Available online at www.sciencedirect.com
ScienceDirect
Neurocomputing 71 (2008) 1857–1865
NEUROCOMPUTING
www.elsevier.com/locate/neucom

An approach for directly extracting features from matrix data and its application in face recognition

Yong Xu[a,*], David Zhang[b], Jian Yang[c], Jing-Yu Yang[c]

[a]Biometrics Computing Center, Shenzhen Graduate School, Harbin Institute of Technology, Sili, Shenzhen 518055, China
[b]Biometrics Research Centre, Department of Computing, Hong Kong Polytechnic University, Kowloon, Hong Kong
[c]School of Computer Science & Technology, Nanjing University of Science & Technology, Nanjing, China

Available online 29 February 2008

Abstract

By formulating two-dimensional principle component analysis (2DPCA) as a mathematical form different from the conventional 2DPCA, we present theoretical basis of 2DPCA and show the theoretical similarities and differences between 2DPCA and PCA. We also show that 2DPCA owns its decorrelation property and the feature vectors extracted from matrices are uncorrelated. We use the proposed mathematical form to show that 2DPCA is the best approach for directly extract features from matrices. We also present in detail 2DPCA Schemes 1 and 2, two schemes to implement the proposed mathematical form. The two schemes transform original images into different spaces, respectively, 2DPCA Scheme 1 enhances the transverse characters of images, whereas the second scheme enhances vertical characters of images. We propose a feature fusion approach for achieving better recognition results by combining the features generated from the two schemes of 2DPCA. The proposed fusion approach is tested on face recognition tasks and is found to be more accurate than both 2DPCA Scheme 1 and 2DPCA Scheme 2.

Keywords: Principal component analysis; Feature extraction; Feature fusion; Face recognition

1. Introduction

Principal component analysis (PCA) [1–10] is a widely applied dimension reduction and feature extraction technique. It has been used in handprint recognition [7], the recognition of man-made objects [6], industrial robotics [18], and image-based recognition systems [4,13]. PCA is generally implemented on image data as follows: first an image matrix is converted into a vector by concatenating its columns or rows. Then eigenvectors of the covariance matrix or correlation matrix of these vectors are used as transforming axes to obtain their principal components. PCA has been shown to be effective [4,5,9,10,13,14,18,19, 21–24,28] but it does suffer from two particular problems. First, if the number of training samples is small and the data are high-dimensional, it is difficult to accurately estimate the covariance (or correlation) matrix. Second, because the one-dimensional vector space derived from images is usually of very large dimensionality, implementation of PCA is usually very time consuming [24,25].

Responding to these drawbacks, two-dimensional PCA (2DPCA), a novel transform technique derived from the PCA technique, directly extracts features from image matrices [25,26]. Note that 2DPCA as the generation matrix takes the covariance matrix (or correlation matrix) of the image matrix rather than the corresponding one-dimensional vector. 2DPCA calculates directly the projection of a matrix onto the transforming axis. 2DPCA is much more efficient than PCA [25], requiring less memory and having a lower computational cost and has obtained promising experimental results in the areas of feature extraction and dimension reduction. A further difference between traditional PCA and 2DPCA is that in PCA every feature that is extracted is a scalar whereas in 2DPCA every extracted feature is a vector, hereafter called a feature vector. However, it is not known whether the approach is theoretically well-founded.

In this paper we will analyze the theoretical basis of 2DPCA and propose a new 2DPCA-based feature fusion

*Corresponding author. Tel.: +86 755 26032458.
E-mail address: laterfall2@yahoo.com.cn (Y. Xu).

0925-2312/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2007.09.021

approach that combines the feature extraction results of two 2DPCA implementation schemes. We will further show that under the proposed mathematical form of 2DPCA, certain fine theoretical properties hold, for example decorrelation.

Some previous literatures also provide valuable investigation of 2DPCA. For example, Wang et al. [15–17] demonstrated that 2DPCA was equivalent to a special case of block-based PCA [3]. Xu et al. [20] constructed two transformation matrices based on the 2DPCA technique, and performed two transforms to obtain features of a matrix. Motivated by the successes of the two-dimensional linear discriminant analysis, Tao et al. [11,12] developed a general tensor discriminant analysis. Ye et al. proposed generalized principal component analysis (GPCA) in [27]. GPCA also works directly with images in their native state, as two-dimensional matrices, by projecting the images to a vector space that is the tensor product of two lower-dimensional vector spaces.

The differences between our 2DPCA schemes and the previous approaches are as follows. Liwei Wang's approach [15,17] seems to be computationally equivalent to our 2DPCA scheme 1. However, he also did not analyze the theoretical basis of 2DPCA whereas we analyze this and indicate the decorrelation property of 2DPCA. Different from Xu's approach [20] of consecutively performing two transforms to obtain features of a matrix, this paper focuses on fusing two classes of image features obtained using two different implementation schemes of 2PCA to improve face recognition performance. That is, the transform matrices generated from 2DPCA Scheme 1 and 2DPCA Scheme 2 were first used to transform image matrices into two classes of features. Then the two classes of features were fused for face recognition by using a matching score fusion approach. Although from the point of view of methodology, GPCA [27] and 2DPCA belong to the same class of technique, that is, they can both extract features directly from a two-dimensional matrix, no closed form solution exists for GPCA and it cannot be proved theoretically that the two 2DPCA schemes presented in this paper are special cases of GPCA.

The rest of the paper is organized as follows. Section 2 formally presents 2DPCA and its theoretical basis and introduces two 2DPCA implementation schemes. Section 3 presents the characteristics of the reconstruction images, respectively, associated with the two implementation schemes. Section 4 proposes the 2DPCA-based feature fusion approach. Section 5 offers a brief Conclusion.

2. Theoretical basis and implementation schemes of 2DPCA

Suppose there are M images and A_1, $A_2,\dots,A_M$ are, respectively, the matrices corresponding to these images. The conventional 2DPCA [25] as the generation matrix takes the following covariance matrix:

$$G_t=\frac{1}{M}\sum_{i=1}^{M}((A_i-\bar{A})^{\mathrm{T}}(A_i-\bar{A})),$$

where $\bar{A}$ is the mean of all the image matrices. Eigenvalues and eigenvectors of G_t should be first determined, and then k eigenvectors associated with the k largest eigenvalues are selected as transforming axes. The conventional 2DPCA projects an image matrix onto these transforming axes, respectively, and regards the resultant k projections (k vectors) as features of the image [25]. One of the advantages of 2DPCA is that it is much more efficient than PCA [25]. In addition, 2DPCA has obtained promising experimental results in the areas of feature extraction and dimension reduction. However, it is not known whether the approach is theoretically well-founded.

2.1. 2DPCA Scheme 1

In this paper, 2DPCA Scheme 1 is referred to as the 2DPCA technique based on the generation matrix $\Sigma_1=E(A^{\mathrm{T}}A)$, where A stands for a two-dimensional matrix. Note that the generation matrix of 2DPCA Scheme 1 is formally different from that of the conventional 2DPCA, thus we say that 2DPCA Scheme 1 formulates the 2DPCA technique as a new mathematical form. This subsection will address this issue of whether the 2DPCA technique is theoretically well-founded. Indeed, our analysis will demonstrate that 2DPCA Scheme 1 is able to produce the minimal reconstruction error and uncorrelated feature vectors. Suppose that non-decreasing eigenvalues of Σ_1 are $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_n$. 2DPCA Scheme 1 takes the r eigenvectors corresponding to the first r largest eigenvalues of Σ_1 as transforming axes to directly extract features from a matrix. Using 2DPCA Scheme 1, we can project a matrix onto a transforming axis to produce a feature vector (column-feature-vector). If 2DPCA Scheme 1 exploits multiple transforming axes for feature extraction, the feature extraction results will be multiple column-feature-vectors, which can form a new matrix. In this sense, we say that 2DPCA Scheme 1 transforms an original matrix into a new matrix with smaller dimension. We begin with the following theorem to analyze theoretical basis of 2DPCA Scheme 1.

Theorem 1. *Measured using mean squared error, 2DPCA is the best technique for directly transforming matrices into feature vectors as feature vectors obtained using the 2DPCA technique allow matrices to be reconstructed with the minimum mean-square reconstruction error.*

Proof. Suppose that image matrix A can be accurately expressed in terms of

$$A=\sum_{i=1}^{n}v_iu_i^{\mathrm{T}},\quad 1\leqslant i,\quad j\leqslant n, \tag{1}$$

where

$$u_i^{\mathrm{T}}u_j=\begin{cases}1 & i=j,\\ 0 & i\neq j,\end{cases}\quad u_i\,(1\leqslant i\leqslant n)$$

are basis vectors, and vectors v_i $(1\leqslant i\leqslant n)$ are the corresponding coefficients. If $\hat{A}=\sum_{i=1}^{r} v_i u_i^{\mathrm{T}} r<n$, then the deviation between $\hat{A}$ and A will be

$$A-\hat{A}=\sum_{i=r+1}^{n} v_i u_i^{\mathrm{T}}$$

and their mean squared error will be

$$\begin{aligned}E(\|A-\hat{A}\|_{\mathrm{F}}^2)&=E[tr((A-\hat{A})^{\mathrm{T}}(A-\hat{A}))]\\&=E[tr((A-\hat{A})(A-\hat{A})^{\mathrm{T}})],\end{aligned} \quad (2)$$

where $\|\cdot\|_{\mathrm{F}}^2$ denotes the square of the Frobenius norm of a matrix.

According to (1), we have

$$(A-\hat{A})(A-\hat{A})^{\mathrm{T}}=\sum_{i=r+1}^{n} v_i v_i^{\mathrm{T}} \quad \text{and} \quad v_i=Au_i.$$

This then allows

$$(A-\hat{A})(A-\hat{A})^{\mathrm{T}}=\sum_{i=r+1}^{n} Au_i u_i^{\mathrm{T}} A^{\mathrm{T}}$$

to be derived. As a result, it is certain that

$$\begin{aligned}E(\|A-\hat{A}\|_{\mathrm{F}}^2)&=E[\mathrm{tr}((A-\hat{A})(A-\hat{A})^{\mathrm{T}}]\\&=\sum_{i=r+1}^{n} u_i^{\mathrm{T}} E(A^{\mathrm{T}}A)u_i.\end{aligned} \quad (3)$$

$\hat{A}$ will be the best approximate to A, only if (3) reaches its minimum. If $\hat{A}$ is the best approximation to A, the corresponding u_i $(1\leqslant i\leqslant n)$ will be the optimal coordinate axes for expressing A.

To obtain the optimal coordinate axes for expressing A, we introduce the following Langragian function

$$f(u_i)=\sum_{i=r+1}^{n} u_i^{\mathrm{T}} E(A^{\mathrm{T}}A)u_i-\sum_{i=r+1}^{n} \lambda_i(u_i^{\mathrm{T}}u_i-1).$$

Obviously $f(u_i)$ and (3) will reach their extremes simultaneously. Moreover, (3) will reach its extreme, if and only if the derivative of $f(u_i)$ with regard to u_i is equal to zero. Equating the derivative of $f(u_i)$ with respect to u_i to zero, we have $\Sigma_1 u_i - \lambda_i u_i=0$, where $\Sigma_1=E(A^{\mathrm{T}}A)$. Thus, under the condition $\Sigma_1 u_i=\lambda_i u_i$, (3) will obtain the extreme value. In other words, if $u_1,u_2,\ldots,u_n$ are eigenvectors of the following eigen-equation

$$\Sigma_1 u=\lambda u, \quad (4)$$

(3) will be minimized. Consequently, if the r eigenvectors associated with the r largest eigenvalues of (4) are used to reconstruct A in terms of

$$\hat{A}=\sum_{i=1}^{r} v_i u_i^{\mathrm{T}}$$

they will produce the minimum deviation between A and $\hat{A}$. This completes the proof of Theorem 1. Indeed, Theorem 1 shows that 2DPCA is the best approach for directly extract features from matrices because 2DPCA allows the information loss caused by the transformation process to be minimized.

The generation matrix of 2DPCA Scheme 1, Σ_1, is usually evaluated by

$$\Sigma_1=\frac{1}{M}\sum_{i=1}^{M}(A_i^{\mathrm{T}}A_i),$$

where M is the number of training images, and A_i is the matrix corresponding to the ith training sample. After the eigenvector u_i of

$$\Sigma_1=\frac{1}{M}\sum_{i=1}^{M}(A_i^{\mathrm{T}}A_i)$$

is obtained, the vector v_i as determined by $v_i=Au_i$ is called the ith feature vector (or column-feature-vector) of the matrix A.

Notice that 2DPCA Scheme 1 has the theoretical basis as shown in Theorem 1; however, when the conventional 2DPCA was originally proposed as a feature extraction approach, no any theoretical analysis was provided [25]. The relationship between 2DPCA Scheme 1 and the conventional 2DPCA [25] is as follows: if images are centered in advance, the generation matrix Σ_1 of 2DPCA Scheme 1 will be evaluated by

$$\frac{1}{M}\sum_{i=1}^{M}((A_i-\bar{A})^{\mathrm{T}}(A_i-\bar{A})),$$

where $\bar{A}$ is the mean of all the matrices. Thus in this case Σ_1 is computationally equivalent to the generation matrix G_t of the conventional 2DPCA. As a result, in this special case, 2DPCA Scheme 1 is computationally equivalent to the conventional 2DPCA, whereas in other cases 2DPCA Scheme 1 is formally different from the conventional 2DPCA.

2.2. 2DPCA Scheme 2

2DPCA Scheme 2 can describes as follows. Suppose that

$$A=\sum_{i=1}^{n} v_i u_i^{\mathrm{T}}, \quad v_i^{\mathrm{T}}v_j=\begin{cases}1 & i=j,\\0 & i\neq j,\end{cases}$$

i.e. $v_1,v_2,\ldots,v_n$ are treated as basis vectors for expressing A and vectors $u_1,u_2,\ldots,u_n$ are the corresponding coefficients $\hat{A}=\sum_{i=1}^{r} v_i u_i^{\mathrm{T}} r<n$, then the mean squared error of between $\hat{A}$ and A is

$$\begin{aligned}E(\|A-\hat{A}\|_{\mathrm{F}}^2)&=E[tr((A-\hat{A})^{\mathrm{T}}(A-\hat{A})]\\&=\sum_{i=r+1}^{n} v_i^{\mathrm{T}} E(AA^{\mathrm{T}})v_i.\end{aligned}$$

Thus, we can conclude that the r optimal basis vectors are the r eigenvectors associated with the r largest eigenvalues of $\Sigma_2 v=\lambda v$, where $\Sigma_2=E(A\ A^{\mathrm{T}})$. The proof of this statement is similar to that of Theorem 1. Σ_2 can be evaluated using

$$\frac{1}{M}\sum_{i=1}^{M}(A_iA_i^{\mathrm{T}}),$$

where M still represents the number of the training samples. The feature extraction scheme that exploits the eigenvectors of Σ_2 as transforming axes is called 2DPCA Scheme 2. $v^{\mathrm{T}}A$ is a feature vector of matrix A and called row-feature-vector. In addition, it can be also proved that if images are centered in advance, 2DPCA Scheme 2 presented above will be computationally equivalent to the alternative form of the conventional 2DPCA presented in Ref. [25].

2.3. Decorrelation property of 2DPCA

In this subsection, we will demonstrate that 2DPCA is also a decorrelation technique, i.e. 2DPCA decorrelates rows or column vectors of the matrix. In the PCA decorrelation technique, if every sample is a p-dimensional vector and the samples are transformed into a novel space using PCA, the components of the data in the resultant space will be statistically uncorrelated to each other. For the image, PCA eliminates the correlation between the pixels to obtain uncorrelated PCA components. Similarly, 2DPCA is able to obtain uncorrelated feature vectors of the original image according to the following definition and theorem.

Definition 1. The correlation coefficient between two feature vectors v_i and v_j is defined as

$$\rho(v_i, v_j) = \mathrm{cov}(v_i, v_j)/(\sqrt{D(v_i)\cdot D(v_j)}),$$

where

$$\mathrm{cov}(v_i, v_j) = E[[v_i - E(v_i)]^{\mathrm{T}}[v_j - E(v_j)]],$$

$$D(v_i) = E[[v_i - E(v_i)]^{\mathrm{T}}[v_i - E(v_i)]],$$

$$D(v_j) = E[[v_j - E(v_j)]^{\mathrm{T}}[v_j - E(v_j)]].$$

This definition will be valid as long as $D(v_i)>0$ and $D(v_j)>0$. If $\rho(v_i,v_j) = 0$, we say that feature-vectors v_i and v_j are uncorrelated with each other.

Theorem 2. *Feature vectors obtained using 2DPCA Schemes 1 or 2 are statistically uncorrelated to each other under the condition that the mean of original image matrices is zero or they have been centered in advance.*

Now we prove this theorem using 2DPCA Scheme 1 and Definition 1. Because $v_i = Au_i$, $v_j = Au_j$ and u_i, u_j are two eigenvectors of Σ_1, we have

$$\begin{aligned}\mathrm{cov}(v_i, v_j) &= E[u_i^{\mathrm{T}}[A^{\mathrm{T}} - E(A^{\mathrm{T}})][A - E(A)]u_j] \\ &= u_i^{\mathrm{T}}E[[A^{\mathrm{T}} - E(A^{\mathrm{T}})][A - E(A)]]u_j,\end{aligned}$$

$$E[[A^{\mathrm{T}} - E(A^{\mathrm{T}})][A - E(A)]],$$

is usually evaluated using

$$G_t = \frac{1}{M}\sum_{i=1}^{M}((A_i - \bar{A})^{\mathrm{T}}(A_i - \bar{A})),$$

so we have $\mathrm{cov}(v_i, v_j) = u_i^{\mathrm{T}}G_tu_j$. If the mean of original image matrices is zero or they have been centered ahead of time, G_t will be computationally equivalent to Σ_1 as defined in Section 2.1. As a result, we have $\mathrm{cov}(v_i, v_j) = u_i^{\mathrm{T}}\Sigma_1u_j$. Since u_j is the jth eigenvector of Σ_1, we have $\Sigma_1u_j = \lambda_ju_j$. Thus it is certain that $\mathrm{cov}(v_i, v_j) = \lambda_ju_i^{\mathrm{T}}u_j$. Obviously, if $i\neq j$, then $\mathrm{cov}(v_i, v_j) = 0$. Because v_i,v_j are feature vectors generated from u_i,u_j, we obtain

$$D(v_i) = \lambda_iu_i^{\mathrm{T}}u_i = \lambda_i, \quad D(v_j) = \lambda_ju_j^{\mathrm{T}}u_j = \lambda_j.$$

If u_i,u_j are eigenvectors corresponding to non-zero eigenvalues, $D(v_i)>0$, $D(v_j)>0$ will be satisfied. As a result, the correlation coefficient presented in Definition 1 is valid for arbitrary two eigenvectors associated with non-zero eigenvalues. Since $i\neq j$ implies $\mathrm{cov}(v_i,v_j) = 0$, $\rho(v_i,v_j) = 0$ is always satisfied for arbitrary v_i and v_j corresponding to non-zero eigenvalues. In other words, using 2DPCA Scheme 1, we can obtain uncorrelated column-feature-vectors. Similarly, using 2DPCA Scheme 2, we can obtain uncorrelated row-feature-vectors. This means that Theorem 2 has been proved and that 2DPCA can be also viewed as a decorrelation technique for eliminating correlation between vectors.

Between 2DPCA and PCA there exist clear similarities and differences. There are two major similarities between 2DPCA and PCA. As we know, the traditional PCA-based feature extraction is the best technique for extracting features from vectors for it allows original vectors to be reconstructed with the minimum mean squared error. Similarly, 2DPCA is the best technique for directly extracting features from matrices as it obtains less reconstruction error than other techniques of directly extracting features from matrices. This is the first similarity between 2DPCA and PCA. The second similarity between 2DPCA and PCA is that they are both decorrelation techniques.

There are three main differences between 2DPCA and PCA. First, 2DPCA extracts features from a matrix by directly projecting the matrix onto the transforming axes whereas PCA extracts features from a matrix by projecting the corresponding one-dimensional vector onto its transforming axes. Indeed, PCA usually transforms an image matrix into a vector whereas 2DPCA converts a matrix into a new matrix. PCA is able to eliminate statistical correlation among matrix elements whereas 2DPCA tries to reduce correlation among row or column vectors of the matrix. Second, 2DPCA extracts features from matrices more efficiently than PCA. Third, implementing 2DPCA based on matrices rather than the corresponding one-dimensional vectors makes it easier to preserve the spatial structure of matrices. In contrast, in PCA, some adjacent elements of a two-dimensional matrix will not be adjacent when the matrix is converted into the corresponding one-dimensional vector for feature extraction using PCA.

3. Further analysis of two classes of 2DPCA features

3.1. Comparison between two classes of 2DPCA features

Based on the two schemes of 2DPCA, we can get two classes of features of an image. The first class of features consists of the row-feature-vectors produced by 2DPCA Scheme 1 and the second class of features consists of the column-feature-vectors produced by 2DPCA Scheme 2. In this subsection we focus on comparison analysis of the two classes of 2DPCA features. We can reconstruct the image using each of the two classes of features. Fig. 1 shows some reconstruction images of a face image, which are generated from 2DPCA Scheme 1. Fig. 2 presents several reconstruction images of the same face generated from 2DPCA Scheme 2. We can see that the reconstruction images associated with the Scheme 1 enhance the transverse characteristic of the face, whereas the reconstruction images associated with the second scheme enhance the vertical characteristic of the face.

Now we explain the difference between reconstruction images, respectively, obtained using the two 2DPCA schemes from the point of view of numerical computation. Each element of one feature vector generated from 2DPCA Scheme 1 is a weighted sum of elements of one row of the image matrix. However, in Scheme 2, every element of a feature vector is a weighted sum of elements of one column of the image matrix. Therefore, the first scheme emphasizes mainly the row vector information of an image matrix, whereas the second scheme emphasizes primarily the column vector information. This is why the image in Fig. 1 appears to enhance the transverse characteristics of the face whereas in Fig. 2 the vertical characteristics are enhanced. These two classes of features thus complement each other in representing the face and we might expect to get a higher accuracy by fusing the two classes of features. An approach that exploits the information of the sum of the elements of each row vector and that of each column vector of the face image for face recognition can be found in [19].

3.2. Feature fusion approach

To fuse the two classes of features, we propose the following feature fusion approach. We first evaluated the distance d_i $(i = 1,2,\ldots,M)$ between the first class of features of a test sample and that of the ith training sample. Here M is the number of training images. Let $D_{\max}$ be the maximum value among these distances between this test sample and all the training samples. Then d_i was normalized by using $d'_i = d_i/D_{\max}$, $i = 1, 2, \ldots, M$. Features generated from Scheme 2 were similarly used to get normalized distance values e'_i, $i = 1,2,\ldots,M$. We calculated t'_i using $t'_i = d'_i + e'_i$, $i = 1, 2, \ldots, M$. We regarded t'_i as the final measurement of the distance between the test sample and the ith training sample. We obtained $t'_{\min}$ using $t'_{\min} = \min t'_i$, $i = 1,2,\ldots,M$. If $t'_l = t'_{\min}$, we classified the test sample into the class which the lth training sample belongs to. In this paper the above approach is called feature fusion approach.

4. Experiments on face recognition tasks

4.1. Experiment on the ORL database

This experiment was performed on the ORL database. All the face images of the ORL face database were

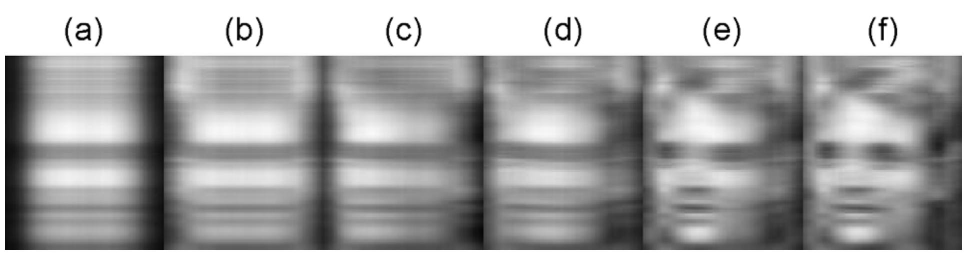


Fig. 1. Reconstructed images obtained using the 2DPCA Scheme 1. (a)–(f) are, respectively, the reconstruction images associated with the first 1,2,3,4,5,6 eigenvectors of the correlation matrix.

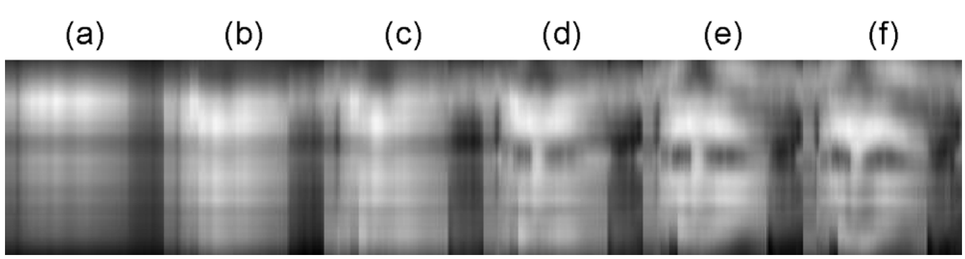


Fig. 2. Reconstructed images obtained using 2DPCA Scheme 2. (a)–(f) are, respectively, reconstructed images associated with the first 1,2,3,4,5,6 eigenvectors of the correlation matrix. Notice that Fig. 1 and Fig. 2 are associated with the same original face image.

obtained against a dark homogeneous background. These images contain various facial expressions (smiling/no smiling, open/closed eyes) and facial detail. The subjects were in up-right, frontal position with tolerance for some tilting and rotation of up to about 20°. For each of the 40 subjects, 10 different images were created. In this experiment the first 4, 5 or 6 face images of all the subjects were, respectively, used as training samples, and the corresponding remaining images were regarded as test samples.

Table 1 shows the classification accuracies of the 2DPCA Scheme 1, Scheme 2, and the feature fusion approach on the ORL database using the first 4,6,8,10 eigenvectors. The feature fusion approach classifies better than both 2DPCA Schemes 1 and 2. Table 2 shows the highest classification accuracies of all the 2DPCA-based approaches and PCA as well as the corresponding feature extraction time. From Table 2, we can see that 2DPCA Schemes 1 and 2 can extract features more efficiently than PCA but they are not as accurate as PCA and the proposed feature fusion approach can perform better than any of the other approaches. Where the first six samples of each subject are used as training samples, the highest classification accuracy of the feature fusion approach is 98.1% whereas the highest classification accuracies of 2DPCA Schemes 1 and 2 and PCA are 93.8%, 91.3% and 95.6%, respectively. The feature fusion approach is also more efficient than PCA-based feature extraction.

Fig. 3 shows the classification accuracies of 2DPCA Schemes 1 and 2, and the feature fusion approach with six training samples per subject on the ORL database. The fact

Table 1
The classification accuracies (%) of 2DPCA Scheme 1 and 2, and the feature fusion approach on the ORL database

Number of training images	2DPCA Scheme 1 The number of eigenvectors				2DPCA Scheme 2 The number of eigenvectors				The feature fusion approach The number of eigenvectors from each scheme			
	4	6	8	10	4	6	8	10	4	6	8	10
4	80.4	82.5	82.5	81.3	78.8	80.0	80.8	81.3	88.3	87.9	88.8	88.8
5	84.0	85.5	85.0	84.0	84.5	83.5	85.5	86.0	92.0	89.5	91.0	91.5
6	92.5	93.8	93.1	91.9	91.3	91.3	91.3	91.3	98.1	97.5	96.9	96.3

Table 2
The highest classification accuracies (%) of all the approaches on the ORL database and the corresponding feature extraction time

Number of training images	PCA	2DPCA Scheme 1	2DPCA Scheme 2	The feature fusion approach
4	88.3 (25.6 s)	82.5 (1.1 s)	81.3 (0.7 s)	89.2 (8.4 s)
5	90.5 (32.9 s)	85.5 (1.3 s)	86.0 (0.9 s)	92.0 (9.8 s)
6	95.6 (37.3 s)	93.8 (1.4 s)	91.3 (0.7 s)	98.1 (11.2 s)

The number in the bracket is the feature extraction time (unit: second) associated with the highest accuracy.

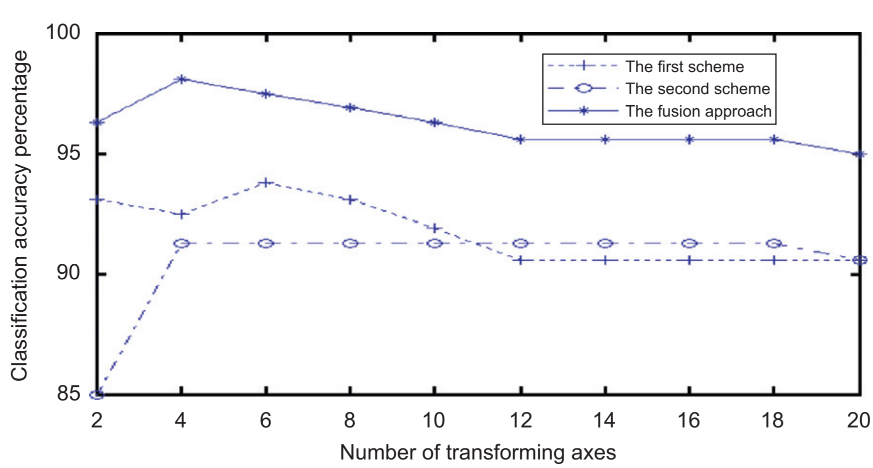


Fig. 3. The classification accuracies on the ORL database of 2DPCA Schemes 1 (denoted by "the first scheme") and 2 (denoted by "the second scheme") and the feature fusion approach with six training samples per class. The y-axis denotes classification accuracy as a percentage. The abscissa represents the number of transforming axes used for feature extraction. Notice that the feature fusion approach uses twice as many transforming axes shown by the abscissa axis.

that the feature fusion approach has the best classification accuracy implies that the two classes of features generated from Schemes 1 and 2 contain complementary image presentation information.

4.2. Experiment on the Yale database

The Yale database contains face images with a variety of expressions such as normal, sad, happy, sleepy, surprised, and winking, all obtained under differing lighting conditions. In some images, the faces wear glasses. The first 1,2, or 3 face images of every subject were, respectively, selected as training samples and the others were used as test samples.

Table 3 shows the classification accuracies of Schemes 1 and 2, and the feature fusion approach on the Yale database. The feature fusion approach has the best recognition accuracy. Table 4 shows the highest classification accuracies of all the approaches and the corresponding feature extraction time on the Yale database. Schemes 1 and 2 are more efficient than PCA but not as accurate. The feature fusion approach is also more accurate than 2DPCA Schemes 1 and 2 and it extracts the corresponding features faster than PCA-based feature extraction (as shown in Table 4). Fig. 4 shows the classification accuracies of 2DPCA Scheme 1, 2DPCA Scheme 2, and the feature fusion approach with three training samples per class on the Yale database. This experiment shows again that the two classes of features obtained using 2DPCA Schemes 1 and 2 contain complementary image presentation information.

Table 3
The classification accuracies (%) of the first scheme, the second scheme and the feature fusion approach of 2DPCA on the Yale database

Number of training images	2DPCA Scheme 1 The number of eigenvectors				2DPCA Scheme 2 The number of eigenvectors				The feature fusion approach The number of eigenvectors from each scheme			
	4	6	8	10	4	6	8	10	4	6	8	10
1	41.3	42.7	46.7	44.7	37.3	39.3	39.3	38.7	50.0	50.7	53.3	54.7
2	55.2	60.0	57.8	56.6	47.4	52.6	51.9	51.9	60.0	65.2	64.4	65.9
3	73.3	70.8	70.0	71.7	67.5	70.0	70.0	69.0	75.0	79.2	79.2	80.0

Table 4
The highest classification accuracies (%) of all the approaches on the Yale database and the corresponding feature extraction time

Number of training images	PCA	2DPCA Scheme 1	2DPCA Scheme 2	The feature fusion approach
1	54.0 (6.0 s)	46.7 (0.3)	39.3 (0.3)	54.7 (2.1 s)
2	65.9 (6.2 s)	65.2 (0.3 s)	52.6 (0.3 s)	65.9 (2.8 s)
3	83.3 (7.0 s)	73.3 (0.3 s)	70.0 (0.3 s)	82.7 (3.6 s)

The number in the bracket is the feature extraction time (unit: second) associated with the highest accuracy.

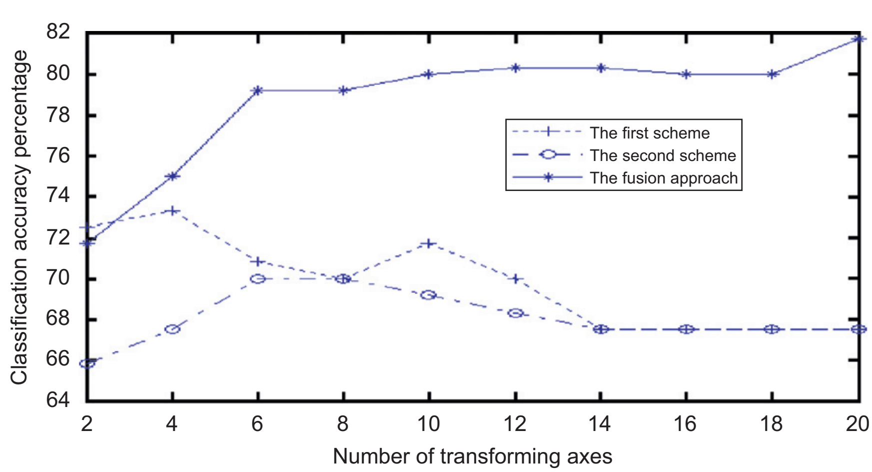


Fig. 4. The classification accuracies on the Yale database of 2DPCA Schemes 1 (denoted by "the first scheme") and 2 (denoted by "the second scheme") and the feature fusion approach with six training samples per class. The y-axis denotes classification accuracy as a percentage. The abscissa represents the number of transforming axes used for feature extraction. Notice that the feature fusion approach uses twice as many transforming axes shown by the abscissa axis.

5. Conclusion

2DPCA, which directly transforms image matrices into feature vectors, is an efficient feature extraction technique for two-dimensional matrix. In this paper, we firstly formulated 2DPCA as a new mathematical form having two implementation schemes and then used the proposed mathematical form to show the theoretical basis of 2DPCA. It is shown that 2DPCA is the best approach for directly extract features from matrices. This means that both 2DPCA Schemes 1 and 2 are effective means of feature extraction and dimension reduction. The two different schemes transform original images into two different spaces. 2DPCA Scheme 1 enhances the transverse characteristic of face images, whereas 2DPCA Scheme 2 enhances the vertical characteristic of face images. In this sense, the two classes of features obtained using 2DPCA Schemes 1 and 2 are complementary for face image presentation. We designed a feature fusion approach to combine the features generated from the first scheme and the second scheme. The experimental results on the ORL and Yale face image databases show that the proposed feature fusion approach is able to greatly improve the classification accuracy of 2DPCA.

Acknowledgements

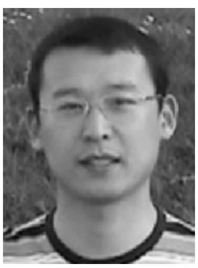
We wish to thank National Natural Science Foundation of China (Nos. 60602038, 60632050, 60620160097) and Natural Science Foundation of Guangdong province, China (No. 06300862) for supporting.

References

[1] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, Machine Press, Beijing, China, 2004.
[2] K. Fukunaga, Introduction to Statistical Pattern Recognition, second ed., Academic Press Inc., New York, 1990.
[3] R. Gottumukkal, V.K. Asari, An improved face recognition technique based on modular PCA approach, Pattern Recognition Lett. 25 (4) (2004) 429–436.
[4] M. Kirby, L. Sirovich, Application of the KL Procedure for the Characterization of Human Faces, IEEE Trans. Pattern Anal. Mach. Intell. 12 (1) (1990) 103–108.
[5] S. Mika, B. Schölkopf, A.J. Smola, K.R. Müller, M. Scholz, G. Rätsch, Kernel PCA and de-noising in feature spaces, in: M.S. Kearns, S.A. Solla, D.A. Cohn (Eds.), Adv. Neural Inf. Process. Systems, vol. 11, MIT Press, Cambridge, MA, 1999, pp. 536–542.
[6] H. Murase, S. Nayar, Visual learning and recognition of 3-D objects from appearance, Int. J. Comput. Vision 14 (1) (1995) 5–24.
[7] H. Murase, F. Kimura, M. Yoshimura, Y. Miyake, An improvement of the auto-correlation matrix in pattern matching method and its applications to handprinted 'HIRAGANA', Trans. IECE. J 64-D (3) (1981).
[8] E. Oja, Neural networks, principal components, and subspaces, Int. J. Neural Systems 1 (1) (1989) 61–68.
[9] S. Romdhani, A. Psarrou, S. Gong, Multi-View nonlinear active shape model using kernel PCA. in: Proceedings of the 10th British Machine Vision Conference, Nottingham, England, 1999, pp. 483–492.
[10] B. Schölkopf, S. Mika, A.J. Smola, G. Rätsch, K.R. Müller, Kernel PCA pattern reconstruction via approximate pre-images, in: Proceedings of the 8th International Conference on Artificial Neural Networks, Berlin, 1998, pp. 147–152.
[11] D. Tao, X. Li, W. Hu, S.J. Maybank, X. Wu, Supervised tensor learning, Knowledge and Information Systems, doi:10.1007/s10115-006-0050-6.
[12] D. Tao, X. Li, X. Wu, S.J. Maybank, General tensor discriminant analysis and gabor features for gait recognition, IEEE Trans. Pattern Anal. Mach. Intell. (29) (2007).
[13] M. Turk, A. Pentland, Face recognition using Eigenfaces, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1991, pp. 586–591.
[14] C. Twining, C. Taylor, The use of kernel principal component analysis to model data distributions, Pattern Recognition 36 (1) (2003) 217–227.
[15] L. Wang, X. Wang, M. Chang, J. Feng, Is two-dimensional PCA a new technique?, Acta Automatica 31 (5) (2005) 782–787.
[16] L. Wang, X. Wang, J. Feng, On image matrix based feature extraction algorithms, IEEE Trans. SMC, Part B. 36 (1) (2006) 194–197.
[17] L. Wang, X. Wang, X. Zhang, J. Feng, The equivalence of two-dimensional PCA to line-based PCA, Pattern Recognition Lett. 26 (1) (2005) 57–60.
[18] J.J. Weng, Crescepton, SHOSLIF, towards comprehensive visual learning, in: S.K. Nayar, T. Poggio (Eds.), Early Visual Learning, Oxford University Press, Oxford, 1996, pp. 183–214.
[19] J. Wu, Z.H. Zhou, Face recognition with one training image per person, Pattern Recognition Lett. 23 (14) (2002) 1711–1719.
[20] A. Xu, X. Jin, Y. Jiang, P. Guo, Complete two-dimensional PCA for face recognition, in: Proceedings of the 18th International Conference on Pattern Recognition, vol. 3, Hong Kong, August 2006, pp. 481–484.
[21] Y. Xu, D. Zhang, F. Song, J.-Y. Yang, Z. Jing, M. Li, A method for speeding up feature extraction based on KPCA, Neurocomputing 70 (4–6) (2007) 1056–1061.
[22] J. Yang, Z. Jin, J.Y. Yang, D. Zhang, Essence of kernel Fisher discriminant: KPCA plus LDA, Pattern Recognition 37 (2004) 2097–2100.
[23] J. Yang, J.Y. Yang, Why can LDA be performed in PCA transformed space?, Pattern Recognition 36 (2) (2003) 563–566.
[24] J. Yang, J.Y. Yang, From image vector to matrix: a straightforward image projection technique—IMPCA vs. PCA, Pattern Recognition 35 (9) (2002) 1997–1999.
[25] J. Yang, D. Zhang, A.F. Frangi, J.Y. Yang, Two dimensional PCA: a new approach to appearance-based face representation and recognition, IEEE Trans. Pattern Anal. Mach. Intell. 24 (1) (2004) 131–137.
[26] J. Yang, D. Zhang, Y. Xu, J.Y. Yang, Two-dimensional discriminant transform for face recognition, Pattern Recognition. 38 (7) (2005) 1125–1129.
[27] J. Ye, R. Janardan, Q. Li, GPCA: an efficient dimension reduction scheme for image compression and retrieval. in: ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD'04), 2004, pp. 354–363.
[28] W. Zhao, R. Chellappa, A. Krishnaswamy, Discriminant analysis of principal components for face recognition, in: Proceedings of International Conference on Automatic Face and Gesture Recognition, Nara, Japan, 1998, pp. 336–341.

Yong Xu was born in Sichuan, China, in 1972. He received his B.S. degree, M.S. degree in 1994 and 1997, respectively. He received the Ph.D. degree in Pattern recognition and Intelligence System at NUST(China) in 2005. Now he works at Shenzhen graduate school, Harbin Institute of Technology. His current interests include biometric, characteristic recognition, machine learning and image processing.

David Zhang graduated in Computer Science from Peking University in 1974. In 1983 he received his M.Sc. in Computer Science and Engineering from (HIT) and then in 1985 his Ph.D. from the same institution. In 1994 he received his second Ph.D. in Electrical and Computer Engineering from the University of Waterloo, Ontario, Canada. Professor Zhang is currently at the Hong Kong Polytechnic University where he is the Founding Director of the Biometrics Technology Center a body supported by the Hong Kong SAR Government. He also serves as Adjunct Professor in Tsinghua University, Shanghai Jiao Tong University, Harbin Institute of Technology, and the University of Waterloo. Professor Zhang's research interests include automated biometrics-based authentication, pattern recognition, and biometric technology and systems. Professor Zhang holds a number of patents in both the USA and China and is a current Croucher Senior Research Fellow.



Jian Yang was born in Jiangsu, China, in 1973. He obtained the B.S. degree, M.S. degree and Ph.D. degree in 1995, 1998 and 2002, respectively. Now, he is a post-doctor at Hong Kong Polytechnic University. His current research interests include biometric such as face recognition, handwritten characteristic recognition and data fusion.

Jing-Yu Yang received the B.S. degree in Computer Science from NUST, Nanjing, China. From 1982 to 1984 he was a visiting scientist at the Coordinated Science Laboratory, University of Illinois at Urbana-Champaign. From 1993 to 1994 he was a visiting professor at the Department of Computer Science, Missuria University. And in 1998, he acted as a visiting professor at Concordia University in Canada. His current research interests are in the areas of pattern recognition, image processing and artificial intelligence, and expert system.