2006 IEEE International Conference on
Systems, Man, and Cybernetics
October 8-11, 2006, Taipei, Taiwan

Median LDA: A Robust Feature Extraction Method for Face Recognition

Jian Yang, David Zhang, *Senior Member, IEEE* and Jing-yu Yang

***Abstract*—In the existing LDA models, class mean vector is always estimated by the class sample average. In small sample size problems such as face recognition, however, the class sample average does not suffice to provide an accurate estimate of the class mean based on a few of given samples, particularly when there are outliers in the sample set. To overcome this weakness, we use the class median vector to estimate the class mean vector in LDA modeling. The class median vector has two advantages over the class sample average: (1) the class median (image) vector preserves useful details in the sample images and (2) the class median vector is robust to outliers that exist in training sample set. The proposed median LDA model is evaluated using three popular face image databases. All experiment results indicate that median LDA is more effective than the common LDA and PCA.**

I. Introduction

Fisher linear discriminant analysis (LDA) is a classical method for feature extraction and dimension reduction [1]. Like principal component analysis (PCA), in the past decade, LDA has been applied to face recognition area successfully. Liu [2] developed a LDA algorithm for face recognition in 1992. A more popular LDA-based face recognition technique, discriminant eigenfeatures [3] or Fisherfaces [4], appeared four-year later. These methods are based on a concise two-phase framework: PCA plus LDA. Subsequent research saw the development of a series of LDA algorithms [5-9]. Chen [5] proposed a more effective way to extract the null-subspace discriminant information of LDA for small sample size problems. Jin [6] proposed an uncorrelated linear discriminant transform for face recognition. Yu [7] suggested a direct LDA algorithm to deal with high dimensional image data. Yang [8] supplied the theoretical justification for the PCA plus LDA framework. Liu and Wechsler [9] put forward enhanced LDA models to improve the generalization power of LDA in face recognition applications. In addition, non-linear discriminant analysis, represented by kernel Fisher discriminant (KFD), has been found to be effective in face identification applications [10, 11].

In LDA (or KFD) models, class mean vectors (i.e., the expectation of the class random samples) are used to character the between-class and within-class scatters. These vectors are generally estimated by the class sample averages (i.e., the average of class random samples). Class sample averages therefore plan a critical role in the construction of the between-class and within-class scatter matrices and finally affect the projection directions of LDA. Since face recognition is typically a small sample size problem, in which only a few of image samples are available for training per class, it is difficult to give an accurate estimate of the class mean using the class sample average, in particular when there are outliers in the sample set. The inaccurate estimate of the class mean must have a negative effect on the robustness of LDA models.

To overcome the weakness of the LDA models mentioned above, in this paper, we will use the class median vector, rather than the class sample average, to estimate the class mean vector in the LDA modeling. The class median vector has two main advantages over the class sample average: (1) the class median (image) vector preserves useful details in the sample images and (2) the class median vector is robust to outliers that exist in training sample set (for example, the images with noise, occlusion, etc). Thus, the median-based LDA model should be more robust than the current sample-average based LDA models. We will demonstrate this by our experiments using three popular face image databases in this paper.

II. Fisher Linear Discriminant Analysis and Its Weakness

A. Outline of LDA

LDA seeks to find a projection axis such that the Fisher criterion (i.e., the ratio of *the between-class scatter to the within-class scatter*) is maximized after the projection of samples. Suppose there are c pattern classes $\omega_1, \omega_2, \cdots, \omega_c$ in N-dimensional pattern vector space. The between-class and within-class scatter matrices $\mathbf{S}_b$ and $\mathbf{S}_w$ are defined by

$$\mathbf{S}_b = \frac{1}{M}\sum_{i=1}^{c} l_i (\boldsymbol{\mu}_i - \boldsymbol{\mu}_0)(\boldsymbol{\mu}_i - \boldsymbol{\mu}_0)^{\mathrm{T}} \tag{1}$$

$$\mathbf{S}_w = \frac{1}{M}\sum_{i=1}^{c}\sum_{j=1}^{l_i} (\mathbf{x}_{ij} - \boldsymbol{\mu}_i)(\mathbf{x}_{ij} - \boldsymbol{\mu}_i)^{\mathrm{T}} \tag{2}$$

Manuscript received March 15, 2006. This research was supported by the National Science Foundation of China under Grants No. 60503026, No. 60332010, No. 60472060, and No. 60473039, and the CERG fund from the HKSAR Government and the central fund from the Hong Kong Polytechnic University. Besides, Dr. Yang was supported by China and the Hong Kong Polytechnic University Postdoctoral Fellowships.

Jian Yang is with Biometric Research Centre, Department of Computing, Hong Kong Polytechnic University, Kowloon, Hong Kong (e-mail: csjyang@comp.polyu.edu.hk).

David Zhang is with Biometric Research Centre, Department of Computing, Hong Kong Polytechnic University, Kowloon, Hong Kong (e-mail: csdzhang@comp.polyu.edu.hk).

Jing-yu Yang is with Department of Computer Science, Nanjing University of Science and Technology, Nanjing 210094, P. R. China (e-mail: yangjy@mail.njust.edu.cn).

1-4244-0100-3/06/$20.00 ©2006 IEEE

where $\mathbf{x}_{ij}$ denotes the *j*-th training sample in class *i*; l_i is the number of training samples in class *i*; $\boldsymbol{\mu}_i$ is the *mean vector* of the *population* in class *i*, i.e. $\boldsymbol{\mu}_i = E(\mathbf{x}_{ij} \mid \omega_i)$, where $E(\cdot)$ is the operator of expectation; and $\boldsymbol{\mu}_0$ is the total *mean vector* of the *population*, i.e. $\boldsymbol{\mu}_0 = E(\mathbf{x}_{ij})$.

Generally, the sample average (or called sample mean) is used as a estimator of the population mean. So, $\boldsymbol{\mu}_i = E(\mathbf{x}_{ij} \mid \omega_i)$ is estimated by the *class sample average*

$$\mathbf{m}_i = \frac{1}{l_i}\sum_{j=1}^{l_i} \mathbf{x}_{ij}\ , \tag{3}$$

and $\boldsymbol{\mu}_0 = E(\mathbf{x}_{ij})$ is estimated by the *total sample average*

$$\mathbf{m}_0 = \frac{1}{M}\sum_{i=1}^{c}\sum_{j=1}^{l_i} \mathbf{x}_{ij}\ . \tag{4}$$

The *sample average* $\mathbf{m}_i$ and $\mathbf{m}_0$ always replace the the population mean $\boldsymbol{\mu}_i$ and $\boldsymbol{\mu}_0$ in the calculation of the matrices $\mathbf{S}_b$ and $\mathbf{S}_w$.

The Fisher criterion is defined by

$$J_F(\mathbf{w}) = \frac{\mathbf{w}^T\mathbf{S}_b\mathbf{w}}{\mathbf{w}^T\mathbf{S}_w\mathbf{w}} \tag{5}$$

The stationary points of $J_F(\mathbf{w})$ are the generalized eigenvectors $\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_d$ of $\mathbf{S}_b\mathbf{w} = \lambda\mathbf{S}_w\mathbf{w}$ corresponding to *d* largest eigenvalues. These stationary points form the coordinate system of LDA. For a given sample **x**, we can get its coordinate by the following linear transform:

$$\mathbf{y} = \mathbf{W}^{\mathrm{T}}\mathbf{x}\ \text{, where } \mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_d) \tag{6}$$

The vector **y** is used to represent the sample **x** for recognition purpose.

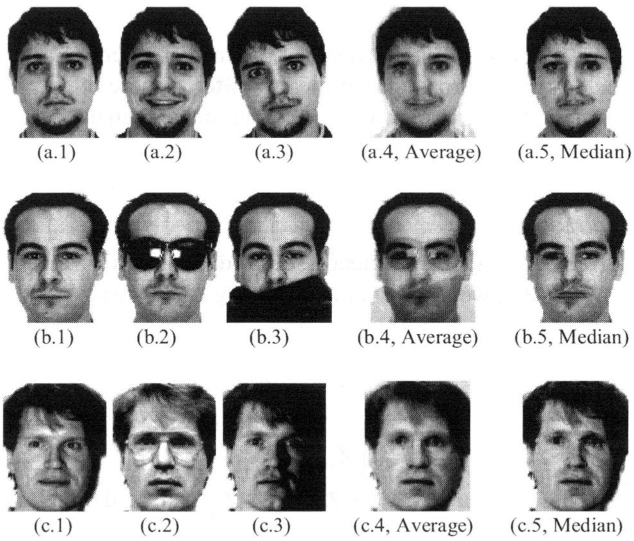
B. Issues of LDA for Small Sample Size Problems

From Equations (1) and (2), we can see that the *class mean vector* plans a central role in the definitions of the between-class and within-class scatter matrices. Thus, the accuracy of its estimate must have substantial effect on the resulting projection directions of LDA. The *class mean* vector (or image) [1], however, is not prone to being estimated accurately by the *class sample average* vector, when there are only a few of image samples available for training per class. In a statistical context, the law of large numbers implies that the average of a random sample from a large population is likely to be close to the mean of the whole population [12]. From this law, we have

$$\mathbf{m}_i = \frac{1}{l_i}\sum_{j=1}^{l_i} \mathbf{x}_{ij} \xrightarrow{\text{Probability}} \boldsymbol{\mu}_i\ \text{, as } l_i \to \infty \tag{7}$$

Therefore, when there is enough amount of training samples in Class *i*, $\boldsymbol{\mu}_i$ can be well estimated by the class *sample average* $\mathbf{m}_i$. However, when there are very few samples available per class, no theory can guarantee that the estimate of $\boldsymbol{\mu}_i$ using Equation (3) is accurate. Besides, there are a lot of instances indicating that the *sample average* may not appear representative of the true central region for skewed data or data with outliers.

In face recognition cases, since each individual only provides several images for training, by averaging these training images, the resulting image is always seriously blurred; some useful details in images are lost. Particularly, when there are outliers in the training sample set, the class mean image might be incorrectly located due to the disturbance of outliers. Fig. 1 shows some examples that the class-average images fail to give an accurate estimate of the true "central tendency" of the face.



(a.1) (a.2) (a.3) (a.4, Average) (a.5, Median)

(b.1) (b.2) (b.3) (b.4, Average) (b.5, Median)

(c.1) (c.2) (c.3) (c.4, Average) (c.5, Median)

Fig. 1 Sample images of some persons and their average and median images (The images in the first and second row are from AR database [17], and those in the third row from Yale database [15]).

III. Median Fisher Linear Discriminant

A. The Concept of Median

In probability theory and statistics, the median defined as a number that separates the higher half of a sample, a population, or a probability distribution from the lower half. It is the middle value in a distribution, above and below which lie an equal number of values. This states that 1/2 of the population will have values less than or equal to the median and 1/2 of the population will have values equal to or greater than the median [13].

To find the median of a finite list of numbers, we need to sort the list into increasing order first. Then, we pick the middle entry value if there are an odd number of observations. Otherwise, we often take the average value of the two middle entry values as the median. Two examples about the choice of

[1] Observe that in image recognition problems, the class mean vector is generated from the class mean image by stacking the columns of the image. Similarly, the class sample average vector is generated from the class sample average image and, the class median vector is from the class median image.

median are given below:

Example 1: With an odd number of data values, for example 7, we have:

Data_SET_1={3.3, 3.0, 10, 3.1, 1, 3.2, 3.4}
Ordered_ Data _SET_1={ 1, 3.0, 3.1, 3.2, 3.3, 3.4, 10}
Median=3.2, Average= 3.857

Example 2: With an even number of data values, for example 8, we have:

Data_SET_2={3.3, 3.0, 10, 3.1, 1, 3.2, 3.4, 3.5}
Ordered_ Data _SET_2={ 1, 3.0, 3.1, 3.2, 3.3, 3.4, 3.5, 10}
Median=(3.2+3.3)/2=3.25, Average=3.8125

Like the sample average, the median can also be used as an estimator of the central tendency such as the population mean. And, it is generally considered that the median is a more robust estimator of the central tendency than the sample average for data with outliers (or skewed data). From the above examples, we can also see that the median does work better than the average when the outliers "1" and "10" exist in the data sets.

A very successful application of the median operator is to filter design. It turns out that median filter is more effective than mean filter for noise removal in an image in many cases [14].

B. Multivariate Statistics: Median Vector and Median Matrix

Given a random sequence of n-dimensional vectors $\mathbf{Z}_1, \mathbf{Z}_2, \cdots \mathbf{Z}_q$, we can form the following data matrix

$$\mathbf{Z} = [\mathbf{Z}_1, \mathbf{Z}_2, \cdots \mathbf{Z}_q] = \begin{bmatrix} Z_{11} & Z_{12} & \cdots & Z_{1q} \\ Z_{21} & Z_{22} & \cdots & Z_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{n1} & Z_{n2} & \cdots & Z_{nq} \end{bmatrix} \tag{8}$$

Then, the *median vector of* $\mathbf{Z}_1, \mathbf{Z}_2, \cdots \mathbf{Z}_q$ can be defined as $\mathbf{M} = (M_1, M_2, \cdots M_n)^{\mathrm{T}}$, where M_j is the median of elements on the j-th row of the data matrix $\mathbf{Z}$. Specifically, if the median operator of a set of numbers is denoted by $Median(\cdot)$, then, $M_j = Median(\{Z_{j1}, Z_{j2}, \cdots, Z_{jq}\})$.

In a similar way, we can define the median matrix of a random sequence of matrices. Actually, we do not need to do so. Instead, we can convert the matrices (by stacking its columns) into vectors, and then get their median vector from the definition of the median vector.

C. Median LDA

Given a set of l_i training sample vectors in Class *i*, we can obtain its *class median* vector $\mathbf{M}_i$ using the definition given in the above subsection. $\mathbf{M}_i$ is used as an estimator of the class mean vector $\boldsymbol{\mu}_i$ in our model. In small sample cases, since each class only provides a few of training samples, the *class median* vector $\mathbf{M}_i$ generally provides more accurate approximation to the true central tendency, in particular when there are outliers in the training samples. Fig. 1 gives some examples of class median images.

From Fig. 1, it can be seen that (1) the *class median* images, preserving more details of the sample images, appear much clearer than the *class sample average* images. (2) As an estimator of the central tendency, *median* is much more robust to outliers than the *sample average.* When there are rotated images, occluded images or images with exceptional lighting in the training sample set, *median* all performs much better than the *sample average*. Specifically, median operator can significantly alleviate the effects of rotations, illuminations and occlusions onto the true central tendency, while the *sample average* cannot.

Besides, it is worthwhile to highlight another merit of median operator for dealing with outliers. Differing from some outlier-removing methods, which just simply remove outliers from the training sample set, the median operator is capable of utilizing the outliers and deriving any valuable information from it. For example, in Fig. 1, image (b.2) or (b.3) can be viewed as outliers with respect to the training sample set of person (b). The median operator, however, does not discard them. It still derives the useful information from the non-occluded parts of these "outlier" images.

Now, let us talk about the estimate of the total *mean vector* $\boldsymbol{\mu}_0$. Although it is possible to use the *total median* vector to estimate $\boldsymbol{\mu}_0$, we still prefer the *total sample average* $\mathbf{m}_0$ as its estimator. There are two justifications for this. First, differing from the size of training samples within class, the number of total training samples is generally large. In such a case, the sample average suffices to provide a satisfying estimate. Second, to consider from the viewpoint of computation, large sample size results in significant increase of computations when median operator is used.

If the *class median* vector $\mathbf{M}_i$ is used to replace $\boldsymbol{\mu}_i$ and $\mathbf{m}_0$ (calculated by Equation (4)) is to replace $\boldsymbol{\mu}_0$ in the construction of the between-class and within-class matrices $\mathbf{S}_b$ and $\mathbf{S}_w$, the resulting LDA model is called Median LDA (MLDA). Due to the advantages of class median (image) vectors over class sample average vectors, MLDA should be more robust than the common LDA algorithms.

Finally, regarding the implementation of MLDA in high-dimensional problems like face recognition, a remark should be made. To avoid the singularity of the within-class scatter matrix in the high-dimensional observation space, we would like to adopt the strategy used in Fisherfaces [4]. That is, PCA is first used for dimension reduction, and then MLDA is performed in the PCA-transformed space.

IV. Experiments

A. Experiment Using the Yale Database

The Yale face database [15] contains 165 images of 15 individuals (each person has 11 different images) under various facial expressions and lighting conditions. Each image was manually cropped and resized to 100×80 pixels in our experiment. The sample images of one person are shown in Fig. 2.

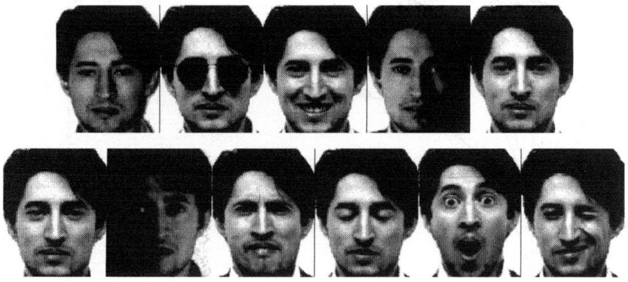
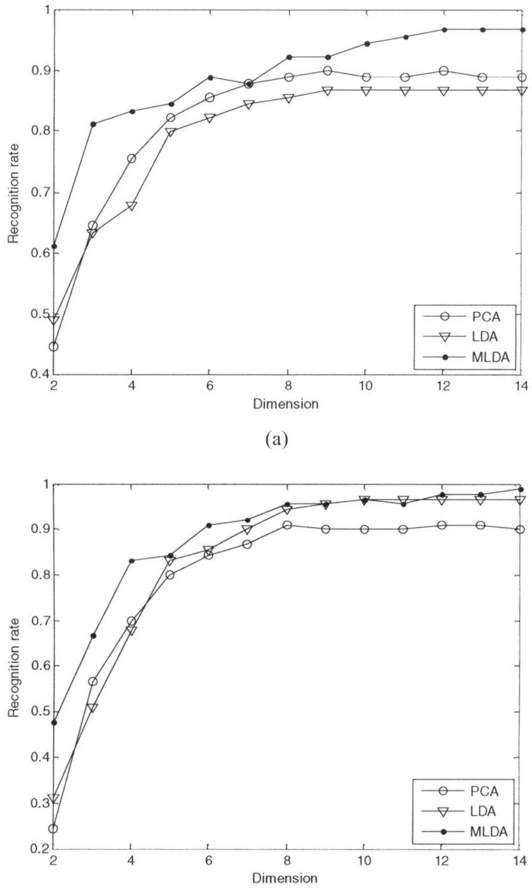


Fig. 2 Sample images of one person in the Yale database.



(a)

(b)

Fig. 3 Recognition rates of PCA, LDA, and MLDA versus the dimensions on the Yale database, (a) the Euclidean distance is used; (b) the cosine distance is used.

The experiment was performed using the first five images (i.e., center-light, w/glasses, happy, left-light, w/no glasses) per class for training, and the remaining six images (i.e., normal, right-light, sad, sleepy, surprised, and wink) for test. Note that in the training set, the images with left-light in the training set can be viewed as outliers. PCA (Eigenfaces) [4], LDA (Fisherfaces) [4], and the proposed MLDA are, respectively, used for feature extraction. In the PCA phase of Fisherface and MLDA, we select the number of principal components as 40. After feature extraction, the nearest neighbor classifiers with Euclidean distance and cosine distance are respectively employed for classification. The recognition rate curves versus the variation of dimensions are illustrated in Fig. 3.

Fig. 3 shows that MLDA significantly outperforms LDA and PCA when Euclidean distance is used, and that MLDA outperforms LDA in most dimensions when cosine distance is used. The maximal recognition rate of MLDA with cosine distance is 98.9% as the dimension is 14, while that of LDA is only 96.7%. These results show MLDA is more robust to outliers than LDA.

Besides, this experiment also show that cosine distance is more effective than Euclidean distance for each method. Thus, we will only use cosine distance in the following experiments.

B. Experiment Using the ORL Database

The ORL (or called AT&T) database [16] contains face images from 40 subjects, each providing 10 different images. For some subjects, the images were taken at different times, varying the lighting, facial expressions and facial details. The size of each image is 92x112 pixels, with 256 grey levels per pixel.

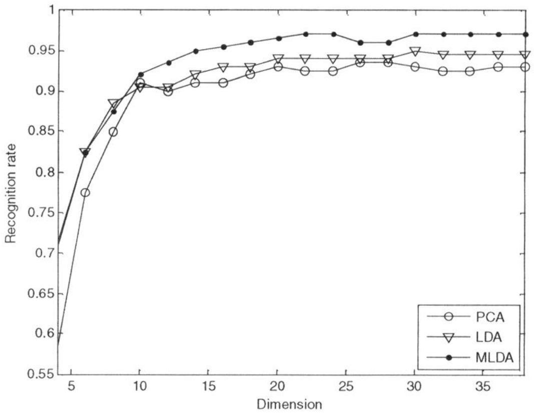


Fig. 4 Recognition rates of PCA, LDA, and MLDA versus the dimensions on the ORL database when cosine distance is used.

TABLE 1
THE MAXIMAL RECOGNITION RATES (%) OF PCA, LDA, AND MLDA USING COSINE DISTANCE ON THE ORL DATABASE AND THE CORRESPONDING DIMENSIONS

Method	PCA	LDA	MLDA
Recognition rate	93.5	95.0	97.0
Dimension	26, 28	30	22, 24, 30-38

In our experiments, the first 5 images of each individual are used for training, and the remaining 5 images are used for test. PCA, LDA and MLDA are, respectively, used for feature extraction. In the PCA phase of LDA and MLDA, the number of principal components is set as 80. Finally, a nearest-neighbor classifier with cosine distance is employed for classification. The recognition rate versus the dimension is plotted in Fig. 4. Fig. 4 indicates MLDA consistently performs better than LDA and PCA as the dimension varies from 10 to 39. Table 1 shows the maximal recognition rate of MLDA is 97.0%, while that of LDA is 95.0%.

C. Experiment Using the AR Database

The AR face [17, 18] contains over 4,000 color face images of 126 people (70 men and 56 women), including frontal views of faces with different facial expressions, lighting conditions and occlusions. The pictures of 120 individuals (65 men and 55 women) were taken in two sessions (separated by two weeks) and each section contains 13 color images. All face images of these 120 individuals are used in our experiment. The face portion of each image is manually cropped and then normalized to 50 x 40 pixels. The sample images of one person are shown in Fig. 5. The details of these images include: (a) neutral expression, (b) smile, (c) anger, (d) scream, (e) left light on, (f) right light on, (g) all sides light on; (h) wearing sun glasses, (i) wearing sun glasses and left light on, (j) wearing sun glasses and right light on, (k) wearing scarf, (l) wearing scarf and left light on, and (m) wearing scarf and right light on.

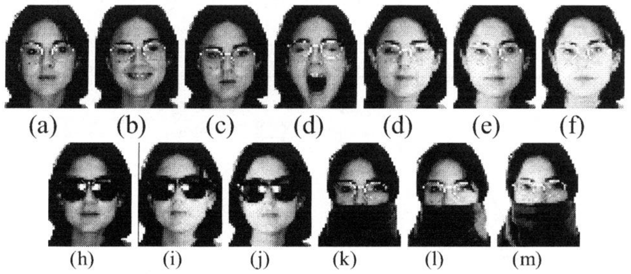
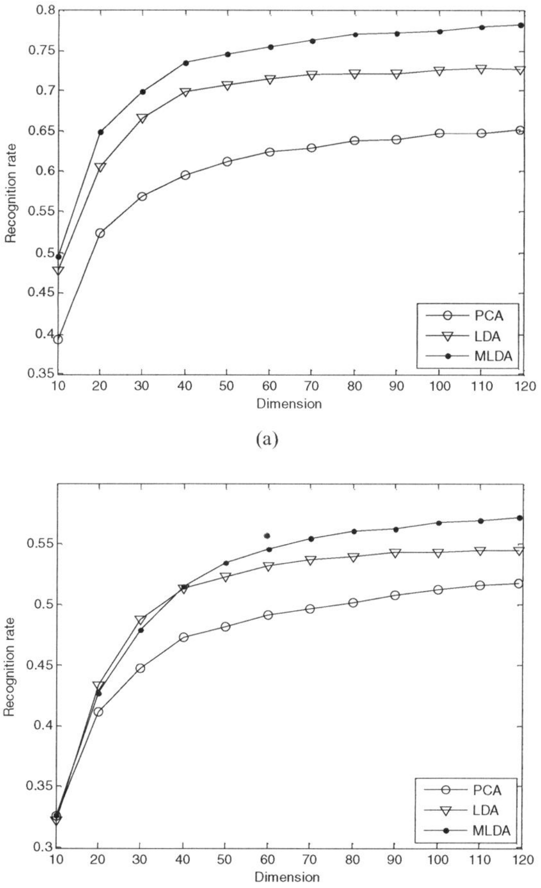


Fig. 5 Sample images of one person in the AR database.

We designed two experiments, depending on whether the occluded images are included in the training sample set or not. In the first experiment, we use three images, i.e. (a), (h) and (k) of each individual for training, the remaining 23 images for test. In the second one, we use another set of three images, i.e. (a), (b) and (c) for training, the remaining 23 images for test. In both experiments, PCA, LDA, and MLDA are, respectively, used for face representation. In the PCA phase of LDA and MLDA, the number of principal components is set as 150. Finally, a nearest-neighbor classifier with cosine distance is employed for classification. The recognition rate curves versus the variation of training sample sizes are shown in Fig. 6, and the maximal recognition rate of each method is listed in Table 2. Fig. 6 and Table 2 shows MLDA significantly outperforms LDA and PCA, whether the training set includes occluded images or not.



(a)

(b)

Fig. 6 Recognition rates of PCA, LDA, and MLDA versus the dimensions on the AR database, (a) the training set includes occluded images; (b) the training set does not include occluded images.

TABLE 2
THE MAXIMAL RECOGNITION RATES (%) OF PCA, LDA, AND MLDA ON THE AR DATABASE WHEN COSINE DISTANCE IS USED

Training set	PCA	LDA	MLDA
{(a), (h), (k)}	65.2	72.8	78.2
{(a), (b), (c)}	51.8	54.5	57.2

V. CONCLUSIONS

In this paper, the class median vector, rather than the class sample average, is used to estimate the class mean vector in the LDA modeling. As an estimator the class mean vector, the class median vector has two main advantages over the class sample average in small sample size cases: (1) the class median (image) vector preserves useful details in the sample images and (2) the class median vector is robust to outliers that exist in training sample set (for example, the images with noise, occlusion, etc). These characteristics make the proposed median LDA (MLDA) model more robust than the common LDA models. We demonstrate the effectiveness of the proposed model using three popular face image databases and show that MLDA outperforms LDA and PCA.

REFERENCES

[1] Andrew Webb, Statistical Pattern Recognition, Arnold, London, 1999.

[2] K. Liu, Y-Q Cheng, J-Y Yang, X. Liu, "An efficient algorithm for Foley-Sammon optimal set of discriminant vectors by algebraic method", *International Journal of Pattern Recognition and Artificial Intelligence*, 1992, 6(5), pp. 817-829.

[3] Daniel L. Swets and John Weng. "Using discriminant eigenfeatures for image retrieval", *IEEE Trans. Pattern Anal. Machine Intell.*, 1996,18(8), pp. 831-836.

[4] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriengman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection", *IEEE Trans. Pattern Anal. Machine Intell.* 1997, 19 (7), pp. 711-720.

[5] L. F. Chen, H. Y. M. Liao, J. C. Lin, M. D. Kao, and G. J. Yu. "A new LDA-based face recognition system which can solve the small sample size problem", *Pattern Recognition*, 2000, 33(10), pp. 1713-1726.

[6] Z. Jin, J.Y. Yang, Z.S. Hu, Z. Lou, Face Recognition based on uncorrelated discriminant transformation, *Pattern Recognition*, 2001,34(7), 1405-1416.

[7] H. Yu, J. Yang. "A direct LDA algorithm for high-dimensional data—with application to face recognition", *Pattern Recognition*, 34(10) (2001) 2067-2070.

[8] J. Yang, J.Y. Yang, "Why can LDA be performed in PCA transformed space?" *Pattern Recognition*, 2003, 36(2), pp. 563-566.

[9] C. J. Liu and H. Wechsler. "Robust coding schemes for indexing and retrieval from large face databases", *IEEE Trans. Image Processing*, 2000, 9(1), 132-137.

[10] M. H. Yang, "Kernel Eigenfaces vs. kernel Fisherfaces: face recognition using kernel methods", *Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition* (RGR'02), Washington D. C., May, 2002, pp. 215-220.

[11] J. Yang, A. Frangi, J.Y. Yang, D. Zhang, and Z. Jin, "KPCA plus LDA: a complete kernel Fisher discriminant framework for feature extraction and recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, 2005, 27 (2), 230-244.

[12] G. R., Grimmett and D. R. Stirzaker. Probability and Random Processes, 2nd Edition, Clarendon Press, Oxford, 1992

[13] Wikipedia, "the free encyclopedia", http://en.wikipedia.org/wiki /Median

[14] A. Marion, "An Introduction to Image Processing", Chapman and Hall, 1991.

[15] Yale face database, http://cvc.yale.edu/projects/yalefaces /yalefaces.html

[16] The AT&T face database, http://www.uk.research.att.com /facedatabase.html

[17] A.M. Martinez and R. Benavente, "The AR Face Database", http://rvl1.ecn.purdue.edu/~aleix/aleix_face_DB.html.

[18] A.M. Martinez and R. Benavente, "The AR Face Database", CVC Technical Report #24, June 1998.