Improvement on PCA and 2DPCA Algorithms for Face Recognition

Vo Dinh Minh Nhat and Sungyoung Lee

Kyung Hee University – South of Korea
{vdmnhat, sylee}@oslab.khu.ac.kr

Abstract. Principle Component Analysis (PCA) technique is an important and well-developed area of image recognition and to date many linear discrimination methods have been put forward. Despite these efforts, there persist in the traditional PCA some weaknesses. In this paper, we propose new PCA-based methods that can improve the performance of the traditional PCA and two-dimensional PCA (2DPCA) approaches. In face recognition where the training data are labeled, a projection is often required to emphasize the discrimination between the clusters. Both PCA and 2DPCA may fail to accomplish this, no matter how easy the task is, as they are unsupervised techniques. The directions that maximize the scatter of the data might not be as adequate to discriminate between clusters. So we proposed new PCA-based schemes which can straightforwardly take into consideration data labeling, and makes the performance of recognition system better. Experiment results show our method achieves better performance in comparison with the traditional PCA and 2DPCA approaches with the complexity nearly as same as that of PCA and 2DPCA methods.

1. Introduction

Principal component analysis (PCA), also known as Karhunen-Loeve expansion, is a classical feature extraction and data representation technique widely used in the areas of pattern recognition and computer vision. Sirovich and Kirby [1], [2] first used PCA to efficiently represent pictures of human faces. Turk and Pentland [3] presented the well-known Eigenfaces method for face recognition in 1991. Since then, PCA has been widely investigated and has become one of the most successful approaches in face recognition [4], [5], [6], [7]. However, Wiskott et al. [10] pointed out that PCA could not capture even the simplest invariance unless this information is explicitly provided in the training data. They proposed a technique known as elastic bunch graph matching to overcome the weaknesses of PCA. Recently, two PCA-related methods, independent component analysis (ICA) and kernel principal component analysis (Kernel PCA) have been of wide concern. Bartlett et al. [11] and Draper et al. [12] proposed using ICA for face representation and found that it was better than PCA when cosines were used as the similarity measure (however, their performance was not significantly different if the Euclidean distance is used). Yang [14] used Kernel PCA for face feature extraction and recognition and showed that the Kernel

Eigenfaces method outperforms the classical Eigenfaces method. However, ICA and Kernel PCA are both computationally more expensive than PCA. The experimental results in [14] showed the ratio of the computation time required by ICA, Kernel PCA, and PCA is, on average, 8.7: 3.2: 1.0.

In all previous PCA-based face recognition technique, the 2D face image matrices must be previously transformed into 1D image vectors. The resulting image vectors of faces usually lead to a high dimensional image vector space, where it is difficult to evaluate the covariance matrix accurately due to its large size and the relatively small number of training samples. Fortunately, the eigenvectors can be calculated efficiently using the SVD techniques and the process of generating the covariance matrix is actually avoided. However, this does not imply that the eigenvectors can be evaluated accurately in this way since the eigenvectors are statistically determined by the covariance matrix, no matter what method is adopted for obtaining them. So recently in [16], a new PCA approach called 2DPCA, is developed for image feature extraction. As opposed to conventional PCA, 2DPCA is based on 2D matrices rather than 1D vectors. That is, the image matrix does not need to be transformed into vector. Instead, an image covariance matrix can be constructed directly using original image matrices. In contrast to the covariance matrix of PCA, the size of the image covariance matrix using 2DPCA is much smaller. As a result, 2DPCA has two important advantages over PCA. First, it is easier to evaluate the covariance matrix accurately. Second, less time is required to determine the corresponding eigenvectors.

However, in face recognition where the data are labeled, a projection is often required to emphasize the discrimination between the clusters. Both PCA and 2DPCA may fail to accomplish this, no matter how easy the task is, as they are unsupervised techniques. The directions that maximize the scatter of the data might not be as adequate to discriminate between clusters. In this paper, our proposed approaches can straightforwardly take into consideration data labeling, which makes the performance of recognition system better. The remainder of this paper is organized as follows: In Section 2, the PCA and 2DPCA methods are reviewed. The idea of the proposed methods and their algorithms are described in Section 3. In Section 4, experimental results are presented on the ORL face databases to demonstrate the effectiveness of our methods. Finally, conclusions are presented in Section 5.

2. PCA and 2D-PCA

In this section, we review the basic notions, essential mathematical background and algorithms of PCA and 2DPCA approaches that are needed for subsequent derivations in next sections.

Theorem 1. *Let A be an* nxn *symmetric matrix. Denoted by* $\lambda_1 \geq \ldots \geq \lambda_n$ *its sorted eigenvalues, and by* $w_1, \ldots, w_n$ *the corresponding eigenvectors. Then* $w_1, \ldots, w_m (m < n)$ *are the maximizer of the constrained maximization problem* $\max tr(W^T AW)$ *subject to* $W^T W = I$.

For the proof, we can reference [18].

Let us consider a set of *N* sample images $\{x_1, x_2, ..., x_N\}$ taking values in an *n*-dimensional image space, and the matrix $A = [\overline{x_1}\,\overline{x_2}...\overline{x_N}] \in \mathbb{R}^{nxN}$ with $\overline{x_i} = x_i - \mu$ and $\mu \in \mathbb{R}^n$ is the mean image of all samples. Let us also consider a linear transformation mapping the original *n*-dimensional image space into an *m*-dimensional feature space, where *m* < *n*. The new feature vectors $y_k \in \mathbb{R}^m$ are defined by the following linear transformation :

$$y_k = W^T \overline{x_k} \text{ and } Y = W^T A \qquad \textbf{(1)}$$

where $k = 1, 2, ..., N$ and $W \in \mathbb{R}^{nxm}$ is a matrix with orthonormal columns.

If the total scatter matrix is defined as

$$S_T = AA^T = \sum_{k=1}^{N} (x_k - \mu)(x_k - \mu)^T \qquad \textbf{(2)}$$

where *N* is the number of sample images, then after applying the linear transformation W^T, the scatter of the transformed feature vectors $\{y_1, y_2, ..., y_N\}$ is $W^T S_T W$. In PCA, the projection W_{opt} is chosen to maximize $tr(W^T S_T W)$. By Theorem 1, we have $W_{opt} = [w_1 w_2 ... w_m]$ with $\{w_i \,|\, i = 1, 2, ..., m\}$ is the set of *n*-dimensional eigenvectors of S_T corresponding to the *m* largest eigenvalues.

In 2DPCA approach, the image matrix does not need to be previously transformed into a vector, so a set of *N* sample images is represented as $\{X_1, X_2, ..., X_N\}$ with $X_i \in \mathbb{R}^{kxs}$. The total scatter matrix is re-defined as

$$G_T = \sum_{i=1}^{N} (X_i - \mu_X)(X_i - \mu_X)^T \qquad \textbf{(3)}$$

with $\mu_X = \frac{1}{N}\sum_{i=1}^{N} X_i \in \mathbb{R}^{kxs}$ is the mean image of all samples. $G_T \in \mathbb{R}^{kxk}$ is also called image covariance (scatter) matrix.

Similarly, a linear transformation mapping the original kxs image space into an mxs feature space, where $m < k$. The new feature matrices $Y_i \in \mathbb{R}^{mxs}$ are defined by the following linear transformation :

$$Y_i = W^T (X_i - \mu_X) \in \mathbb{R}^{mxs} \qquad \textbf{(4)}$$

where $i = 1, 2, ..., N$ and $W \in \mathbb{R}^{kxm}$ is a matrix with orthonormal columns. And $W_{opt} = [w_1 w_2 ... w_m]$ with $\{w_i | i = 1, 2, ..., m\}$ is the set of *n*-dimensional eigenvectors of G_T corresponding to the *m* largest eigenvalues.

After a transformation by 2DPCA, a feature matrix is obtained for each image. Then, a nearest neighbor classifier is used for classification. Here, the distance between two arbitrary feature matrices Y_i and Y_j is defined by using Euclidean distance as follows :

$$d(Y_i, Y_j) = \sqrt{\sum_{u=1}^{k} \sum_{v=1}^{s} (Y_i(u,v) - Y_j(u,v))^2} \tag{5}$$

Given a test sample Y_t, if $d(Y_t, Y_c) = \min_j d(Y_t, Y_j)$, then the resulting decision is Y_t belongs to the same class as Y_c.

3. Our proposed approaches

In the following part, we present our proposed methods. Firstly we will take a look at some necessary background. Let $A, B \in \mathbb{R}^{mxn}$, then A_{ci} and A_{rj} are i^{th} column vector and j^{th} row vector of matrix A. The Euclidean distance between A and B is defined as follows :

$$d(A,B)^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} (A_{ij} - B_{ij})^2 \tag{6}$$

The *Laplacian* is a key entity for describing pairwise relationships between data elements. This is a symmetric positive-semidefinite matrix, characterized by having zero row and column sums.

Lemma 1. *Let L be an* nxn *Laplacian, and let* $B \in \mathbb{R}^{mxn}$. Then we have the following equation : $tr(BLB^T) = \sum_{i<j} -L_{ij} d(B_{ci}, B_{cj})^2$.

Proof. Let $z = [z_1 z_2 ... z_n]^T \in \mathbb{R}^n$ then we have

$$\begin{aligned} z^T Lz &= \sum_i L_{ii} z_i^2 + 2\sum_{i<j} L_{ij} z_i z_j = \\ &= \sum_{i<j} -L_{ij}(z_i^2 + z_j^2) + 2\sum_{i<j} L_{ij} z_i z_j = \sum_{i<j} -L_{ij}(z_i - z_j)^2 \end{aligned} \tag{7}$$

By applying (5) we have

$$tr(BLB^T) = \sum_{k=1}^{m} B_{rk} L B_{rk}^T = \sum_{i<j} \sum_{k=1}^{m} -L_{ij} (B_{ki} - B_{kj})^2 = \sum_{i<j} -L_{ij} d(B_{ci}, B_{cj})^2 \quad \textbf{(8)}$$

Proof is done. Now we show that PCA finds the projection that maximizes the sum of all squared pair-wise distances between the projected data elements.

Theorem 2 . *PCA computes the m-dimensional project that maximizes*

$$\sum_{i<j} d(y_i, y_j)^2 \quad \textbf{(9)}$$

Proof. Before proving this Theorem, we define a NxN unit Laplacian, denoted by L^u , as $L^u = N\delta_{ij} - 1$, with δ_{ij} is the Kronecker delta (defined as 1 for $i = j$ and as 0 otherwise). We have

$$AL^u A^T = A(NI_N - U)A^T = NS_T - AUA^T = NS_T \quad \textbf{(10)}$$

with I_N is identity matrix and U is a matrix of all ones. The last equality is due to the fact that the coordinates are centered.

By Lemma 1, we get

$$\begin{aligned} tr(W^T S_T W) &= \frac{1}{N} tr(W^T AL^u A^T W) \\ &= \frac{1}{N} tr(YL^u Y^T) = \frac{1}{N} \sum_{i<j} d(y_i, y_j)^2 \end{aligned} \quad \textbf{(11)}$$

Maximizing $\frac{1}{N}\sum_{i<j} d(y_i, y_j)^2$ is maximizing $\sum_{i<j} d(y_i, y_j)^2$ .Proof is done.

Formulating PCA as in (9) implies a straightforward generalization—simply replace the unit Laplacian with a general one in the target function. In the notation of Theorem 2, this means that the m-dimensional projection will maximize a weighted sum of squared distances, instead of an unweighted sum. Hence, it would be natural to call such a projection method by the name weighted PCA (WPCA). Let us formalize this idea. Let be $\{wt_{ij}\}_{i,j=1}^{N}$ symmetric nonnegative pair-wise weights, with measuring how important it is for us to place the data elements i and j further apart in the low dimensional space. By convention, $wt_{ij} = 0$ for $i = j$. For this reason, these weights will be called dissimilarities in the context of weighted PCA. Normally, they are either supplied from an external source, or calculated from the data coordinates, in order to reflect any desired relationships between the data elements. Let define NxN Laplacian $L_{ij}^w = \begin{cases} \sum_{i \neq j} wt_{ij} & i = j \\ -wt_{ij} & i \neq j \end{cases}$ and $wt_{ij} = \begin{cases} 0 & x_i, x_j \in same \quad class \\ 1/d(x_i, x_j) & other \end{cases}$

Proposition 1 . *The m-dimensional project that maximizes*

$$\sum_{i<j} w_{ij} d(y_i, y_j)^2 \tag{12}$$

is obtained by taking the direction vectors to be the m highest eigenvectors of the matrix $AL^w A^T$ *.*

Proof. According to Lemma 1, we have

$$tr(W^T AL^w A^T W) = tr(YL^w Y^T) = \sum_{i<j} w_{ij} d(y_i, y_j)^2 \tag{13}$$

Now, we have weighted PCA and it seeks for the m-dimensional projection that maximizes $\sum_{i<j} wt_{ij} d(y_i, y_j)^2$. And this is obtained by taking the m highest eigenvectors of the matrix $AL^w A^T$. Now, we still have one thing need solving in this approach. It is how to get the eigenvectors of $AL^w A^T \in \mathbb{R}^{nxn}$, because this is a very big matrix. Let D be the N eigenvalues diagonal matrix of $A^T AL^w \in \mathbb{R}^{NxN}$ and V be the matrix whose columns are the corresponding eigenvectors, we have

$$A^T AL^w V = VD \Leftrightarrow AL^w A^T (AL^w V) = (AL^w V) D \tag{14}$$

From (14), we see that $AL^w V$ is the matrix whose columns are the first N eigenvectors of $AL^w A^T$ and D is the diagonal matrix of eigenvalues.

Until this time, we now can apply the idea of “weighted PCA” into 2DPCA approach. Let define A_i as follows :

$$A_i = [((X_1)_{ci} - (\mu_X)_{ci}) \ldots \quad ((X_N)_{ci} - (\mu_X)_{ci})] \in \mathbb{R}^{kxN} \tag{15}$$

and B_i be a matrix which is formed by all the column i^{th} of each matrix Y_i

$$B_i = [(Y_1)_{ci} \ldots \quad (Y_N)_{ci}] \in \mathbb{R}^{mxN} \tag{16}$$

The image scatter matrix G_T could be re-written as follow :

$$\begin{aligned} G_T &= \sum_{i=1}^{N} (X_i - \mu_X)(X_i - \mu_X)^T \\ &= \sum_{i=1}^{N} \sum_{j=1}^{s} (X_i^{(j)} - \mu_X^{(j)})(X_i^{(j)} - \mu_X^{(j)})^T \\ &= \sum_{i=1}^{s} A_i A_i^T \end{aligned} \tag{17}$$

Similarly, we show that 2DPCA also finds the projection that maximizes the sum of all squared pair-wise distances between the projected data .

Theorem 3 . *2DPCA computes the m-dimensional project that maximizes*

$$\sum_{i<j} d(Y_i, Y_j)^2 \qquad \textbf{(18)}$$

Proof. By Lemma 1, we get

$$\begin{aligned} tr(W^T G_T W) &= \frac{1}{N} tr(\sum_{i=1}^{s} W^T A_i L^u A_i^T W) \\ &= \frac{1}{N} tr(\sum_{i=1}^{s} B_i L^u B_i^T) = \frac{1}{N} \sum_{l=1}^{s} \sum_{i<j} d((Y_i)_{cl}, (Y_j)_{cl})^2 \qquad \textbf{(19)} \\ &= \frac{1}{N} \sum_{i<j} d(Y_i, Y_j)^2 \end{aligned}$$

Proof is done.

Proposition 2 . *The m-dimensional project that maximizes*

$$\sum_{i<j} w_{ij} d(Y_i, Y_j)^2 \qquad \textbf{(20)}$$

is obtained by taking the direction vectors to be the m highest eigenvectors of the matrix $\sum_{i=1}^{s} A_i L^w A_i^T$.

Proof. By Lemma 1, we get

$$\begin{aligned} tr(W^T (\sum_{i=1}^{s} A_i L^w A_i^T) W) &= tr(\sum_{i=1}^{s} W^T A_i L^w A_i^T W) \\ &= \frac{1}{N} tr(\sum_{i=1}^{s} B_i L^w B_i^T) = \frac{1}{N} \sum_{l=1}^{s} \sum_{i<j} w_{ij} d((Y_i)_{cl}, (Y_j)_{cl})^2 \qquad \textbf{(21)} \\ &= \sum_{i<j} w_{ij} d(Y_i, Y_j)^2 \end{aligned}$$

Proof is done.

Similar to the weight PCA approach we proposed above, the weighted 2DPCA (2DWPCA) seeks for the m-dimensional projection that maximizes $\sum_{i<j} wt_{ij} d(Y_i, Y_j)^2$. And this is obtained by taking the m highest eigenvectors of the matrix $\sum_{i=1}^{s} A_i L^w A_i^T$.

4. Experimental results

This section evaluates the performance of our propoped algorithms WPCA and 2DWPCA compared with that of the original PCA and 2DPCA algorithms based on using ORL face database. In the ORL database, there are ten different images of each of 40 distinct subjects. For some subjects, the images were taken at different times, varying the lighting, facial expressions (open / closed eyes, smiling / not smiling) and facial details (glasses / no glasses). All the images were taken against a dark homogeneous background with the subjects in an upright, frontal position (with tolerance for some side movement).

In our experiments, we tested the recognition rates with different number of training samples. $k(k = 2,3,4,5)$ images of each subject are randomly selected from the database for training and the remaining images of each subject for testing. For each value of k, 30 runs are performed with different random partition between training set and testing set. And for each k training sample experiment, we tested the recognition rates with different number of dimensions , d , which are from 2 to 10.

Table 1& *2* shows the average recognition rates (%) with ORL database. In *Table 1* two method PCA and WPCA are evaluated, while in *Table 2* the 2DPCA and 2DWPCA are performed also.

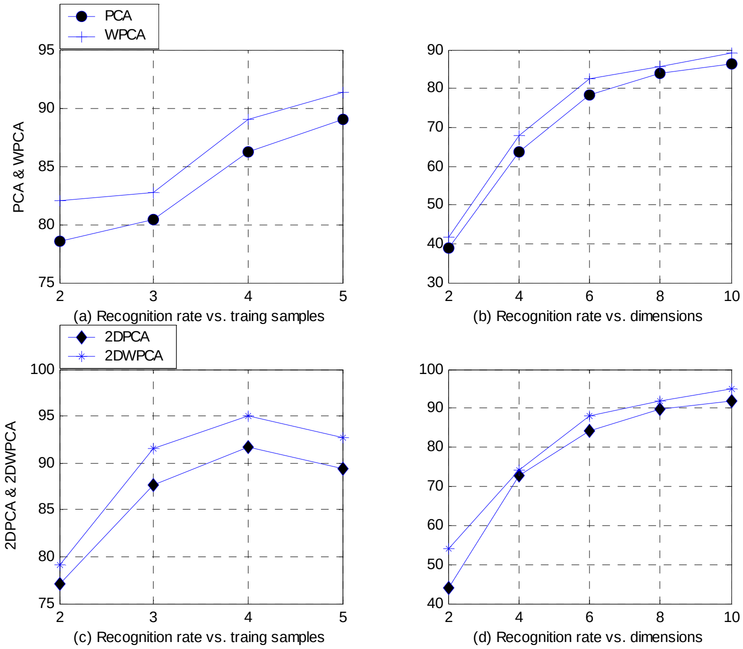
Table 1. The recognition rates with PCA and WPCA

D	2		4		6		8		10	
K	PCA	WPCA	PCA	WPCA	PCA	WPCA	PCA	WPCA	PCA	WPCA
2	39.69	**44.24**	61.56	**62.11**	69.69	**71.22**	78.13	**81.35**	78.49	**82.05**
3	40.36	**44.84**	66.79	**68.49**	70.00	**72.75**	78.21	**82.09**	80.36	**82.72**
4	38.75	**41.62**	63.75	**67.86**	78.33	**82.35**	83.75	**85.76**	86.25	**89.03**
5	37.00	**41.33**	68.00	**72.57**	79.50	**84.57**	85.50	**88.97**	89.00	**91.39**

Table 2. The recognition rates with 2DPCA and 2DWPCA

D	2		4		6		8		10	
K	2DPCA	2DWPCA	2DPCA	2DWPCA	2DPCA	2DWPCA	2DPCA	2DWPCA	2DPCA	2DWPCA
2	41.56	**43.95**	59.33	**63.37**	67.48	**70.18**	71.93	**74.44**	77.11	**79.14**
3	43.5	**46.17**	75.17	**78.89**	79.33	**81.62**	82.67	**85.47**	87.67	**91.49**
4	44.1	**54.2**	72.67	**74.11**	84.1	**88.13**	89.81	**91.72**	91.71	**95.06**
5	58.22	**60.3**	73.78	**76.01**	84.89	**85.55**	88.22	**89.92**	89.33	**92.77**

In below figure, we plot the graphs to make us see the recognition results of those methods intuitively. Two upper graphs are performed on PCA and WPCA methods, while the two lower ones are evaluated with 2DPCA and 2DWPCA methods. In recognition rate vs. training samples test, we choose the dimension d=10, and in recognition rate vs. dimension test, we choose the training sample k=4. We can see that our method achieves the better recognition rate compared to the traditional PCA and 2DPCA.



(a) Recognition rate vs. traing samples

(b) Recognition rate vs. dimensions

(c) Recognition rate vs. traing samples

(d) Recognition rate vs. dimensions

Fig. 1. The recognition rate (%) graphs which compare PCA & WPCA, 2DPCA & 2DWPCA

5. Conclusions

A new PCA-based methods for face recognition has been proposed in this paper. The proposed PCA-based methods can outperform the traditional PCA and 2DPCA methods. Both PCA and 2DPCA may fail to emphasize the discrimination between the clusters, no matter how easy the task is, as they are unsupervised techniques. The directions that maximize the scatter of the data might not be as adequate to discriminate between clusters. So we proposed new PCA-based schemes which can straightforwardly take into consideration data labeling, and makes the performance of recognition system better. The effectiveness of the proposed approaches can be seen through our experiments based on ORL face databases. Perhaps, this approach is not a novel technique in face recognition, however it can improve the performance of traditional PCA and 2DPCA approaches whose complexity is less than LDA or ICA approaches.

Reference

[1] L. Sirovich, M. Kirby: Low-Dimensional Procedure for Characterization of Human Faces. J. Optical Soc. Am., Vol. 4. (1987) 519-524.

[2] M. Kirby, L. Sirovich: Application of the KL Procedure for the Characterization of Human Faces. IEEE Trans. Pattern Analysis and Machine Intelligence. Vol. 12. (1990) 103-108.

[3] M. Turk, A. Pentland: Eigenfaces for Recognition. J. Cognitive Neuroscience. Vol. 3. (1991) 71-86.

[4] A. Pentland: Looking at People: Sensing for Ubiquitous and Wearable Computing. IEEE Trans. Pattern Analysis and Machine Intelligence. Vol. 22. (2000) 107-119.

[5] M.A. Grudin: On Internal Representations in Face Recognition Systems. Pattern Recognition. Vol. 33. (200) 1161-1177.

[6] G.W. Cottrell, M.K. Fleming: Face Recognition Using Unsupervised Feature Extraction. Proc. Int’l Neural Network Conf. (1990) 322-325.

[7] D. Valentin, H. Abdi, A.J. O’Toole, G.W. Cottrell: Connectionist Models of Face Processing: a Survey. Pattern Recognition. Vol. 27. (1994) 1209-1230.

[8] P.S. Penev, L. Sirovich: The Global Dimensionality of Face Space. Proc. Fourth IEEE Int’l Conf. Automatic Face and Gesture Recognition. (2000) 264- 270.

[9] L. Zhao, Y. Yang: Theoretical Analysis of Illumination in PCA-Based Vision Systems. Pattern Recognition. Vol. 32. (1999) 547-564.

[10] L. Wiskott, J.M. Fellous, N. Kru¨ ger, C. von der Malsburg: Face Recognition by Elastic Bunch Graph Matching. IEEE Trans. Pattern Analysis and Machine Intelligence. Vol. 19. (1997) 775-779.

[11] M.S. Bartlett, J.R. Movellan, T.J. Sejnowski: Face Recognition by Independent Component Analysis. IEEE Trans. Neural Networks. Vol. 13. (2002) 1450-1464.

[12] B.A. Draper, K. Baek, M.S. Bartlett, J.R. Beveridge: Recognizing Faces with PCA and ICA. Computer Vision and Image Understanding: special issue on face recognition, in press.

[13] P.C. Yuen, J.H. Lai: Face Representation Using Independent Component Analysis. Pattern Recognition. Vol. 35. (2002) 1247-1257.

[14] M.H. Yang: Kernel Eigenfaces vs. Kernel Fisherfaces: Face Recognition Using Kernel Methods. Proc. Fifth IEEE Int’l Conf. Automatic Face and Gesture Recognition (RGR’02). (2002) 215-220.

[15] Koren Y., Carmel L.: Robust linear dimensionality reduction. IEEE Transactions on Visualization and Computer Graphics. Vol 10. (2004) 459 – 470.

[16] Jian Yang, Zhang D., Frangi A.F., Jing-yu Yang: Two-dimensional PCA: a new approach to appearance-based face representation and recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol 26. (2004) 131 – 137.

[17] “The ORL face database” http://www.uk.research.att.com/facedatabase.html

[18] K. Fukunaga: Introduction to statistical pattern recognition. Academic Press, new York, 2 edition, 1990.