2010 International Conference on System Science, Engineering Design and Manufacturing Informatization

Feature selection using principal component analysis

Fengxi Song, Zhongwei Guo, Dayong Mei
Department of Automation and Simulation
New Star Research Inst. of Applied Tech. in Hefei City
Hefei, China
songfengxi@yahoo.com, Guo_zw@126.com, meidy0924@126.com

Abstract—**Principal component analysis (PCA) has been widely applied in the area of computer science. It is well-known that PCA is a popular transform method and the transform result is not directly related to a sole feature component of the original sample. However, in this paper, we try to apply principal components analysis (PCA) to feature selection. The proposed method well addresses the feature selection issue, from a viewpoint of numerical analysis. The analysis clearly shows that PCA has the potential to perform feature selection and is able to select a number of important individuals from all the feature components. Our method assumes that different feature components of original samples have different effects on feature extraction result and exploits the eigenvectors of the covariance matrix of PCA to evaluate the significance of each feature component of the original sample. When evaluating the significance of the feature components, the proposed method takes a number of eigenvectors into account. Then it uses a reasonable scheme to perform feature selection. The devised algorithm is not only subject to the nature of PCA but also computationally efficient. The experimental results on face recognition show that when the proposed method is able to greatly reduce the dimensionality of the original samples, it also does not bring the decrease in the recognition accuracy.**

Keywords-feature selection; principal component analysis; face recognition

1. INTRODUCTION

Feature selection has been widely applied in a number of branches of compute science including computer vision, pattern recognition and machine learning. It is known that feature selection is an important way to reduce the dimensionality of high-dimensional patterns. Feature selection has the following merits: first, it usually allows the whole method to be implemented computationally more efficient. Second, it usually achieves the increase in the accuracy or right rate of the method.

In the past years many concepts and approaches such as the mutual information [1], feature similarity [2], loss-margin of nearest neighbor classification [3], ant colony optimization [4] and genetic algorithms [5] have been proposed for feature selection. Feature selection also has been exploited to address a series of real-world problems such as clustering [6], cancer detection [7] and face recognition [8]. People also studied how to optimally integrate feature selection with a specific classifier such as SVM [9].

In this paper, we propose, for the first time, to exploit principal components analysis (PCA) for feature selection. PCA has been widely used in a variety of fields such as image processing, pattern recognition, data compression, data mining, machine learning and computer vision. The idea and method proposed in this paper is not only absolutely novel but also simple and intuitive. As we know, PCA is a powerful data representation method, being able to capture the most variable data components of samples [10]. The principal components analysis methodology has been applied to face recognition [10-13], image denoising [14] and machine learning [15,16,17], etc. PCA was also used as a baseline method when researchers tested some new methods. On the other hand, almost all these studies on PCA are focused on its applications in the field of feature extractions. In other words, the majority of the applications of PCA is to use PCA to transform samples into a new space and to use lower-dimensional representation from the new space to denote the sample. As the feature extraction result is an integrated reflection of the original feature components of samples, superficially it seems that PCA is not able to perform feature selection and no one has ever made this attempt. However, in this paper, we show, for the first time, applying PCA to feature selection is still feasible and we can use PCA to select a number of feature components from all the features components of original samples. We achieve this by viewing the PCA transform from a viewpoint of numerical analysis. The method proposed in this paper is significant both in theory and application.

The rest of the paper is organized as follows: Section 2 describes our method. Section 3 provides the analysis of our method. Section 4 presents the experimental results. Finally section 5 offers our conclusions.

II. DESCRIPTION OF THE PROPOSED METHOD

This section presents the proposed method that uses PCA to perform feature selection. As we know, both feature extraction and feature selection usually have a common goal of dimension reduction. We start

978-0-7695-4223-2/10 $26.00 © 2010 IEEE
DOI 10.1109/ICSEM.2010.14

from PCA-based feature extraction. Supposing that x is an eigenvector of the covariance matrix of PCA, we know that the feature extraction result, with respect to x, of an arbitrary sample vector a is

$$z = a^T x = \sum_{i=1}^{N} a_i x_i \qquad (1)$$

where $x = [x_1 ... x_N]^T$, $a = [a_1 ... a_N]^T$, and N is the dimensionality of samples vectors.

It is clear that the absolute value of x_i ($i = 1,2,..., N$) is able to statistically evaluate the contribution, to the feature extraction result, of the *i*th feature component of samples. It is easy to know the smaller the absolute value of x_i, the less the contribution of the *i*th feature component of samples. Indeed, if the absolute value of x_k is small enough, removing $a_k x_k$ from $\sum_{i=1}^{N} a_i x_i$ will almost not take effect on the feature extraction result. When a feature component is little important for feature extraction, we can also speculate that in the original space, this feature component is also not important. As a result, if the absolute value of x_k is small enough, we can consider that the k th feature component of samples is not important and can be deleted. We also note that there are always multiple eigenvectors, so we propose to take more than one eigenvector into account when evaluating the significance of one feature component. As a result, we devise the following algorithm to perform feature selection:

Step 1. We calculate the covariance matrix of PCA using the original training samples. We then solve all the eigenvectors and eigenvalues.

Step 2. We select the eigenvectors corresponding to the first m largest eigenvalues and denote these eigenvectors by $V_1, ..., V_m$, respectively.

Step 3. We calculate the contribution, to the feature extraction result, of the *j*th feature component as follows:

$$c_j = \sum_{p=1}^{m} | V_{pj} |, \qquad (2)$$

where V_{pj} denotes the *j*th entry of V_p, $j = 1,2,..., N$, $p = 1,2,..., m$. $| V_{pj} |$ stands for the absolute value of V_{pj}.

Step 4. We sort c_j in the descending order and use d_j to store the order, where $j = 1,2,..., N$. For example, if c_s and c_t are respectively the first and second largest among all the c_j, $j = 1,2,..., N$, then we should let $d_1 = s$ and $d_2 = t$, which means that the *s*th and *t*th feature components of the original samples are the two most important features. If *n*-dimensional features are required, then the feature selection result will be the d_1 th, ,d_2 th, …, d_n th feature components.

It is clear that after our method carries out the above feature selection procedure, it can use a classifier to classify the test samples. If our method is implemented in this way, we refer to it as scheme 1 of our method. On the other hand, after our method carries out the above feature selection procedure, it can also use PCA to transform the selected features into a lower-dimensional space and then classify test samples in the new space. We refer to this implementation of our method as scheme 2 of our method.

III. Insight into the proposed method

The proposed method indeed deals with the feature selection issue from a viewpoint of numerical analysis. It exploits the eigenvector to evaluate the contribution, to the feature extraction result, of each feature component. As shown in Eq.(1), the feature extraction result of an arbitrary sample is a linear combination of all the feature components of this sample and the entries of the eigenvector are the coefficients. That is, in the linear combination, the coefficient of the i th feature component of the sample is indeed the i th entry of the eigenvector. As a result, if the i th entry of the eigenvector has a very small absolute value, the i th feature component of all the samples will statistically have a little effect on the feature extraction result. Usually, more than one feature component that have a small absolute value (have a little effect). For example, if samples are images such as face images, samples will be very high-dimensional and experimental results show that a large number of feature components have a little effect. Thus, as our method discards all the feature components that have a little effect, it can greatly reduce the dimensionality of the original samples. It should be pointed out that our method performs feature selection at a very low computational cost, whereas conventional feature selection methods usually have a high computational complexity.

Since PCA-based feature extraction always exploits a number of eigenvectors, our method also uses m eigenvectors corresponding to the first m largest eigenvalues to evaluate the significance of different feature components of the original samples.

There are the following potential rationales: first, according to the essence of PCA, the eigenvector corresponding to a larger eigenvalue is able to capture more representative information of samples [10], so it is reasonable to use the eigenvectors corresponding to large eigenvalues rather than small eigenvalues. Second, if we use only one eigenvector to evaluate the significance of different feature components of the original samples, the evaluation result will not be representative. As shown early, while our method exploits multiple eigenvectors to perform feature extraction, it also uses multiple eigenvectors to evaluate the significance of different feature components. This allows it to obtain robust evaluation result.

Fig. 1 shows six original face images from the ORL database and the images corresponding to the feature selection results of our method. In this figure, the pixels that were selected as final features by the feature selection algorithm were set to the gray values of the pixels of the original image and the pixels that were not selected by the feature selection algorithm were set to zeroes. Fig. 2 shows five original face images from the AR database and the images corresponding to the feature selection results of our method. The first 13 images per class were used as training samples. The values of the pixels were set using the same scheme in Fig. 1. Fig. 3 shows five original face images from the Feret database and the images corresponding to the feature selection results of our method. The first four images per class were used as training samples.

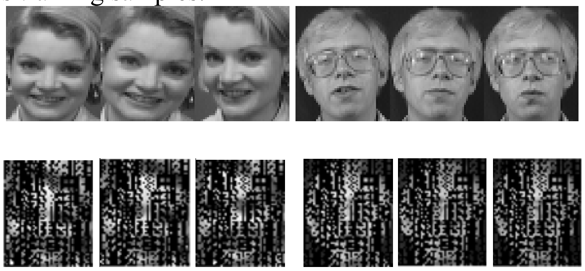

Figure 1. Six original face images from the ORL database and the images corresponding to the feature selection results of our method. The first row shows the six original face images. The second row shows the images corresponding to the feature selection result. The dimension of the selected features is 1000 whereas the original samples have the dimension of 2576.

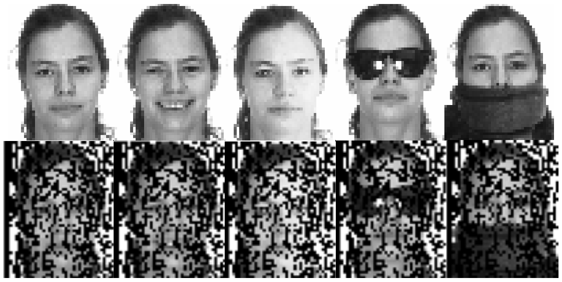

Figure 2. Five original face images from the AR database and the images corresponding to the feature selection result of our method. The first row shows the five original face images. The second row shows the images corresponding to the feature selection result. The dimension of the selected features is 1000 whereas the original samples have the dimension of 2000.

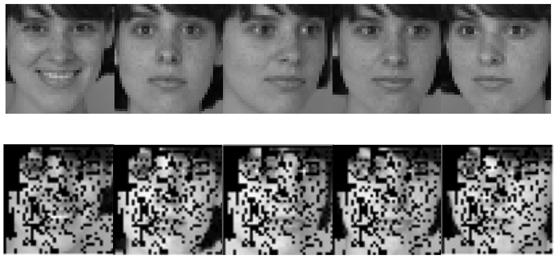

Figure 3. Five original face images from the Feret database and the images corresponding to the feature selection result of our method. The first row shows the five original face images. The second row shows the images corresponding to the feature selection result. The dimension of the selected features is 1000 whereas the original samples have the dimension of 1600.

IV. Experimental results

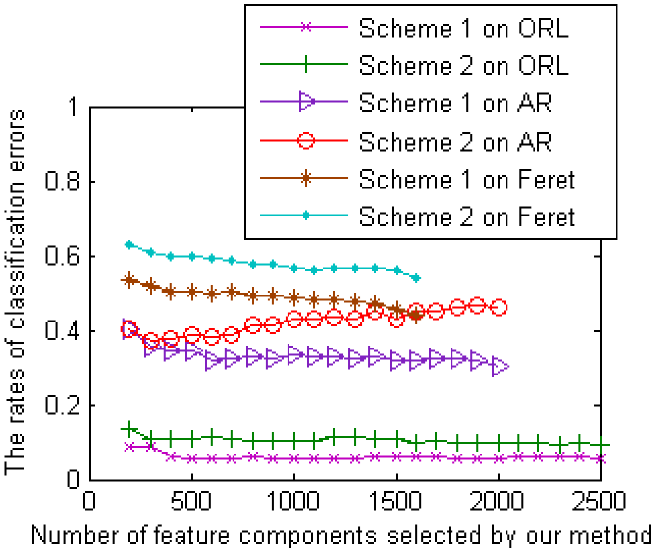

Figure 4. The rates of classification errors of scheme 1 and scheme 2 of our method on the ORL, AR and Feret face databases. In this figure, "scheme 1" and "scheme 2" denote scheme 1 of our method and scheme 2 of our method, respectively. Scheme 2 of our method first performed feature selection and then exploited PCA to transform the selected features into 200-dimensional ones.

We used the ORL [18], AR [19] and Feret [20] face databases to test our method. The first six samples per class of the ORL database were used as training samples and the remaining samples were used as test samples. As for the AR and Feret database, the first 16 and four samples per class were respectively used as training samples and the remaining samples were used as test samples. The nearest neighbor classifier was adopted for classifying the test samples. The images in all these face database were first resized into a half of the original size by using the downsampling method in [21]. Fig. 4 shows the rates of classification errors of scheme 1 of our method and scheme 2 of our method on the ORL, AR and Feret databases, respectively.

TABLE 1. THE RATES OF CLASSIFICATION ERRORS OF THE ORIGINAL SAMPLES OBTAINED USING THE NN CLASSIFIER

Databases	AR	Feret	ORL
Rates of classification errors	0.5047	0.4583	0.05

Table 1 shows the rates of classification errors on the original samples obtained using the nearest neighbor classifier. Fig. 4 and Table 1 show that the feature components selected by our method can obtain a lower classification error rate than the original samples. For example, while the classification error rate on the original samples of the AR database is 0.5047, scheme 1 of our method obtains a classification error rate of 0.3192 when it selects 600 feature components from all the feature components of the original samples of the AR database. we also see that scheme 1 of our method outperforms scheme 2 of our method.

V. CONCLUSION

PCA is a popular transform method and superficially the transform result is not directly related with a sole feature component of original samples; however, the proposed method clearly shows that PCA has the potential to perform feature selection and is able to select a number of important individuals from all the feature components. It achieves this by viewing the PCA transform as a numerical analysis problem and devises the algorithm from the viewpoint of numerical computation. The experimental results on face recognition show that when our method is able to greatly reduce the dimensionality of the original samples, it is also able to improve the classification accuracy.

ACKNOWLEDGMENT

This work is supported by the National Science Foundation of China under grant no. 60975006 and by Anhui Provincial Natural Science Foundation under grant no. 090412052.

REFERENCES

[1] H.W. Liu, J. G. Sun, L. Liu, H. J. Zhang. Feature selection with dynamic mutual information. *Pattern Recognition*, 42 (7):1330-1339, July 2009.

[2] P. Mitra, C. A. Murthy, S. K. Pal. Unsupervised Feature Selection Using Feature Similarity. *IEEE Trans. Pattern Anal. Mach. Intell*. 24(3): 301-312, 2002.

[3] Y. Li, B. L Lu. Feature selection based on loss-margin of nearest neighbor classification. *Pattern Recognition* 42(9): 1914-1921, 2009.

[4] M. H. Aghdam, N. G. Aghaee, M. E. Basiri, Text feature selection using ant colony optimization, *Expert Syst*. Appl. 36(3): 6843-6853, 2009.

[5] H. Yoshida, R. Leardi, K. Funatsu and K. Varmuza, Feature selection by genetic algorithms for mass spectral classifiers, *Anal. Chim.* Acta 446 (2001):485–494.

[6] H. Zeng, Y. M. Cheung. A new feature selection method for Gaussian mixture clustering. *Pattern Recognition* 42(2): 243-250, 2009.

[7] Y. Sun, C. F. Babbs, E.J. Delp. A comparison of feature selection methods for the detection of breast cancers in mammograms: adaptive sequential floating search vs. genetic algorithm, *Proceedings of the 27th Annual Conference of the IEEE Engineering in Medicine and Biology*, Shanghai, 2005.

[8] H. K. Ekenel, B. Sankur. Feature selection in the independent component subspace for face recognition. *Pattern Recognition* Letters 25(12): 1377-1388, 2004.

[9] M. H. Nguyen, F. d. Torre. Optimal feature selection for support vector machines, *Pattern Recognition*, 43(3): March 2010.

[10] Y. Xu, D. Zhang, J..Y. Yang. A feature extraction method for use with bimodal biometrics. *Pattern Recognition*, 43:1106–1115, 2010.

[11] P. N. Belhumeur, J. P. Hespanha, D. J. Kriegman, Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. IEEE Trans. *Pattern Anal. Machine Intell.* 19 (7): 711–720, 1997.

[12] M. Kirby, L. Sirovich. Application of the KL Procedure for the Characterization of Human Faces. *IEEE Trans. Pattern Anal. Machine Intell*., 12 (1):103-108 ,1990.

[13] Y. Xu, D. Zhang, J. Yang, J. Y. Yang, An approach for directly extracting features from matrix data and its application in face recognition, *Neurocomputing*, 71 (10-12):1857-1865, 2008.

[14]L. Zhang, R. Lukac, X. L. Wu, D. Zhang. PCA-Based Spatially Adaptive Denoising of CFA Images for Single-Sensor Digital Cameras. *IEEE Transactions on Image Processing* 18(4): 797-812, 2009.

[15] Y. Xu, D. Zhang , F. X. Song, J. Y. Yang, Z. Jing , M. Li, A method for speeding up feature extraction based on KPCA, *Neurocomputing*, 70(4-6): 1056-1061, 2007.

[16] Y. Xu, C. Lin, W. Zhao, Producing computationally efficient KPCA-based feature extraction for classification problems, Electronics Letters (SCI), 46(6), 452–453, 2010..

[17] Y. Xu, D. Zhang, Represent and fuse bimodal biometric images at the feature level: complex-matrix-based fusion scheme, Optical Engineering, 49(3), 2010, doi:10.1117/1.3359514.

[18]*http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html.*

[19] J. Yang, D. Zhang, X. Yong, J. Y. Yang, Two-dimensional Discriminant Transform for Face Recognition, *Pattern Recognition*, 38(7): 1125–1129, 2005.

[20] P. J. Phillips, H. Moon, S.A. Rizvi, P. J. Rauss. The FERET evaluation methodology for face-recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 22(10):1090–1104, 2000.

[21] Y. Xu, Z. Jin, Down-sampling face images and low-resolution face recognition. *The third international conference on innovative computing, information and control,* Dalian, China, 18-20: 392-395. June 2008.