduction techniques transform data set $\mathbf{X} = [x_1, x_2, \dots, x_n]$ into a new data set $\mathbf{Y} = [y_1, y_2, \dots, y_n]$ with dimension d ($d \leq D$), while retaining the geometry of the data as much as possible. In the following subsection, we will review PCA, LDA and LPP in brief.

3.1 PCA

Principal component analysis (PCA) [2] is a well-known and widely used linear dimensionality reduction technique. PCA aims to produce a low dimensional representation of high dimensional data that preserves the greatest sources of variation within the data set. This is achieved by performing a linear transformation of the data, projecting it onto the axes of greatest variance, called the principal components. The resulting low dimensional features are uncorrelated and ordered such that the greatest variance by any projection of the data set is accounted for by the first dimension, the second greatest variance by the second dimension, and so on.

In order to find a linear mapping \mathbf{M} , PCA maximizes the following objective function:

$$J_F(\mathbf{M}) = \text{trace}(\mathbf{M}^T \text{cov}(\mathbf{X})\mathbf{M}) \quad (3)$$

where $\text{cov}(\mathbf{X})$ is the sample covariance matrix of the data $\mathbf{X} = [x_1, x_2, \dots, x_n]$. Then, PCA solves the following eigenproblem:

$$\text{cov}(\mathbf{X})\mathbf{M} = \lambda\mathbf{M} \quad (4)$$

The d principal eigenvectors of the covariance matrix form the linear mapping \mathbf{M} . And then the low dimensional data representations are computed by $\mathbf{Y} = \mathbf{X}\mathbf{M}$. Here, \mathbf{X} is assumed to be centered, i.e. have zero mean. In face recognition, x_i represents a face image, and the eigenvectors are so-called eigenfaces.

3.2 LDA

Linear discriminant analysis (LDA) [3] is to seek the discriminant vectors such that the ratio of the between-class scatter to the within-class scatter is maximized. Let $x_i \in R^D$ be D -dimensional samples and $l_i \in \{1, 2, \dots, c\}$ be associated class labels, where n

is the number of samples and c is the number of classes. Let $y_i \in R^d$ ($1 \leq d \leq D$) be the low dimensional data representation of a sample x_i , where d is the dimension of the embedding space. Then the between-class scatter matrix S_b and the within-class scatter matrix S_w are constructed as follows:

$$S_b = \sum_{i=1}^c l_i (m_i - m_0)(m_i - m_0)^T \quad (5)$$

$$S_w = \sum_{i=1}^c \sum_{j=1}^{l_i} (x_i^{(j)} - m_i)(x_i^{(j)} - m_i)^T \quad (6)$$

where $x_i^{(j)}$ is the j th sample of class i ($i=1,2,\dots,c$), m_i is the mean vector of the samples in class i , and m_0 is the mean vector of all samples.

The LDA method tries to find the projected matrix that maximizes the ratio of the between-class scatter matrix to the within-class scatter matrix in the projected space:

$$J_F(\mathbf{V}) = \max \frac{\text{trace}(\mathbf{V}^T \mathbf{S}_b \mathbf{V})}{\text{trace}(\mathbf{V}^T \mathbf{S}_w \mathbf{V})} \quad (7)$$

where \mathbf{V} can be obtained via the generalized eigenvalue problem:

$$\mathbf{S}_b \mathbf{V} = \lambda \mathbf{S}_w \mathbf{V} \quad (8)$$

where the eigenvectors \mathbf{V} corresponds to the d largest eigenvalues λ . Then the d -dimensional representation is $\mathbf{Y} = \mathbf{XV}$. Since the between-class scatter matrix S_b has at most rank $c-1$, LDA can find at most $c-1$ meaningful features. This is an essential limitation of LDA for dimensionality reduction.

3.3 LPP

While PCA aims to preserve the global structure of the data, LPP [13] seeks to preserve the local (i.e., neighborhood) structure of the data by learning a locality preserving submanifold.

Based on the spectral graph theory, LPP constructs a weighted graph $G = (v, \varepsilon, \mathbf{P})$, where v is the set of all points, ε is the set of edges connecting the points and \mathbf{P} is a similarity matrix with weights characterizing the likelihood of two points. The objective function of LPP is as follows:

$$\min_w \sum_{ij} \|y_i - y_j\|^2 P_{ij} \quad (9)$$

where $y_i = \mathbf{W}^T x_i$, $i=1,2,\dots,n$, and $\mathbf{P} = (P_{ij})_{n \times n}$ is a similarity matrix which is defined as follows:

$$P_{ij} = \begin{cases} \exp(-\|x_i - x_j\|^2 / t) & \text{if } x_i \text{ is among kNN of } x_j \\ & \text{or if } x_i \text{ is among kNN of } x_i \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

With simple formulation, the objective function is equivalent to minimizing

$$\begin{aligned} & \frac{1}{2} \sum_{ij} \|y_i - y_j\|^2 P_{ij} \\ &= \frac{1}{2} \sum_{ij} \|\mathbf{W}^T x_i - \mathbf{W}^T x_j\|^2 P_{ij} \\ &= \mathbf{W}^T \mathbf{X}(\mathbf{D} - \mathbf{P})\mathbf{X}^T \mathbf{W} \\ &= \mathbf{W}^T \mathbf{X} \mathbf{L} \mathbf{X}^T \mathbf{W} \end{aligned} \quad (11)$$

where \mathbf{D} is a diagonal matrix with its entries being the row sums of \mathbf{P} , i.e., $d_{ii} = \sum_j p_{ij}$, and $\mathbf{L} = \mathbf{D} - \mathbf{P}$ is the Laplacian matrix.

In order to remove the arbitrary scaling factor in the embedding, LPP imposes a constraint as follows:

$$\mathbf{W}^T \mathbf{X} \mathbf{D} \mathbf{X}^T \mathbf{W} = \mathbf{1} \quad (12)$$

This constraint sets the mapping (embedding) scale and makes the vertices with high similarities to be mapped nearer to the origin. Finally, the minimization problem reduces to

$$\min_w \frac{\mathbf{W}^T \mathbf{X} \mathbf{L} \mathbf{X}^T \mathbf{W}}{\mathbf{W}^T \mathbf{X} \mathbf{D} \mathbf{X}^T \mathbf{W}} \quad (13)$$

The optimal w is given by the minimum eigenvalue solution to the following generalized eigenvalue problem:

$$\mathbf{X} \mathbf{L} \mathbf{X}^T \mathbf{W} = \lambda \mathbf{X} \mathbf{D} \mathbf{X}^T \mathbf{W} \quad (14)$$

That is, LPP seeks a transformation matrix \mathbf{W} such that nearby data pairs in the original space are kept close in the embedding space. Thus, LPP tends to preserve the local structure of the data. In our experiment, the neighbour number of KNN is set to 1 and the parameter t is empirically set to 5.

4 LFDA

Local Fisher Discriminant Analysis (LFDA) [12] finds a transformation matrix \mathbf{T} such that an embedded representation y_i of a sample x_i is given by

$$y_i = \mathbf{T}^T x_i \quad (15)$$

where \mathbf{T}^T denotes the transpose of a matrix \mathbf{T} . Let n_l be the number of samples in class l :

$$\sum_{l=1}^c n_l = n \quad (16)$$

Let $\mathbf{S}^{(bw)}$ and $\mathbf{S}^{(lb)}$ be the local within-class scatter matrix and the local between-class scatter matrix:

$$\mathbf{S}^{(bw)} = \frac{1}{2} \sum_{i,j=1}^n \mathbf{W}_{i,j}^{(bw)} (x_i - x_j)(x_i - x_j)^T \quad (17)$$

$$\mathbf{S}^{(lb)} = \frac{1}{2} \sum_{i,j=1}^n \mathbf{W}_{i,j}^{(lb)} (x_i - x_j)(x_i - x_j)^T \quad (18)$$

$$\mathbf{W}_{i,j}^{(bw)} = \begin{cases} \mathbf{A}_{i,j} / n_l & \text{if } l_i = l_j \\ 0 & \text{if } l_i \neq l_j \end{cases} \quad (19)$$

$$\mathbf{W}_{i,j}^{(lb)} = \begin{cases} \mathbf{A}_{i,j}(1/n - 1/n_l) & \text{if } l_i = l_j \\ 1/n & \text{if } l_i \neq l_j \end{cases} \quad (20)$$

where \mathbf{A} is a affinity matrix between x_i and x_j . Using the local scaling heuristic, \mathbf{A} is defined as

$$\mathbf{A}_{i,j} = \exp(-\|x_i - x_j\|^2 / \sigma_i \sigma_j) \quad (21)$$

where σ_i is the local scaling around x_i and defined by $\sigma_i = \|x_i - x_i^{(k)}\|$, and $x_i^{(k)}$ is the k -th nearest neighbor of x_i . A heuristic choice of $k=7$ has shown to be the best performance.

The LFDA transformation matrix \mathbf{T}_{LFDA} is defines as

$$\mathbf{T}_{LFDA} = \arg \max_{\mathbf{T} \in \mathbb{R}^{D \times d}} [\text{trace}(\mathbf{T}^T \mathbf{S}^{(lb)} \mathbf{T} (\mathbf{T}^T \mathbf{S}^{(bw)} \mathbf{T})^{-1})] \quad (22)$$

That is, LFDA seeks a transformation matrix \mathbf{T} such that nearby data pairs in the same class are made close and the data pairs in different classes are

separated from each other; far apart data pairs in the same class are not imposed to be close.

5 Support Vector Machines

Support vector machines (SVM) [17] are based on the statistical learning theory of structural risk management which aims to limit the empirical risk on the training data and on the capacity of the decision function. The basic concept of SVM is to transform the input vectors to a higher dimensional space by a nonlinear transform, and then an optimal hyperplane which separates the data can be found.

Given training data set $(x_1, y_1), \dots, (x_l, y_l), y_i \in \{-1, 1\}$, to find the optimal hyperplane, a nonlinear transform, $Z = \Phi(x)$, is used to make training data become linearly dividable. A weight w and offset b satisfying the following criteria will be found:

$$\begin{cases} w^T z_i + b \geq 1, & y_i = 1 \\ w^T z_i + b \leq -1, & y_i = -1 \end{cases} \quad (23)$$

The above procedure can be summarized to the following:

$$\min_{w,b} \Phi(w) = \frac{1}{2} (w^T w) \quad (24)$$

subject to $y_i (w^T z_i + b) \geq 1, \quad i = 1, 2, \dots, n$

If the sample data is not linearly dividable, the following function should be minimized.

$$\Phi(w) = \frac{1}{2} w^T w + C \sum_{i=1}^l \xi_i \quad (25)$$

whereas ξ can be understood as the error of the classification and C is the penalty parameter for this term.

By using Lagrange method, the decision function of $w_0 = \sum_{i=1}^l \lambda_i y_i z_i$ will be

$$f = \text{sgn}[\sum_{i=0}^l \lambda_i y_i (z_i^T z_i) + b] \quad (26)$$

From the functional theory, a non-negative symmetrical function $K(u, v)$ uniquely defines a Hilbert space H , where K is the rebuild kernel in the space H :

$$K(u, v) = \sum_i \alpha \varphi_i(u) \varphi_i(v) \quad (27)$$

This stands for an internal product of a characteristic space:

$$z_i^T z = \Phi(x_i)^T \Phi(x) = K(x_i, x) \quad (28)$$

Then the decision function can be written as:

$$f = \text{sgn}[\sum_{i=1}^l \lambda_i y_i K(x_i, x) + b] \quad (29)$$

The development of a SVM classification model depends on the selection of kernel function. There are several kernels that can be used in SVM models. These include linear, polynomial, radial basis function (RBF) and sigmoid function.

$$K(x_i, x_j) = \begin{cases} x_i^T x_j & \text{Linear} \\ (\gamma x_i^T x_j + \text{coefficient})^{\text{degree}} & \text{Polynomial} \\ \exp(-\gamma |x_i - x_j|^2) & \text{RBF} \\ \tanh(\gamma x_i^T x_j + \text{coefficient}) & \text{Sigmoid} \end{cases} \quad (30)$$

Many real-world data sets involve multi-class problem. Since SVMs are inherently binary classifiers, the binary SVMs are needed to extend to be multi-class SVMs for multi-class problem. Currently, there are two types of approaches for building multi-class SVMs. One is the “single machine” approach, which attempts to construct multi-class SVMs by solving a single optimization problem. The other is the “divide and conquer” approach, which decomposes the multi-class problem into several binary sub-problems, and builds a standard SVM for each. The most popular decomposing strategy is probably the “one-against-all”. The “one-against-all” approach consists of building one SVM per class and aims to distinguish the samples in a single class from the samples in all remaining classes. Another popular decomposing strategy is the “one-against-one”. The “one-against-one” approach builds one SVM for each pair of classes. When applied to a test point, each classification gives one vote to the winning class and the point is labeled with the class having most votes. In practice, the “one-against-one” approach is more effective than the “one-against-all” approach due to its computation simplicity and comparable performance.

6 Facial Expression Database

The popular JAFFE facial expression database [14] used in this study contains 213 facial images from 10 Japanese female. Each image has a resolution of 256×256 pixels. The head is almost in frontal pose. The number of images corresponding to each of the 7 categories of expression (neutral, happiness, sadness, surprise, anger, disgust and fear) is almost the same. A few of them are shown in Fig. 3.



Fig.3 Examples of facial images from JAFFE database

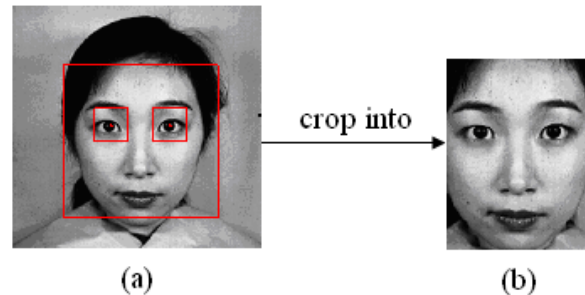


Fig.4 (a) Two eyes location, (b) the final cropped image of 110×150 pixels

As done in [7, 8, 15], we normalized the faces to a fixed distance of 55 pixels between the centers of two eyes. Generally, it is observed that the width of a face is roughly two times of the distance, and the height is roughly three times. Therefore, based on the centers of two eyes, facial images of 110×150 pixels were cropped from original image. To locate the centers of two eyes, automatic face registration was performed by using a robust real-time face detector based on a set of rectangle haar-like features [16]. From the results of automatic face detection including face location, face width and face height, two square bounding boxes for left eye and right eye were automatically created respectively. Then, the center locations of two eyes can be quickly worked out in terms of the centers of two square bounding boxes for left eye and right eye. Fig.4 shows the process of two eyes location and the final cropped

image. No further alignment of facial features such as alignment of mouth was performed. Additionally, there was no attempt made to remove illumination changes due to LBP's gray-scale invariance.

7 Experiments and Results Analysis

The cropped facial images of 110×150 pixels contain facial main components such as mouth, eyes, brows and noses. For simplicity, we applied LBP operator to the whole region of the cropped facial images. As done in [7, 8], we selected the 59-bin operator $LBP_{P,R}^{u2}$, and divided the 110×150 pixels face images into 18×21 pixels regions, giving a good trade-off between recognition performance and feature vector length. Thus face images were divided into 42 (6×7) regions, and represented by the LBP histograms with the length of 2478 (59×42). The reduced feature dimension is limited to the range [2, 20]. We used the LIBSVM package [18] to implement SVM algorithm with radial basis function (RBF) kernel, kernel parameter optimization, one-against-one strategy for multi-class classification problem. All extracted LBP features were normalized by a mapping to [0, 1] before anything else.

To testify the performance of LFDA, we use the JAFFE database to perform two types of facial expression recognition experiments: person-dependent experiments and person-independent experiments. For person-dependent experiments, the training data and testing data have the same person with different images. A more challenging application is to create a person-independent facial expression recognition system since the facial expression recognition system in real-world sceneries should be work for recognizing new person's expressions. Therefore, for person-independent experiments, each person only lies in either training data or testing data so that the persons

in training data are guaranteed to be independent to the persons in testing data.

7.1 System Structure

In order to clarify the scheme of how to employ dimensionality reduction techniques like LFDA on facial expression recognition tasks, Fig.5 shows the basic structure of a facial expression recognition system based on dimensionality reduction techniques. As shown in Fig.5, we can see that this system consists of three main components: feature extraction, feature dimensionality reduction and facial expression recognition. In the feature extraction stage, the original facial images from the JAFFE facial expression database are divided into two parts: training data and testing data. The corresponding LBP features for training data and testing data are extracted. The result of this stage is the extracted facial feature data represented by a set of high dimensional LBP features. The second stage aims at reducing the size of LBP features and generating the new low dimensional embedded features with dimensionality reduction techniques, such as LFDA, PCA, LDA and LPP. It is noted that for the mapping of testing data, the low dimensional embedded mapping of training data is needed to be learnt. This is realized by using the out-of-sample extensions of dimensionality reduction methods. Due to the linearity, the out-of-sample extensions of all used linear dimensionality reduction methods, i.e., LFDA, PCA, LDA and LPP, are performed by multiplying testing data with the linear mapping matrix with a straightforward method. The last stage in this system is in the low dimensional embedded feature space the trained SVM classifier are used to predict the accurate facial expression categories on testing data and the recognition results are given.

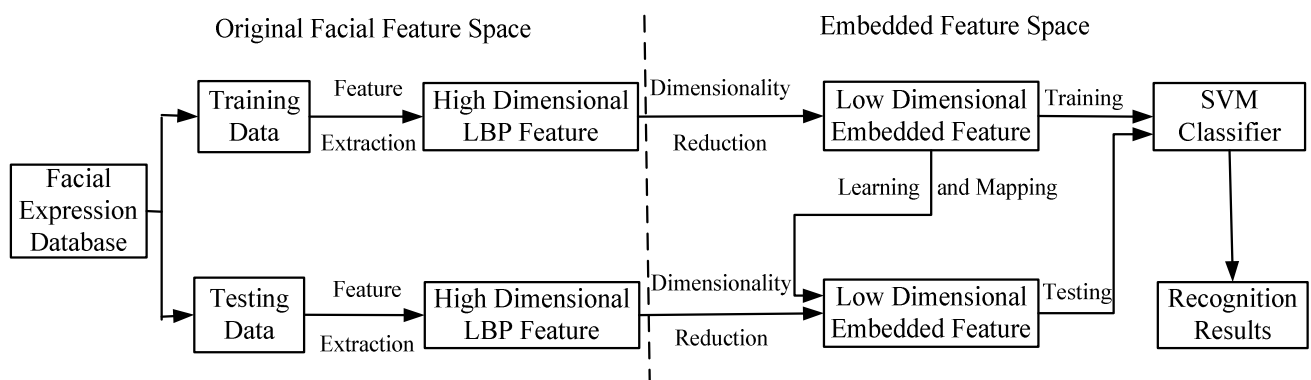


Fig.5 The basic structure of a facial expression recognition system based on dimensionality reduction

7.2 Person-dependent Experiments

To evaluate the performance of the algorithms for person-dependent facial expression recognition on the JAFFE database, a 10-fold stratified cross validation scheme was performed for facial expression recognition experiments and the average recognition results were reported. In 10-fold cross validation, the original samples are randomly partitioned into ten subsets. Of the ten subsets, one subset is retained as the validation data for testing the model, and the remaining nine subsets are used as training data. This process is then repeated ten times, with each of the ten subsets used exactly once as the validation data. Then the average result across ten folds is computed. The person-dependent recognition results of different dimensionality reduction methods, i.e., PCA, LDA, LPP and LFDA, are given in Fig.6. It is pointing out that the reduced dimension of LDA is set to the range [2, 6] because LDA can find at most 6 (less than 7 categories of expression) meaningful embedded features due to the rank deficiency of the between-class scatter matrix [3]. The best accuracy for different methods with corresponding reduced dimension is presented in Table 1. Note that the ‘‘Baseline’’ method denotes that the result is obtained on the original 2478 dimensional LBP features without any dimensionality reduction.

From the results in Fig.6 and Table 1, we can make the following observations. First, LFDA obtains the highest accuracy of 90.7% with 11 reduced features, outperforming the other methods, i.e., Baseline, LDA, LPP and PCA. This indicates that LFDA is capable of extracting the most discriminative low dimensional embedded data representations for facial expression recognition. Second, LDA performs better than PCA and LPP, since LDA is a supervised dimensionality reduction method and can extract the low dimensional embedded data representations with higher discriminative power than PCA and LPP. Third, PCA outperforms LPP. This may be caused by the fact PCA retains information relevant to variation while reducing redundant information so that PCA is more capable of extracting discriminative information than LPP. Finally, there is no significant improvement on facial expression recognition performance if more reduced feature dimensions are used. This shows that in our experiments it is acceptable that the reduced target feature dimension is confined to the range [2, 20].

To further explore the recognition accuracy of each expression when LFDA performs best, Table 2 gives the confusion matrix of 7-class facial expression recognition results in person-dependent

case. From Table 2 we can observe that four expressions, i.e., anger, joy, disgust and neutral, are classified with more than 90% accuracy, while other three expressions, sad, surprise and fear, are discriminated with relatively low accuracy (less than 90%).

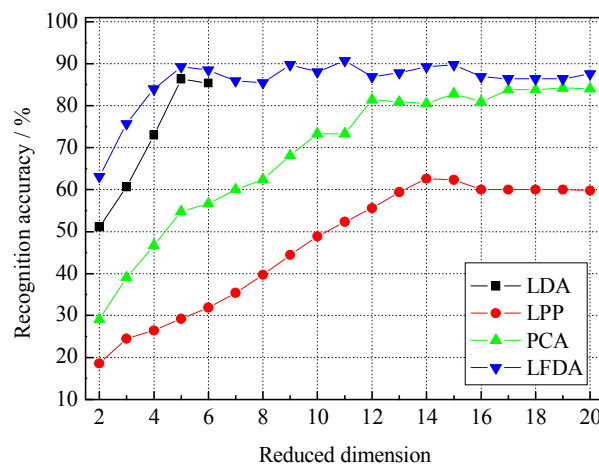


Fig.6 Person-dependent recognition results versus reduced dimension

Table 1 The best recognition accuracy (%) in person-dependent case for different methods with corresponding reduced dimension

Methods	Dimension	Accuracy
Baseline	2478	87.32
LDA	5	86.33
LPP	14	62.62
PCA	19	84.24
LFDA	11	90.70

Compared with previously reported results [8-11] in person-dependent case on the JAFFE database, in our work based on LBP and LFDA the best recognition accuracy of 90.7% with 11 reduced features is highly competitive. In [8], similar to our experimental settings, they obtained the best accuracy of 81% with SVM and LBP features. In [9], they extracted the local texture information by applying LBP to facial feature points; the shape information was also considered as the pair direction. In addition, they used LBP with the entire image to get global texture information. Combining these three types of features, with the nearest neighbour classifier they reported an accuracy of 83%. In [10], the recognition accuracy of 85.57% was achieved by

using SVM and LBP features, but they did not performed 10-fold cross-validation. In [11], by using LBP features and the linear programming technique, they reported an accuracy of 93.8%. Nevertheless,

they preprocessed the images by using the CSU Face Identification Evaluation System [19] to exclude non-face area with an elliptical mask.

Table 2 Confusion matrix of 7-class facial expression obtained by LFDA in person-dependent case

	Anger (%)	Joy (%)	Sad (%)	Surprise (%)	Disgust (%)	Fear (%)	Neutral (%)
Anger	92.38	0	5.76	0.08	0.16	0.23	1.39
Joy	0	96.22	0.26	0.55	0.21	0.19	2.57
Sad	0	1.43	84.15	0.18	1.24	2.57	11.43
Surprise	0.16	2.54	0.13	88.31	0	2.82	6.04
Disgust	1.38	0.14	4.63	0	90.76	3.09	0
Fear	0.19	0	5.69	2.28	1.12	87.47	3.25
Neutral	0	0	2.45	0.89	0	1.03	95.63

7.3 Person-independent Experiments

To evaluate the performance of the algorithms for person-independent facial expression recognition on the JAFFE database, we firstly split the whole 213 facial images into ten groups according to the persons the JAFFE database contains, with each group including all the seven expressions of one distinct person. Then the so-called leave-one-person-out cross validation strategy is used in the experiments. That is, each time, facial expression images of one person are used for training and all the images of the remaining persons are used for testing. Repeat the process for each person. The average is the final recognition rate. Fig.7 gives the person-independent recognition results of different dimensionality reduction methods. Table 3 presents the best accuracy for different methods with corresponding reduced dimension.

As shown in Fig.7 and Table 3, we can see that LFDA still performs best among all used methods for person-independent facial expression recognition. In detail, LFDA gives the highest accuracy of 65.91% with 10 reduced features. In addition, compared the person-independent recognition results in Fig.7 with the person-dependent recognition results in Fig.6, we can observe that the recognition accuracy in person-independent case are much lower than the recognition accuracy in person-dependent case. More precisely, we can get the best accuracy of about 90% for person-dependent facial expression

recognition, while about 65% for person-independent facial expression recognition. The results of about 65% accuracy in person-independent case are reasonable since human beings themselves normally can only recognize expressions with an accuracy of about 60% [20].

Table 4 presents the confusion matrix of 7-class expression recognition results in person-independent case when using LFDA to obtain the best performance. As shown in Table 4, we can see that neutral is identified best with an accuracy of 86.43%, whereas the other five expressions are classified with less than 80% accuracy.

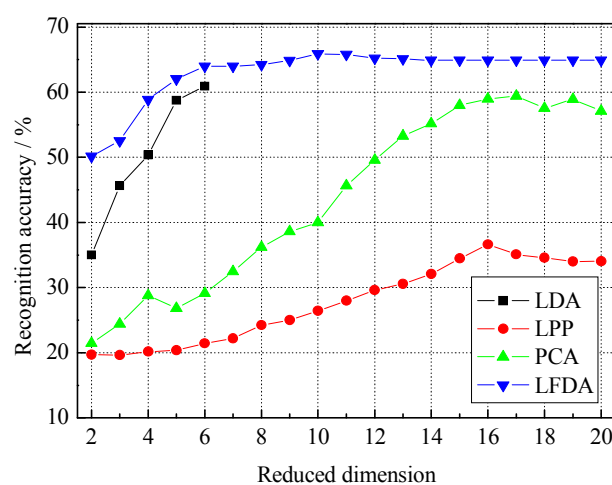


Fig.7 Person-independent recognition results versus reduced dimension

Table 3 The best recognition accuracy (%) in person-independent case for different methods with corresponding reduced dimension

Methods	Dimension	Accuracy
Baseline	2478	61.28
LDA	6	60.92
LPP	16	36.61
PCA	17	59.42
LFDA	10	65.91

Compared with the previously reported results [21-23] in person-independent case on the JAFFE database, the recognition accuracy of about 65% in our experiments is still comparable. In [21], they used the method of PCA+LDA to extract the

statistical features, and obtained the difference of statistical features for every expression. Finally, by using the difference of statistical features and the nearest neighbour classifier they reported the highest accuracy of 62.78%. In [22], based on two dimensional facial expression feature extraction methods, including two dimensional principal component analysis (2DPCA), two dimensional linear discriminant analysis (2DLDA) and generalized low rank approximation of matrices (GLRAM), with SVM classifier they achieved the recognition accuracy of 63.1%, 60.5%, 61.4% for 2DPCA, 2DLDA, and GLRAM, respectively. In [23], base on the 2nd-order gray-level raw pixels and the encoded 3rd-order tensor-formed Gabor features of facial expression images, they employed the orthogonal tensor neighbourhood preserving embedding (OTNPE) algorithm for dimensionality reduction, and obtained about 50% accuracy with the nearest neighbour classifier.

Table 4 Confusion matrix of 7-class facial expression obtained by LFDA in person-independent case

	Anger (%)	Joy (%)	Sad (%)	Surprise (%)	Disgust (%)	Fear (%)	Neutral (%)
Anger	64.65	0	12.48	0.13	8.04	2.87	11.83
Joy	0.14	61.76	15.04	0.07	0.21	0	22.78
Sad	3.22	2.97	66.32	7.68	8.24	5.45	6.12
Surprise	0.06	5.13	2.34	72.41	0.05	2.17	17.84
Disgust	4.97	1.03	18.94	0.31	56.82	12.79	5.14
Fear	1.09	0.98	9.85	10.24	13.61	52.97	11.26
Neutral	0	0.72	7.26	4.2	0.09	1.3	86.43

8 Conclusions

Facial expression recognition has attracted more and more attention duo to its important applications in a wide range of areas. One key step in facial expression recognition is to extract the low dimensional discriminative features before the feature data are fed into classifier for classification. In this paper, we presented a new method of facial expression recognition based on LBP and LFDA. The experiment results on the popular JAFFE facial expression database indicate that LFDA performs better than PCA, LDA as well as LPP, and obtains

the promising performance of 90.7% accuracy with 11 reduced features. This is attributed to the fact that LFDA has the better ability than PCA, LDA and LPP to extract the low dimensional discriminative embedded data representations for facial expression recognition. In the future, it's an interesting task to employ LFDA to construct a real time facial expression recognition system for natural human-computer interaction.

Acknowledgments

This work is supported by Zhejiang Provincial Natural Science Foundation of China (Grant No.Z1101048, No. Y1111058).

References:

- [1] Y Tian, T Kanade, and J Cohn, Facial expression analysis, *Handbook of face recognition*, Springer, 2005.
- [2] M A Turk, and A P Pentland, Face recognition using eigenfaces, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 1991, pp. 586-591.
- [3] P N Belhumeur, J P Hespanha, and D J Kriegman, Eigenfaces vs. fisherfaces: Recognition using class specific linear projection, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 19, No. 7, 1997, pp. 711-720.
- [4] M J Lyons, J Budynek, and S Akamatsu, Automatic classification of single facial images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 21, No. 12, 1999, pp. 1357-1362.
- [5] T Ojala, M Pietikinen, and T M Enp, Multiresolution gray scale and rotation invariant texture analysis with local binary patterns, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 24, No. 7, 2002, pp. 971-987.
- [6] T Ojala, M Pietikinen, and D Harwood, A comparative study of texture measures with classification based on featured distributions, *Pattern Recognition*, Vol. 29, No. 1, 1996, pp. 51-59.
- [7] C Shan, S Gong, and P McOwan, Robust facial expression recognition using local binary patterns, *Proc. IEEE International Conference on Image Processing*, 2005, pp. 370-373.
- [8] C Shan, S Gong, and P McOwan, Facial expression recognition based on Local Binary Patterns: A comprehensive study, *Image and Vision Computing*, Vol. 27, No. 6, 2009, pp. 803-816.
- [9] X Feng, B Lv, Z Li and et al., A Novel Feature Extraction Method for Facial Expression Recognition, *Proc. Joint Conference on Information Sciences*, 2006.
- [10] S Liao, W Fan, A Chung, and et al., Facial expression recognition using advanced local binary patterns, tsallis entropies and global appearance features, *Proc. IEEE International Conference on Image Processing*, 2006, pp. 665-668.
- [11] X Feng, M Pietikainen, and A Hadid, Facial expression recognition with local binary patterns and linear programming, *Pattern Recognition and Image Analysis*, Vol. 15, No. 2, 2005, pp. 546-548.
- [12] M Sugiyama, T Idé, S Nakajima and et al., Semi-supervised local Fisher discriminant analysis for dimensionality reduction, *Machine learning*, Vol. 78, No. 1, 2010, pp. 35-61.
- [13] X He, and P Niyogi, Locality preserving projections, *Advances in neural information processing systems (NIPS)*, MIT Press, 2003.
- [14] M Lyons, S Akamatsu, M Kamachi, and et al., Coding facial expressions with Gabor wavelets, *Proc. Third IEEE International Conference on Automatic Face and Gesture Recognition*, 1998, pp. 200-205.
- [15] Y Tian, Evaluation of face resolution for expression analysis, *Proc. first IEEE Workshop on Face Processing in Video*, 2004, pp. 82-82.
- [16] P Viola, and M Jones, Robust real-time face detection, *International Journal of Computer Vision*, Vol. 57, No. 2, 2004, pp. 137-154.
- [17] V Vapnik, *The nature of statistical learning theory*, Springer, 2000.
- [18] Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines, 2001. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [19] D Bolme, M Teixeira, J Beveridge, and B Draper, The CSU Face Identification Evaluation System User's Guide: Its Purpose, Feature and Structure, *Proc. 3rd International Conference on Computer Vision Systems*, 2003, pp. 304-313.
- [20] T Jinghai, Y Zilu, and Z Youwei, The contrast analysis of facial expression recognition by human and computer, *Proc. 8th International Conference on Signal Processing*, 2006, pp. 1649-1653.
- [21] G Xue, and Z Youwei, Facial Expression Recognition Based on the Difference of Statistical Features, *Proc. 8th International Conference on Signal Processing*, 2006, pp. 16-20.
- [22] Y Zilu, L Jingwen, and Z Youwei, Facial expression recognition based on two dimensional feature extraction, *Proc. 9th International Conference on Signal Processing*, 2008, pp. 1440-1444.
- [23] S Liu, and Q Ruan, Orthogonal tensor neighborhood preserving embedding for facial expression recognition, *Pattern Recognition*, Vol. 44, 2011, pp. 1497-1513.