Unified Principal Component Analysis with Generalized Covariance Matrix for Face Recognition

Shiguang Shan[1], Bo Cao[2], Yu Su[1,3], Laiyun Qing[4], Xilin Chen[1], Wen Gao[1,3]
[1] Key Laboratory of Intelligent Information Processing, Chinese Academy of Sciences, Beijing,China
[2] Microsoft Advanced Technology Center, Sigma Center, Haidian District, Beijing, China
[3] School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China
[4] Graduate School of Chinese Academy of Sciences, Beijing, China
{sgshan, bcao, ysu, lyqing, xlchen, wgao}@jdl.ac.cn

Abstract

Recently, 2DPCA and its variants have attracted much attention in face recognition area. In this paper, some efforts are made to discover the underlying fundaments of these methods, and a novel framework called Unified Principal Component Analysis (UPCA) is proposed. First, we introduce a novel concept, named Generalized Covariance Matrix (GCM), which is naturally derived from the traditional Covariance Matrix (CM). Each element of GCM is a generalized covariance of two random vectors rather than two scalar variables in CM. Based on GCM, the UPCA framework is proposed, from which the traditional PCA and its 2D counterparts can be deduced as special cases. Furthermore, under the UPCA framework, we not only revisit the existing 2D PCA methods and their limitations, but also propose two new methods: the grid-sampling method (GridPCA) and the intra-group correlation reduction method. Extensive experimental results on the FERET face database support the theoretical analysis and validate the feasibility of the proposed methods.

1. Introduction

As one of the most successful face recognition methods, Eigenfaces [1] manipulates 1D image vectors formed by concatenating directly the rows of the original 2D face image. Despite its success, one main drawback limits its usability: it is difficult to estimate the covariance matrix stably due to the high dimension of the image vectors and the relatively small size of the training set. Recently, many methods have been developed to overcome this difficulty. Among them, 2DPCA and its variants have attracted much attention. In common, these methods treat the face images as 2D matrices rather than 1D vectors.

In 2DPCA [2] (or IMPCA [3]), face images were directly treated as 2D matrices, based on which some variants were proposed. In [4], the authors proposed to transform the transpose of the face images into image matrices, while in [6] DiaPCA proposed to transform the original face images into diagonal face images by rotating each row one pixel to its right. Inspired by the idea of 2D process, MatPCA [5] was proposed by matrixizing 1D input into 2D matrices and then applying the 2D PCA.

Usually, compared with PCA, 2D PCA methods need more coefficients for image representation. To solve this problem, several alternatives were proposed. For instance, the Bilateral-projection-based 2DPCA (B2DPCA) [7] reduces the redundancies among both rows and columns of the face images by projecting them to the left- and right-multiplying projection matrices. Similarly, two-directional 2DPCA ($(2D)^2$PCA) [4] and Bidirectional PCA (BD-PCA) [8] utilizes the two projection matrices obtained from 2DPCA and A2DPCA[4] respectively.

After investigating carefully the existing 2D PCA methods, we notice that, in spite of the popularity and success of 2D PCA methods in the last few years, there are three open fundamental problems on 2D PCA methods:

1. Why can the 2D PCA methods outperform their 1D counterpart, say, the traditional PCA?
2. What are the reasons behind that well interprets the performance variance of different 2D PCA methods?
3. Is there any underlying theory unifying these 2D PCA methods, based on which potentially better 2D PCA methods can be proposed?

Some efforts do have been made to answer the first question. For instance, according to [2] [4] [6] [7], the advantage of the 2DPCA over PCA is attributed to the smaller covariance matrix (CM) and thus its more stable estimation. This argument actually can be cast back to the overfitting risk of PCA, since 2DPCA is equivalent to the PCA performed on the rows of all the face images [7] or the line-based PCA [9]. This implies that the training set be significantly enlarged. Besides, some researchers believe that the spatial information embedded in the face images are better preserved in 2D PCA methods [7].

These arguments can partly explain the superior performances of the 2D PCA methods over the traditional PCA, but fail to clearly interpret their essential differences, as well as the differences between the variants of 2D PCA methods. This paper tries to fill this gap by proposing a

978-1-4244-2243-2/08/$25.00 ©2008 IEEE

Unified Principal Component Analysis (UPCA) method. The main contributions of this paper are as follows:

1. A definition of Generalized Covariance Matrix (GCM) is proposed. We show that it is a natural extension of the traditional covariance matrix (CM).
2. Based on GCM, a unified framework called Unified Principal Component Analysis (UPCA) is presented, which offers a unified view for understanding and explaining both PCA and the 2D PCA.
3. Deducing from UPCA, we further propose the grid-sampling methods and the intra-group correlation reduction methods to achieve better performances.

The remainder of this paper is organized as follows: In Section 2, the GCM are described. The UPCA is presented in Section 3. In Section 4, the previous 2D PCA methods are revisited and two new methods are proposed. Experiments and discussions are presented in Section 5. Finally, conclusions are presented in Section 6.

2. Generalized Covariance Matrix

In this section, we briefly review the definition of the traditional CM and the difficulty of its stable estimation. To solve this difficulty, we introduce the novel concepts of GCM, which can be more stably estimated.

2.1. Traditional Covariance Matrix

In statistics and probability theory, the Covariance Matrix (CM) is a matrix of covariance between elements of a random vector. Consider two random vectors $\mathrm{X}=[\mathrm{x}_1\ \mathrm{x}_2 \ldots \mathrm{x}_m]^T$ and $\mathrm{Y}=[\mathrm{y}_1\ \mathrm{y}_2 \ldots \mathrm{y}_n]^T$, whose entries have finite variance (to be concise, all the random vectors are assumed to be centered in the paper). The cross covariance matrix $\mathbf{\Sigma}_{\mathrm{XY}}$ between X and Y is the $m \times n$ matrix as follows:

$$\mathbf{\Sigma}_{\mathrm{XY}} = E\left[\mathrm{XY}^T\right]. \tag{1}$$

The covariance matrix of X is defined as

$$\mathbf{\Sigma} = E\left[\mathrm{XX}^T\right]. \tag{2}$$

Generally, $\mathbf{\Sigma}$ indicates the dispersion of the random vector X's distribution. Its trace, $\mathrm{tr}(\mathbf{\Sigma})$, the sum of the diagonal components of $\mathbf{\Sigma}$, is the sum of the variances of each random variable in X. Thus, it is a reasonable measure of the total scatter of X.

Suppose that there are N observations $X_1, X_2, \ldots, X_N$ of X, the sample CM is given by

$$\Sigma = \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)\left(X_i - \overline{X}\right)^T. \tag{3}$$

where $\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$.

2.2. Partitioned Covariance Matrix

In order to study the relationship between groups of random variables, CM can be partitioned. If we consider only the simplest case that each group has equal number (say p) of random variables, the random vector X in Equ.(2) can be partitioned as follows

$$\begin{aligned}\mathrm{X} &= \left[\mathrm{x}_{a_{11}} \cdots \mathrm{x}_{a_{1p}} \,|\, \mathrm{x}_{a_{21}} \cdots \mathrm{x}_{a_{2p}} \,|\cdots|\, \mathrm{x}_{a_{k1}} \cdots \mathrm{x}_{a_{kp}}\right]^T, \\ &= \left[\mathrm{Y}_1^T \quad \mathrm{Y}_2^T \quad \cdots \quad \mathrm{Y}_k^T\right]^T\end{aligned} \tag{4}$$

where each sub-vector Y_i contains p grouped random variables, and $k \times p$ is equal to m, the dimension of X.

Then the partitioned CM can be written as

$$\mathbf{\Sigma} = E\left[\mathrm{XX}^T\right] = \begin{bmatrix} \mathbf{\Sigma}_{\mathrm{Y_1Y_1}} & \mathbf{\Sigma}_{\mathrm{Y_1Y_2}} & \cdots & \mathbf{\Sigma}_{\mathrm{Y_1Y_k}} \\ \mathbf{\Sigma}_{\mathrm{Y_2Y_1}} & \mathbf{\Sigma}_{\mathrm{Y_2Y_2}} & \cdots & \mathbf{\Sigma}_{\mathrm{Y_2Y_k}} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{\Sigma}_{\mathrm{Y_kY_1}} & \mathbf{\Sigma}_{\mathrm{Y_kY_2}} & \cdots & \mathbf{\Sigma}_{\mathrm{Y_kY_k}} \end{bmatrix}. \tag{5}$$

where $\mathbf{\Sigma}_{\mathrm{Y_iY_j}}$ is a $p \times p$ cross covariance matrix between the random vectors Y_i and Y_j.

2.3. Generalized Covariance Matrix

The dimension of the partitioned CM can be greatly reduced if a single scalar instead of the cross covariance matrix $\mathbf{\Sigma}_{\mathrm{Y_iY_j}}$ is used to measure the overall covariance of Y_i and Y_j. One natural choice for this purpose is the trace, $\mathrm{tr}\left(\mathbf{\Sigma}_{\mathrm{Y_iY_j}}\right)$, which is also known as the generalized covariance of Y_i and Y_j. Formally, let $\sigma_{\mathrm{Y_iY_j}} = \mathrm{tr}\left(\mathbf{\Sigma}_{\mathrm{Y_iY_j}}\right)$, then, from Equ.(5), $\mathbf{\Sigma}$ can be simplified as $\mathbf{G}$:

$$\mathbf{G} = \begin{bmatrix} \sigma_{\mathrm{Y_1Y_1}} & \sigma_{\mathrm{Y_1Y_2}} & \cdots & \sigma_{\mathrm{Y_1Y_k}} \\ \sigma_{\mathrm{Y_2Y_1}} & \sigma_{\mathrm{Y_2Y_2}} & \cdots & \sigma_{\mathrm{Y_2Y_k}} \\ \vdots & \vdots & & \vdots \\ \sigma_{\mathrm{Y_kY_1}} & \sigma_{\mathrm{Y_kY_2}} & \cdots & \sigma_{\mathrm{Y_kY_k}} \end{bmatrix} \in R^{k\times k}. \tag{6}$$

We can reform the partitioned m-dimension random vector $\mathrm{X} = \left[\mathrm{Y}_1^T\ \mathrm{Y}_2^T \cdots \mathrm{Y}_k^T\right]^T$ in Equ.(4) to a random matrix $\mathbf{Y} = \left[\mathrm{Y}_1\ \mathrm{Y}_2 \cdots \mathrm{Y}_k\right]^T \in R^{k\times p}$, $\mathbf{G}$ can be rewritten concisely as

$$\mathbf{G} = E\left[\mathbf{Y}\mathbf{Y}^T\right]. \tag{7}$$

Equ.(7) is the same as Equ.(2) except that each element of $\mathbf{G}$ is a *generalized covariance* of two random vectors, whereas each element of $\mathbf{\Sigma}$ is a covariance of two random scalar variables. From this point of view, we call $\mathbf{G}$ the Generalized Covariance Matrix (GCM). Similarly, given N observations $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_N$ of $\mathbf{Y}$, the sample GCM is

$$G = \frac{1}{N-1}\sum_{i=1}^{N}\left(\boldsymbol{Y}_i - \overline{\boldsymbol{Y}}\right)\left(\boldsymbol{Y}_i - \overline{\boldsymbol{Y}}\right)^T. \tag{8}$$

where $\overline{\boldsymbol{Y}} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{Y}_i$.

In the above definition of GCM, how to group the random variables is not restricted if only each group contains equal number of elements. Therefore, in this sense, the image CM [2] [3], matrix CM [5] and diagonal CM [6] can all be regarded as special cases of GCM.

Thus, by grouping the elements of X in different

manners, a set of $\mathbf{G}$, denoted by $\mathrm{S_G(X)}$, can be constructed. Easy to see that, when $p=1$, $\mathbf{G}=\mathbf{\Sigma}$, i.e., GCM recedes to traditional CM. It is also easy to prove that, for any $\mathbf{G}\in\mathrm{S_G(X)}$, $\mathrm{tr}(\mathbf{G})=\mathrm{tr}(\mathbf{\Sigma})$, which means that the total dispersion of the random variables is preserved regardless of the different forms of GCM. However, one can also see that it compresses the covariance matrix between two random vectors into a single scalar, and thus turns into a "coarse" representation of the traditional CM. As p increases, GCM denotes a fine-to-coarse representation of the dispersion of the random variables.

In many real world applications, the number of samples used to estimate the CM is often far less than the feature dimension, which leads to unstable estimations. For GCM, however, the estimation may be more stable attribute to its lower dimension. Therefore, when we choose among the different forms of GCM, there is a tradeoff between accuracy and stability of the estimation of covariance.

3. Unified Principal Component Analysis

In this section, we describe the UPCA based on GCM and some principles guiding the grouping of the random variables for UPCA.

3.1. PCA Based on CM

PCA is commonly used for feature extraction, which pursues an orthonormal transformation matrix $\mathbf{W}_{opt}$ that maximizes the total scatter of the extracted feature vectors. The criterion to be maximized is as follows:

$$J(\mathbf{W})=\mathrm{tr}\left(\mathbf{W}^T\mathbf{\Sigma}\mathbf{W}\right). \tag{9}$$

The optimal $\mathbf{W}$ can be chosen as follows:

$$\mathbf{W}_{opt}=\arg\max_{\mathbf{W}}\mathrm{tr}\left(\mathbf{W}^T\mathbf{\Sigma}\mathbf{W}\right)=\left[w_1 \quad w_2 \quad \ldots \quad w_m\right]. \tag{10}$$

where $\{w_i \mid i=1,2,\ldots,m\}$ are the eigenvectors of $\mathbf{\Sigma}$ corresponding to the m largest eigenvalues. The extracted low-dimensional feature vector can be obtained by

$$\mathrm{X}_f=\mathbf{W}_{opt}^T\mathrm{X}. \tag{11}$$

Thus, uncorrelated random variables are obtained, i.e.,

$$\mathbf{\Sigma}_{\mathrm{X}_f}=E\left[\mathrm{X}_f\mathrm{X}_f^T\right]=\mathrm{Diag}\left(\lambda_1 \quad \lambda_2 \quad \cdots \quad \lambda_m\right). \tag{12}$$

where $\{\lambda_i \mid i=1,2,\cdots,m\}$ is the set of m largest eigenvalues of Σ. When m is fixed, $\mathrm{tr}\left(\mathbf{W}_{opt}^T\mathbf{\Sigma}\mathbf{W}_{opt}\right)=\sum_{i=1}^{m}\lambda_i$ indicates the energy preserved by the m eigenvectors, which is invariant despite the different forms of CM (e.g. the permutation of the elements of X).

3.2. UPCA Based on GCM

As we can see from the above analysis, the GCM is in some sense equivalent to the traditional CM; therefore, $\mathbf{G}$ can be substituted for $\mathbf{\Sigma}$ in PCA, which results in a new kind of PCA. We call this method Unified PCA (UPCA).

Formally, suppose that $\mathrm{X}=\left[\mathrm{x}_1\ \mathrm{x}_2\ \ldots\ \mathrm{x}_m\right]^T$ is an m-dimensional random vector and reformed to a $k\times p$ random matrix $\mathbf{Y}=\left[\mathrm{Y}_1\ \mathrm{Y}_2\cdots\mathrm{Y}_k\right]^T$, where $\mathrm{Y}_i=\left[\mathrm{x}_{i1}\ \mathrm{x}_{i2}\cdots\mathrm{x}_{ip}\right]^T$ and $m=k\times p$. Then, one can create GCM as described in Section 2.3. With GCM, the orthonormal transformation matrix $\mathbf{W}_{opt}$ that maximizes the total scatter of the projected feature matrices can then be pursued by maximizing the following criterion:

$$J(\mathbf{W})=\mathrm{tr}\left(\mathbf{W}^T\mathbf{G}\mathbf{W}\right). \tag{13}$$

That is to say, we estimate the optimal $\mathbf{W}$ by the same way as in Equ.(10):

$$\mathbf{W}_{opt}=\arg\max_{\mathbf{W}}\mathrm{tr}\left(\mathbf{W}^T\mathbf{G}\mathbf{W}\right)=\left[w_1 \quad w_2 \quad \ldots \quad w_m\right]. \tag{14}$$

where $\{w_i \mid i=1,2,\ldots,m\}$ is the eigenvectors of $\mathbf{G}$ corresponding to its m largest eigenvalues. Then, the extracted feature matrix can be obtained by

$$\mathbf{Y}_f=\mathbf{W}_{opt}^T\mathbf{Y}. \tag{15}$$

The dimension of the extracted feature matrix $\mathbf{Y}_f$ is $m\times p$, which is p times of that of PCA based on CM, if both methods keep m eigenvectors in their $\mathbf{W}_{opt}$. And, similar to Equ.(12), the GCM after UPCA can written as:

$$\mathbf{G}_{\mathbf{Y}_f}=E\left[\mathbf{Y}_f\mathbf{Y}_f^T\right]=\mathrm{Diag}\left(\lambda_1 \quad \lambda_2 \quad \cdots \quad \lambda_m\right). \tag{16}$$

with $\{\lambda_i \mid i=1,2,\cdots,m\}$ the m largest eigenvalues of $\mathbf{G}$.

Compared with PCA, the maximum of the criterion in Equ.(13) varies with the different forms of $\mathbf{G}$. On account of this point, a more general criterion can be adopted:

$$J(\mathbf{W},\mathbf{G})=\mathrm{tr}\left(\mathbf{W}^T\mathbf{G}\mathbf{W}\right),\ \mathbf{G}\in\mathrm{S_G(X)}. \tag{17}$$

In this new criterion, $\mathbf{G}$ is taken as one of the parameters to be optimized, which implies the selection of grouping strategy by determining the number of variables in each group (i.e. p) and how to divide the variables into groups.

With GCM and the above criterion, we propose the Unified Principal Component Analysis (UPCA), which not only chooses the specific form of $\mathbf{G}$ but also optimizes the orthonormal transformation matrix $\mathbf{W}_{opt}$ that maximizes the total scatter.

Clearly, when $\mathbf{G}=\mathbf{\Sigma}$ (or $p=1$), UPCA degenerates to the traditional PCA. However, as mentioned above, in case of insufficient training samples, the estimation of $\mathbf{G}$ may be unstable when $p=1$. Given a fixed training set, as P increases, the estimation of $\mathbf{G}$ can be more and more stable. But, in this case, we need solve an optimization problem in the condition of $p>1$, which is too difficult to obtain a close-form solution. The difficulty lies in that, compared with PCA, $\mathrm{tr}\left(\mathbf{W}_{opt}^T\mathbf{G}\mathbf{W}_{opt}\right)=\sum_{i=1}^{m}\lambda_i$ in UPCA is not invariant to different forms of $\mathbf{G}$, let alone p. Fortunately, the key of this optimization problem is to choose a specific

grouping strategy (in another word, choose a specific form of $\mathbf{G}$ from $\mathrm{S_G}(\mathrm{X})$), which can be guided by some basic principles as discussed in the next subsection.

3.3. Principles of Variable Grouping

Before introducing the principles of variable grouping strategy, we firstly analyze the correlations between the features extracted by UPCA.

Definition 1: Given two *n*-dimensional random vectors, $\mathrm{X}=[\mathrm{x}_1\ \mathrm{x}_2 \ldots \mathrm{x}_n]^T$ and $\mathrm{Y}=[\mathrm{y}_1\ \mathrm{y}_2 \ldots \mathrm{y}_n]^T$, the random vectors X and Y are called *pseudo uncorrelated*, iff the following equation holds:

$$\mathrm{tr}(\mathbf{\Sigma}_{\mathrm{XY}})=\mathrm{tr}\left(E\left(\mathrm{XY}^T\right)\right)=0. \tag{18}$$

From Equ.(16), it is easy to conclude that the different row vectors of $\mathbf{Y}_f$ are pseudo uncorrelated, that is, UPCA leads to an $m \times p$ feature matrix $\mathbf{Y}_f$ which is comprised of pseudo uncorrelated random vectors. In contrast to traditional PCA based on CM, in which any two extracted random variables are uncorrelated, there still exist two kinds of correlation in the feature matrix $\mathbf{Y}_f$ if $p>1$:

1. The *inter-group* correlation: the random variables belonging to different rows of $\mathbf{Y}_f$ are correlated.
2. The *intra-group* correlation: the random variables belonging to the same row of $\mathbf{Y}_f$ are correlated.

Additionally, with the increase of p, these two kinds of correlation also increase. So, there is a tradeoff between the uncorrelation of the extracted random variables and the stability of the estimation of GCM, which actually reflects the tradeoff between the accuracy and the stability of the GCM estimation.

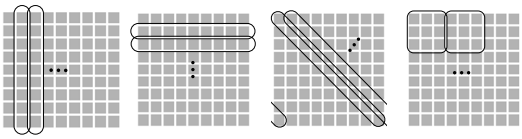
With the above analysis, we give the following three principles for choosing the grouping strategy.

Principle I: the random variables in the same group should be as uncorrelated as possible, because UPCA can not reduce the intra-group correlations.

Principle II: except for the corresponding (i.e., with the same position in different groups) random variables, the random variables in different groups should be as uncorrelated as possible, because UPCA cannot extract the really uncorrelated feature vectors but only the pseudo ones.

Principle III: the corresponding random variables in different groups should be as correlated as possible, which makes UPCA more effective for dimensionality reduction.

With the above three principles, one can deduce various algorithms from UPCA. In the next section, we show that the existing 2D PCA methods can be reformulated within this general framework. In addition, two novel 2D PCA methods are deduced for face recognition.



Column-based (2DPCA) Row-based (A2DPCA) Diagonal-based (DiaPCA) Block-based (MatPCA)

Fig.1. The grouping strategies of existing 2D PCA methods. The grey panes represent an enlarged 9×9 face image matrix, each pane corresponding to a pixel (a random variable) in the image. The panes in a round rectangle constitute a group.

4. Derivatives of UPCA for Face Recognition

In this section, the existing 2D PCA methods for face recognition are first revisited from the point of view of UPCA. Then, two new derivative methods are described.

4.1. Revisiting Existing 2D PCA Methods

Most 2D PCA methods mentioned in Section 1 directly use the original 2D image matrices to form the GCM, which actually means that the strategy is to group the random variables by row or column of the input image. The image covariance matrix of 2DPCA [2] [3] is defined:

$$\mathbf{G}_t = E\left[(\mathbf{A}-E\mathbf{A})^T(\mathbf{A}-E\mathbf{A})\right]. \tag{19}$$

where $\mathbf{A}$ is the image matrix.

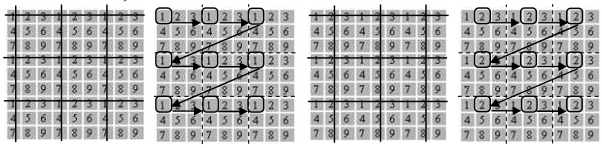
Comparing Equ.(7) with Equ.(19), it can be seen that $\mathbf{A}$ is a transpose of the 2D feature matrix $\mathbf{Y}$. Intuitively, from the viewpoint of grouping strategy, the existing 2D PCA methods are illustrated in Fig.1. Specifically, in 2DPCA, each *column* of the image is treated as one random variable group. Similarly, A2DPCA [4] treats a *row* as a random variable group and DiaPCA [6] treats a diagonal of the image matrix as a random variable group. MatPCA [5] also suggests using the image matrix directly, although re-matrixization (e.g. block partition) is allowed.

All the existing methods adopt the natural orders of the random variables by their positions in the input image. They can be easily analyzed with the help of the three principles described in Section 3.3 and the characteristics of face images. For normalized face images, neighboring pixels are highly correlated. As a result, all the existing methods suffer from the relatively high intra-group correlation, which somewhat violates the *Principle I*, especially for the block-based method, since the random variables in a block are closely located and thus more correlated than other existing methods.

On the other hand, the column-based, row-based and diagonal-based methods well conform to the *Principle III*, because the corresponding random variables in each group are from the same row or column of the image. However, for the block-based method, the correlations between the corresponding random variables in different groups are relatively weak, because these random variables spread over the whole face image.

4.2. GridPCA: Grid-sampling Methods

Based on the characteristics of normalized face images, a grouping strategy, named grid-sampling, is proposed, which can result in lower intra-group correlation and higher inter-group correlation. Specifically, in the method, a virtual rectangular grid is overlaid on the image, and the points at the intersections of gridline are sampled (see Fig.2(a_1)). The intensities corresponding to these sampled pixels are packed into one group (see Fig.2(a_2)). Then, the overlaid grid slides by one pixel in the horizontal or vertical direction (see Fig.2(b_1) and Fig.2(b_2)). At each new position, grid-sampling is performed and a new group of random variables is obtained. Finally, the number of different random variable groups is equal to the number of pixels included in each block and the number of variables in each group equals to the number of intersections. Hereinafter, we call the method GridPCA.



(a_1) (a_2) (b_1) (b_2)

Fig.2. The proposed grid-sampling grouping method. The number in each grey pane indicates the group index that this pane belongs to. (a_1): the image is overlaid by a 3×3 foursquare grid. (a_2): the random variables on the intersections are packed into Group 1 in the order as the arrows indicate. (b_1): the next position of the grid. (b_2): the pixels in Group 2.

In our method, on one hand, the corresponding random variables in different groups are from the same block, as indicated by the dashed lines in Fig.2(a_2) and (b_2). They are closely located in the image and thus highly correlated. On the other hand, the random variables in a group spread over the whole image, which implies lower intra-group correlation. So, the proposed grid-sampling strategy conforms pretty well to the *Principle I* and *Principle III*.

For the purpose of comparison, the grid-sampling strategy is also slightly modified in order to conflict with the Principle III on purpose. This is done by circularly rotating the random variables in each group one element. Thus, the corresponding random variables in different groups are no longer sampled nearby and thus with low correlations. This method is called RGridPCA. Due to its severe confliction with *Principle III*, its performance is expected to be not as good as the GridPCA method.

4.3. Intra-group Correlation Reduction Methods

For a specific grouping strategy, there is still room to improve. For example, we can further reduce the correlation between the random variables in each group by some statistical methods such as traditional PCA in order to satisfy the Principle I and reserve most information embedded in the original variables.

For k random variable groups, there are two ways to construct the PCA subspaces: one is to construct k PCA subspaces each corresponding to one random variable group; the other is to construct one PCA subspace taking all the random variable groups into account. These two methods has been investigated in [10] and called Block Specific PCA (BSPCA) and Block Universal PCA (BUPCA) respectively. Subpattern-based PCA (SpPCA) [11] is also an application of BSPCA.

In the context of UPCA, both BSPCA and BUPCA can be applied to either the original feature matrix $\mathbf{Y} \in R^{k \times p}$ or the extracted feature matrix $\mathbf{Y}_f \in R^{m \times p}$, which is named as prefix or postfix intra-group correlation reduction methods respectively. For prefix intra-group correlation reduction, BUPCA is a better choice. Specifically, one BUPCA subspace is constructed from all the training samples of the k random variable groups. Then, all p-dimensional random variable vectors are projected into this universal lower dimensional BUPCA subspace. Therefore, the correlation between the corresponding random variables in different groups is likely conserved, while the correlation between the intra-group variables is mostly reduced. In contrast, in BSPCA, one group-specific subspace is constructed for each random variable group. So, the correlation between the corresponding random variables in the reduced vector is very low, which violates Principle III. Therefore, its performance is expected to be not as good as that of BUPCA.

In addition, we can also reduce the intra-group correlations that might exist in the final feature matrix $\mathbf{Y}_f$. As a result, most 2D PCA methods utilizing this technique can represent face images by much fewer coefficients.

5. Experiments

We empirically evaluate the proposed methods, and compared with existing methods including direct correlation (DC), PCA, 2DPCA, A2DPCA, DiaPCA, MatPCA and RGridPCA on the FERET [12] database, using nearest neighbor classifier with Euclidean distance. For face recognition, these methods are developed mainly for dimension reduction and image representation. The recognition rates of these methods vary with the dimension of the extracted feature matrix. In our experiments, we compare not only the recognition rates of each method at different dimensions but also the cumulated eigenvalues, which can indicate the energy preserved from the training set.

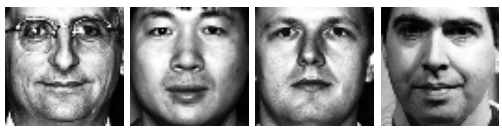
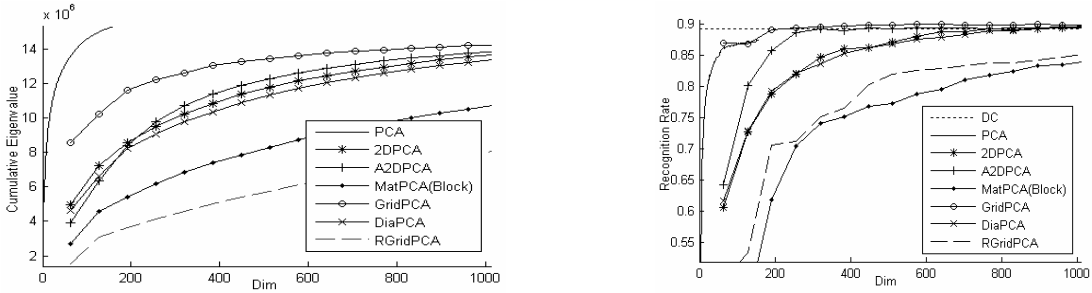
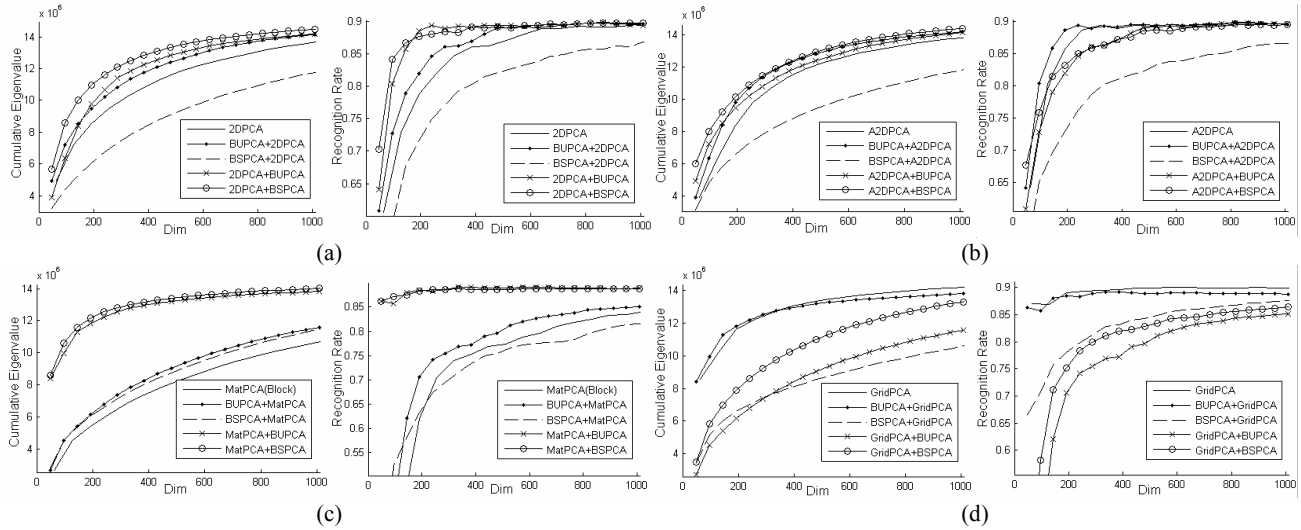


Fig.3. Example normalized face images in the FERET database.



(a) (b)

Fig.4. (a) comparison of the cumulative eigenvalues of 2D PCA methods; (b) comparison of the recognition rates of 2D PCA methods



(a) (b)

(c) (d)

Fig.5. The influences of the prefix and postfix intra-group correlation reduction methods on the cumulative eigenvalues and recognition rates of four 2D methods. (a) 2DPCA (b) A2DPCA (c) MatPCA(Block) (d) GridPCA.

5.1. Experiment Setup

All the face images are geometrically normalized to 64×64 and preprocessed by histogram equalization. Fig.3 shows some examples of the normalized face images. So, it is clear that $p = 64$ in 2DPCA, A2DPCA and DiaPCA. Correspondingly, we choose 8×8 blocks for MatPCA and 8×8 grid for our GridPCA. Thus, all the 2D PCA methods have $p = 64$ random variables in each group. Then, the prefix and postfix intra-group correlation reduction methods (BSPCA and BUPCA) are applied to 2DPCA, A2DPCA, MatPCA and GridPCA respectively and denoted as BSPCA+2DPCA, BUPCA+A2DPCA, etc. for prefix intra-group correlation reduction, and 2DPCA+BSPCA, A2DPCA+BUPCA, etc. for postfix ones. The dimensions of BSPCA and BUPCA subspaces are set to 48, keeping most energy in the variables group.

As described in Section 2.3, the 2D PCA methods are most appropriate for small sample size problem. Therefore, only 160 images, far less than the dimension ($64 \times 64 = 4096$) of the original features, are selected randomly from the standard FERET Gallery set to form the training set (TS). Then, 200 images are selected randomly from the remaining to form the gallery set (GS) and the 200 images of corresponding subjects from the FERET FB probe set form the probe set (PS). These images are all frontal faces and cover only expression variations. To reduce the randomness of the experimental results, the average results of 10 trials are reported.

5.2. Experimental Results

Fig.4(a) shows the cumulative eigenvalues curves for PCA, 2DPCA, A2DPCA, MatPCA, GridPCA, DiaPCA and RGridPCA against the preserved feature dimension. It is worth noting that, for the 2D PCA methods whose p value is greater than 1, the dimension of the extracted feature matrix is n (a nature number) times of p.

It is clear that, with the same dimension, PCA keeps far more energy than all 2D PCA methods, which coincides with the analysis in Section 2.3 and 3.1 that CM is the finest one among all forms of GCM and the random variables in the low-dimension feature vectors after PCA are uncorrelated. Our proposed GridPCA is also

remarkably better than the others, because its grouping strategy implies lower intra-group correlations and higher correlations between the corresponding random variables in different groups. The 2DPCA, A2DPCA and DiaPCA are similar since they are essentially row or column based. As is expected, MatPCA(Block) and RGridPCA are much worse than others, since the former violates Principle I and III and the latter conflicts with Principle III.

Recognition rates of the methods against the feature dimension are shown in Fig.4(b). One can see that, when the dimension is very low, PCA outperforms the others. But, when the dimension gets larger, the performances of the 2D PCA methods increase rapidly. Especially for GridPCA, the recognition rate is comparable with that of PCA even under lower dimension. Some 2D PCA methods outperform DC when the dimension is high enough, because they can remove noises by discarding the eigenvectors corresponding to small eigenvalues [13]. The peak recognition rate of PCA is lower than those of some 2D PCA methods, which can be explained by the stability of estimation as stated in Section 2.3.

Fig.5 demonstrates the influences of the intra-group correlation reduction (CR) methods on the 2D PCA methods. As can be seen from Fig.5, BUPCA is more appropriate for prefix intra-group CR than BSPCA; and the performances of 2DPCA, A2DPCA and MatPCA are improved by utilizing it. These results coincide with the analysis in Section 4.3. For MatPCA, the postfix intra-group CR methods promote the performance more evidently than the prefix ones, as shown in Fig.5(c). For GridPCA, it has weak intra-group correlations and strong correlations between corresponding random variables. In this situation, the intra-group CR methods may cause significant information loss in each group and affect the correlations between the corresponding random variables. So, the intra-group CR methods do not suit the GridPCA.

6. Conclusions and Future Work

This paper has made an attempt to find an underlying fundamental theory unifying the numerous recently emerging 2D PCA methods. Firstly, the generalized covariance matrix (GCM) is proposed by generalizing the traditional and partitioned covariance matrix. By taking the different forms of GCM into account, UPCA is proposed and the principles for variables grouping are discussed. Then, the existing 2D PCA methods are revisited from the viewpoint of UPCA. Finally, the grid-sampling method (GridPCA) and the intra-group correlation reduction method are deduced. Experimental results on the FERET face database not only support our theoretical analysis of GCM and UPCA, but also validate the feasibility of the proposed methods.

Though the principles of UPCA are founded, they only provide some heuristic guidelines to design and tune the 2D PCA methods. Future researches may focus on some general methods to find the optimal grouping strategy.

Acknowledgement

This paper is partially supported by National Natural Science Foundation of China under contract No.60332010, 60772071, and 60673091; Hi-Tech R&D Program of China under contract No.2006AA01Z122 and No.2007AA01Z163; 100 Talents Program of CAS; and ISVISION Technology Co. Ltd.

References

[1] M. A. Turk and A. P. Pentland, "Face recognition using eigenfaces," *IEEE Int. Conf. of Computer Vision and Pattern Recognition*, 1991.

[2] J. Yang, D. Zhang, A. F. Frangi, and J.-y. Yang, "Two-dimensional PCA: a new approach to appearance-based face representation and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, pp. 131-137, 2004.

[3] J. Yang and J.Y. Yang, "From image vector to matrix: a straightforward image projection technique-IMPCA vs. PCA," *Pattern Recognition*, vol. 35, pp. 1997-1999, 2002.

[4] D. Zhang and Z.-H. Zhou, "(2D)2PCA: Two-directional two-dimensional PCA for efficient face representation and recognition," *Neurocomputing*, vol.69, pp.224-231, 2005.

[5] S. Chen, Y. Zhu, D. Zhang, and J.-Y. Yang, "Feature extraction approaches based on matrix pattern: MatPCA and MatFLDA," *Pattern Recognition Letter*, vol. 26, pp. 1157-1167, 2005.

[6] D. Zhang, Z.-H. Zhou, and S. Chen, "Diagonal principal component analysis for face recognition," *Pattern Recognition*, vol. 39, pp. 140-142, 2006.

[7] H. Kong, L. Wang, E. K. Teoh, X. Li, J.-G. Wang, and R. Venkateswarlu, "Generalized 2D principal component analysis for face image representation and recognition," *Neural Networks*, vol. 18, pp. 585-594, 2005.

[8] W. Zuo, D. Zhang, and K. Wang, "Bidirectional PCA with assembled matrix distance metric for image recognition," *IEEE Transactions on Systems, Man and Cybernetics, Part B*, vol. 36, pp. 863-872, 2006.

[9] L. Wang, X. Wang, X. Zhang, and J. Feng, "The equivalence of two-dimensional PCA to line-based PCA," *Pattern Recognition Letter*, vol. 26, pp. 57-60, 2005.

[10] L. Wang, X. Wang, M. Chang, and J. Feng, "Is two-dimensional PCA a New Technique?," *ACTA AUTOMATICA SINICA*, vol. 31, pp. 782-787, 2005.

[11] S. Chen and Y. Zhu, "Subpattern-based principle component analysis," *Pattern Recognition*, vol. 37, pp. 1081-1083, 2004.

[12] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, "The FERET database and evaluation procedure for face-recognition algorithms," *Image and Vision Computing*, vol. 16, pp. 295-306, 1998.

[13] X. Wang and X. Tang, "A unified framework for subspace face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, pp. 1222-1228, 2004.