Face Recognition Using Selected 2DPCA Coefficients

Alessandro L. Koerich, Luiz E. S. de Oliveira
Electrical Engineering Dept. / Computer Science Dept.
Federal University of Paraná (UFPR)
Curitiba, PR, Brazil
alessandro.koerich@ufpr.br, lesoliveira@inf.ufpr.br

Alceu S. Britto Jr.
Postgraduate Program in Informatics (PPGIa)
Pontifical Catholic University of Paraná (PUCPR)
Curitiba, PR, Brazil
alceu@ppgia.pucpr.br

Abstract—Face recognition based on principal component analysis (PCA) has provided successful results. This leads researchers to propose several variants of PCA such as the two-dimensional PCA (2DPCA). The results reported using this technique have demonstrated that it has an enormous potential as feature extractor for face recognition. However, the main drawback is the high number of coefficients produced. In this paper we propose to use a feature selection algorithm to analyze and to discard coefficients that are not relevant to the face recognition task. Experimental results on the ORL and the Yale databases have shown that the number of coefficients extracted by the 2DPCA can be reduced in about ten times while improving recognition rate.

Keywords-Feature selection, PCA, MOGA.

I. Introduction

Face recognition has been an active research field for the last years. Some potential applications for face recognition include bankcard identification, access control, mug shots searching, screening, security monitoring, and surveillance systems. Besides that, face recognition continues to attract researchers from image processing, pattern recognition, and computer vision [1]. Several attempts have been made to improve the reliability of face recognition systems. One very successful approach for face recognition is eigenfaces [2] which is based on PCA. Since then, several researchers have been investigating PCA and proposing successful approaches for face recognition based on such an approach [1].

Recently, other techniques based on PCA have been introduced in the literature. Bartlett et al. [3] use independent component analysis (ICA) for face recognition. They point out that ICA performs better than PCA when the cosine distance is used as similarity measure. Yang [4] used Kernel PCA for face recognition and demonstrated that Kernel PCA surpasses the classical PCA. However, the gain in performance has an associated increase in the computational cost. Yang [4] shows the ratio of the computational time required by ICA, Kernel PCA, and PCA is, on average, 8.7:3.2:1.0.

In PCA-based methods, the 2D face images must be previously transformed into 1D image vectors. This concatenation often leads to a high dimensional vector-space where it is computationally expensive to evaluate the covariance matrix accurately due to its large size and relatively small number of training samples. Yang et al [5] proposed a technique called two-dimensional PCA (2DPCA), which directly computes the eigenvectors from an image covariance matrix avoiding the matrix-to-vector conversion. Since the image covariance matrix has the size of the width of the images, the 2DPCA evaluates the image covariance matrix more accurately and computes the corresponding eigenvectors more efficiently than the PCA computes the covariance matrix of the vectors that result from converting the image matrices into vectors. Moreover, when dealing with recognition problems, the accuracy for 2DPCA is higher than PCA [5, 6] and the extraction of image features is computationally more efficient using 2DPCA than PCA. The latter makes it a strategy that should be considered when dealing with face recognition. One drawback of the 2DPCA lies in the fact that it needs more coefficients for image representation than PCA. Yang and Zhang [5] argue that this problem can be alleviated by using PCA after 2DPCA for further dimensional reduction. Zhang and Zhou [6[proposed a technique called (2D)2PCA to reduce the number of coefficients by simultaneously considering row and column directions of the original image. The fact is that 2DPCA has been proved to be a good feature extractor for face recognition, but it generates a great number of coefficients.

In this paper we investigate the hypothesis that not all coefficients produced by the 2DPCA are relevant to the face recognition task. We propose a methodology based on feature selection to find the most discriminant coefficients extracted by the 2DPCA. Comprehensive experiments on the ORL and the Yale databases have shown that the number of coefficients can be drastically reduced while improving slightly the recognition rate.

II. 2DPCA

First, consider an $m \times n$ random image matrix A. Let $X \in \Re^{n \times d}$ be a matrix with orthogonal columns, $n \geq d$. Projecting A onto X yields an $m \times d$ matrix $Y = AX$. In 2DPCA, the total scatter of the projected samples is used to determine a good projection matrix X. The following criterion is adopted

$$\begin{aligned} J(X) &= trace\left\{E\left[(Y-EY)(Y-EY)^T\right]\right\} \\ &= trace\left\{E\left[(AX-E(AX))(AX-E(AX))^T\right]\right\} \\ &= trace\left\{X^T E\left[(A-EA)^T(A-EA)\right]X\right\} \end{aligned} \quad (1)$$

where E denotes de expected value and the last term in (1) results from the fact that $trace(AB) = trace(BA)$, for any two matrices. Let us define the image covariance matrix $G_t = E[(A\text{-}EA)^T (A\text{-}EA)]$, which is an n×n nonnegative definite matrix from its definition. Suppose that there are M training images samples in total, the j-th training image is denoted by an m×n matrix $A_j(j = 1, 2, \ldots, M)$, and the average image of all training samples is denoted by

$$\overline{A} = \frac{1}{M}\sum_{j=1}^{M} A_j \quad (2)$$

Then, G_t can be evaluated by

$$G_t = \frac{1}{M}\sum_{j=1}^{M}\left(A_j - \overline{A}\right)^T\left(A_j - \overline{A}\right) \quad (2)$$

It has been demonstrated [7] that the optimal value for the projection matrix X_{opt} is composed of the orthogonal eigenvectors $X_1,\ldots,X_d$ of G_t corresponding to the d largest eigenvalues, i.e., $X_{opt} = [X_1,\ldots,X_d]$. Since the size of G is only n×n, computing its eigenvectors is quite fast.

A. Feature Extraction and Classification

The optimal projection vectors axes aforementioned are used for feature extraction. For a given image sample A, let $Y_k = AX_k$ ($k = 1,2,\ldots,d$). Then, we get a family of projected feature vectors, $Y_1,\ldots,Y_d$. Those vectors are used to form an m×d matrix $B = [Y_1,\ldots,Y_d]$, which is called the feature matrix of the image sample A. For example, suppose the image size is 100×100, then the number of coefficients is 100×d. It has been demonstrated that d should be set to no less than five to satisfy accuracy [6]. This leads us to a very large number of coefficients.

Once the features have been extracted, a nearest neighbor classifier is used for classification. The distance between two arbitrary feature matrices, $B_i = \left[Y_1^{(i)}, Y_2^{(i)}, \ldots, Y_d^{(i)}\right]$ and $B_j = \left[Y_1^{(j)}, Y_2^{(j)}, \ldots, Y_d^{(j)}\right]$ is defined by

$$d\left(B_i, B_j\right) = \sum_{k=1}^{d}\left\|Y_k^{(i)} - Y_k^{(j)}\right\|_2 \quad (3)$$

where $\left\|Y_k^{(i)} - Y_k^{(j)}\right\|_2$ denotes the Euclidean distance between the two principal component vectors $Y_k^{(i)}$ and $Y_k^{(j)}$. Consider for example a training set $B_1, B_2,\ldots,B_M$ composed of M images, and that each of these samples is assigned a given class ω_k. Given a test sample B, if $d(B, B_l) = min\ d(B, B_j)$ and $B_l \in \omega_k$, then the resulting decision is $B \in \omega_k$.

III. Feature Selection

An important issue in constructing classifiers is the selection of the best discriminative features. Many applications involve the representation of patterns by hundred of features. However, it has been observed that beyond a certain point, the inclusion of additional features leads to a worse performance both in terms of accuracy and computational complexity [8]. Moreover, the choice of features to represent the patterns affects several aspects of the pattern recognition problem such as accuracy, required learning time, and the necessary number of samples for training.

In the context of practical applications, feature selection presents a multi-criterion optimization function, such as, number of features and accuracy of classification. It has been demonstrated that multi-objective genetic algorithms offer a particularly attractive approach to solve this kind of problems. We have used the strategy proposed by Oliveira et al. [9] to perform feature selection. It is based on a powerful multiobjective genetic algorithm (MOGA) called Non-Dominated Sorting Algorithm (NSGA). Differently of a single genetic algorithm, NSGA produces a set of potential solutions known as Pareto-optimal solution. This allows the user to experiment different trade-offs between the objectives being optimized. The idea behind the NSGA is that a ranking selection method is used to emphasize good points and a niche method is used to maintain stable subpopulations of good points. It differs from simple genetic algorithm only in the way the selection operator works. Before the selection is performed, the population is ranked based on an individual's nondomination. The nondominated individuals present in the population are first identified from the current population. Then, all these individuals are assumed to constitute the first nondominated front in the population and assigned a large dummy fitness value. The same fitness value is assigned to give an equal reproductive potential to all these nondominated individuals. To maintain the diversity in the population, these classified individuals are then shared with their dummy fitness values. Sharing is achieved by performing selection operation using degraded fitness values obtained by dividing the original fitness value of an individual by a quantity proportional to the number of individuals around it. Thereafter, the population is reproduced according to the dummy fitness values. Since individuals in the first front have the maximum fitness value, they get more copies than the rest of the population. The efficiency of NSGA lies in the way multiple objectives are reduced to a dummy fitness function using nondominated sorting procedures. At the end, the algorithm produces a set of potential solutions that can be chosen by a decision maker. To support this choice, a good strategy lies in using an independent validation set to avoid an overfitted solution.

Fig. 1 shows a typical Pareto-optimal front for the feature selection problem. If we analyze only the Pareto-front, the best trade-off between the number of features and the error rate is the solution S_1. However, by analyzing the validation curve, we can observe that such a solution supplies a poor generalization on an unknown database. We can also observe that the accuracy/complexity trade-off that has the best generalization on the validation set is the solution S_2. Therefore, S_2 is the solution we should pick to avoid overfitting.

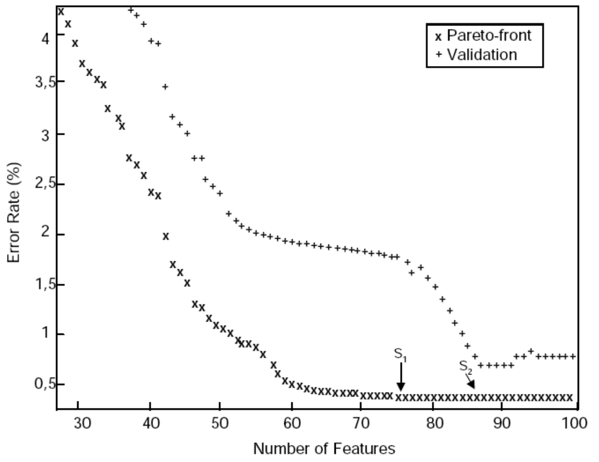
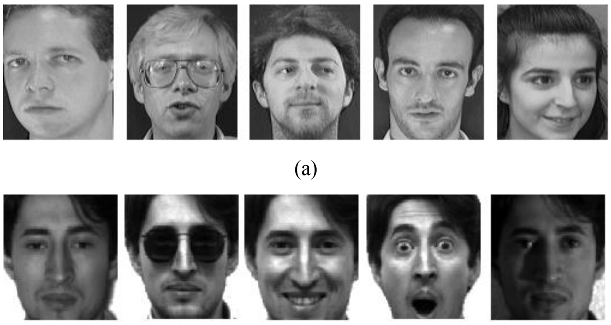


Figure 1. Example of a Pareto-optimal from produced by the MOGA

IV. Experiments and Analysis

Two databases were used to evaluate the performance of the proposed approach: the ORL database (Fig. 2a) is composed of 400 images (112×92 pixels), 10 different images from 40 individuals; the Yale database (Fig. 2b) contains 165 images of 15 individuals (11 images per individual) under various facial expressions and lighting conditions. The images were cropped and resized to 110×100 pixels in this experiment. The data was divided into four datasets, which are used for training, searching, validation of the Pareto, and testing, respectively.



(a)

(b)

Figure 2. Face samples from (a) ORL database, (b) YALE database

We have applied the feature selection on the feature matrix extracted by the 2DPCA method to demonstrate that 2DPCA provides very discriminant features but with a great number of unnecessary coefficients. The database was divided into 4 subsets: the training set, is composed of 160 images (four images of 40 individuals); the searching set, which is used to compute the fitness during the search, is composed of 40 images (one image of 40 individuals); the validation subset, is also composed of 40 images (one image of 40 individuals), and it is used to find the best solution in the Pareto front while avoiding the overfitted ones; the testing subset contains 160 images (four images of 40 individuals).

The first experiment provides a baseline to further evaluate the impact of feature selection. We have reproduced the results reported in [5]. In this way, we have used the first five individuals for training (200 images) and the other five for testing (200 images). We have tried different values for *d* and the minimum value we found without losing performance was *d* = 5, i.e., 112×5 coefficients. Using *d* = 5, we have achieved a recognition rate of 91.5%, which is similar to the results reported in [6] but with much more coefficients (*d* = 27). Since our testing set contains just four individual (160 images), we also computed the recognition rate for this testing set. In such a case, we achieved 91.0% of recognition rate, using five individuals for training (200 images).

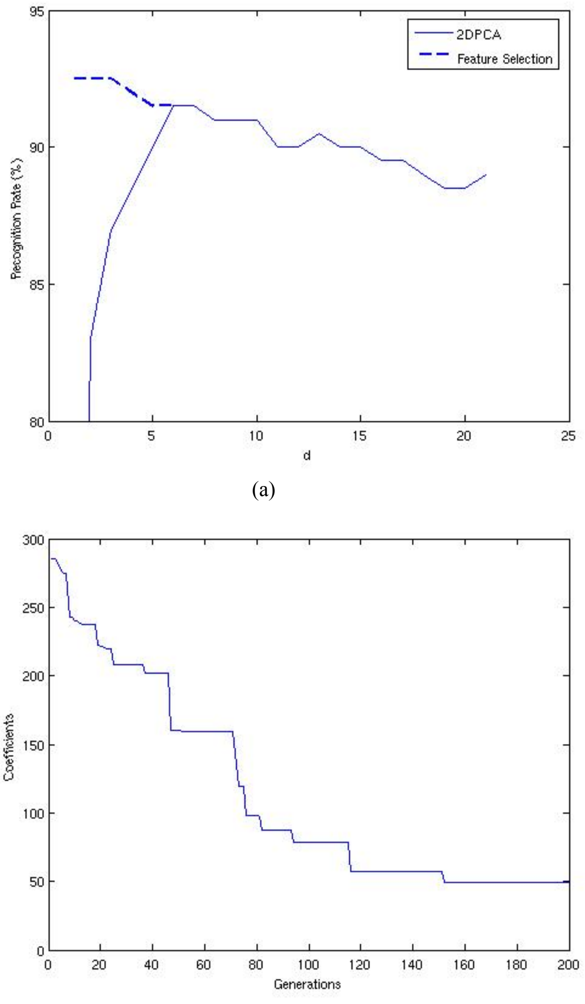
The NSGA is based on bit representation, one-point crossover, bit-flip mutation, and roulette wheel selection (with elitism). The following parameters were employed: population size = 100, number of generations = 200, probability of crossover = 0.8, probability of mutation = 0.002, and niche distance (σ_{share}) = 0.5. In these experiments the first objective was to minimize the number of coefficients and the second was to minimize the classification error. The latter was computed on the searching set through a nearest neighbor algorithm. The experiments were replicated ten times to better compare the results.

TABLE I. Comparison between the conventional 2DPCA and the 2DPCA after feature selection (FS)

Strategy	**Database**			
	ORL		**Yale**	
	# Coeff.	***Rec. Rate (%)***	***# Coeff.***	***Rec. Rate (%)***
2DPCA	560	91.0	550	84.0
2DPCA + FS	48	92.5	55	86.7

We have observed that the population converges to a specific region of the search space with few coefficients that provides a good performance. It also finds some solutions with very few coefficients but poor performance. This is due to the concept of nondominated sorting used during the search. Those solutions are discarded by using the validation set with all the solutions in the Pareto-front, achieving the best solution with only 48 coefficients. The recognition rate on the testing set for this case is 92.5%. In summary, after feature selection, the number of components was drastically reduced and the recognition rate slightly improved (Tab. 1). This corroborates to our claim that the 2DPCA works fine as feature extractor, but it generates a great number of unnecessary or even correlated coefficients.

Fig. 3a compares the number of coefficients and the performance of the 2DPCA and the 2DPCA with feature selection. As stated before, the best setup that we have found was for *d* = 5. The same behavior is observed for the performance of the 2DPCA with feature selection. Fig. 3a shows the performance for the feature selection (dashed line) just for *d* < 5. This is because the initial feature set used for feature selection was composed of 560 coefficients (*d* = 5). Fig. 3b corroborates to our findings in the sense that a great number of unnecessary coefficients are extracted by the 2DPCA. This figure reports the convergence speed of the MOGA (the average of 10 trials). It can be easily observed that after little iteration the algorithm removes a great number of coefficients. Another interesting observation is that the selected coefficients are uniformly distributed among the family of projected feature vectors. This shows that complementary information can be found in the entire feature space.



(a)

(b)

Figure 3. (a) Recognition rate versus d values for the conventional 2DPCA and the 2DPCA with feature selection, (b) Convergence speed of the MOGA (average of 10 trials).

In the experiments with the Yale database we have used the same protocol described before. The only difference is that the training set contains five individuals instead of four. The results follow the same trend we have observed in the previous experiments, i.e., a considerable reduction in the number of coefficients while improving slightly the recognition rate. It is difficult to compare the recognition rates achieved due to the small size of database. The difference reported in Tab. 1 means that more two images were recognized (65/75 instead of 63/75). However, the most important result that we have achieved is the reduction in the number of coefficients. After analyzing the misclassified images, we can observe that the features extracted by 2DPCA are not powerful enough to absorb very different lighting conditions, even after feature selection.

V. CONCLUSIONS

In this paper we propose the use of feature selection to find the most discriminative coefficients extracted by the 2DPCA technique. The approach used in this work takes into account a multi-objective genetic algorithm to perform feature selection. It generates a set of non-dominated solution, which is known as Pareto-optimal solutions. The best solution is then chosen based on a validation set, which helps avoiding an overfitted solution. We have demonstrate through experiments on two different databases (ORL and Yale) that the number of coefficients can be reduced considerably (about 10 times) while improving slightly the recognition rates. In summary, the 2DPCA performs well as feature extractor but it generates a great number of unnecessary or correlated features that can be easily removed by feature selection to optimize face recognition performance.

REFERENCES

[1] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, “Face recognition: A literature survey,” ACM Comp. Surveys, vol.35, no.4, pp.399-458, 2003.

[2] M. Turk and A. Pentland, “Eigenfaces for recognition,” J. Cognitive Neuroscience, vol.3, no.1, pp.71-86, 1991.

[3] M. S. Bartlett, J. R. Movellan, and T. J. Sejnowski, “Face recognition by independent component analysis,” IEEE Trans. Neural Networks, vol.13, no.6, pp.1450-1464, 2002.

[4] M. H. Yang, “Kernel eigenface vs. kernel fisherfaces: Face recognition using kernel methods,” in 5th IEEE Int. Conf. Aut. Face and Gesture Recognition, pp.21-220, 2002.

[5] J. Yang, D. Zang, A. F. Frangi, and J-Y. Yang, “Twodimensional PCA: A new approach to appearence-based face representation and recognition,” IEEE Trans. PAMI, vol.26, no.1, pp.131-137, 2004.

[6] D. Zhang and Z-H. Zhou, “2d2pca: 2-directional 2-dimensional PCA for efficient face representation and recognition,” Neuro-Computing, vol.69, pp.224-231, 2005.

[7] J. Yang and J. Y. Yang, “From image vector to matrix: A straightforward image prediction technique-IMPCA vs PCA,” Patt. Recognition, vol.35, no.9, pp.1997-1999, 2002.

[8] G. V. Trunk, “A problem of dimensionality: A simple example,” IEEE Trans. PAMI, vol.1, no.3, pp.306-307, 1979.

[9] L. S. Oliveira, R. Sabourin, F. Bortolozzi, and C. Y. Suen, “A methodology for feature selection using multi-objective genetic algorithms for handwritten digit string recognition,” Intl J. Patt. Recog. Art. Intelligence, vol.17, no.6, pp.903-930, 2003.