Comparing and Combining Spatial Dimension Reduction Methods in Face Verification

Licesio J. Rodríguez-Aragón, Cristina Conde, Ángel Serrano, and Enrique Cabello

Universidad Rey Juan Carlos, c\ Tulipán, s/n,
E-28933, Móstoles, Madrid, Spain
{licesio.rodriguez.aragon, cristina.conde, angel.serrano, enrique.cabello}@urjc.es
http://frav.escet.urjc.es

Abstract. The problem of high dimensionality in face verification tasks has recently been simplified by the use of underlying spatial structures as proposed in the 2DPCA, 2DLDA and CSA methods. Fusion techniques at both levels, feature extraction and matching score, have been developed to join the information obtained and achieve better results in verification process. The application of these advances to facial verification techniques using different SVM schemes as classification algorithm is here shown. The experiments have been performed over a wide facial database (FRAV2D including 109 subjects), in which only one interest variable was changed in each experiment. For training the SVMs, only two images per subject have been provided to fit in the small sample size problem.

1 Introduction

Many face recognition techniques have been developed over the past few decades. Some of them are based on dimension reduction methods such as Eigen Face method [1], using principal component analysis to obtain a most expressive subspace for face representation. Others are based on Fisher Face method [2], that uses linear discriminant analysis to obtain the most discriminant subspace. These methods as other improved variations treat input images as vectors [3]. Some recent works have begun to treat images as two dimensional matrices.

These new Spatial Dimension Reduction Methods give way to a set of vectors (2DPCA, 2DLDA) or even smaller matrices (CSA), instead of producing a single feature vector over a new lower dimensional space. To obtain the highest verification rate, different approaches have been proposed combining features either in a fusion at extraction level or at matching score level for each method, and then combining the best obtained results from the different methods [4]. SVM has been used as the classifier throughout all the experiments.

Another problem present in face recognition tasks is learning in the small sample size problem [5], what is of great practical interest. In face recognition it is not easy to collect a database with a large number of images of each individual.

E. Corchado et al. (Eds.): IDEAL 2006, LNCS 4224, pp. 645–653, 2006.

Furthermore, it is often difficult to collect training or testing data to cover all possible variations as illumination, pose, expression, occlusion, or even age. A set of different experiments with only two training images per subject to fit the small sample size problem has been carried out.

2 Feature Extraction

Traditional feature extraction techniques require that 2D face images are vectorized into a 1D row vector to then perform the dimension reduction [6]. Recently, Two-Dimensional PCA (2DPCA), Two-Dimensional LDA (2DLDA) and Coupled Subspace Analysis (CSA) have been developed for bidimensional data feature extraction. These methods are based on 2D matrices rather than 1D vectors, preserving spatial information. As base line method Principal Component Analysis [1] has been used.

2.1 Two-Dimensional Principal Component Analysis

Given a set of images $I_1, I_2, \ldots, I_N$ of height h and width w, the consideration of images $I_{h\times w}$ as 1D vectors instead as 2D structures is not the right approach to retain spatial information. Pixels are correlated to their neighbours and the transformation of images into vectors produces a loss of information preserving the dimensionality.

The idea recently presented as a variation of traditional PCA, is to project an image $I_{h\times w}$ onto X^{2DPCA} by the following transformation [7,8],

$$Y_{h\times 1} = I_{h\times w} \cdot X^{2DPCA}_{w\times 1}. \tag{1}$$

As result, a h dimensional projected vector Y, known as projected feature vector of image I, is obtained. The total covariance matrix S_X over the set of projected feature vectors of training images $I_1, I_2, \ldots, I_N$ is considered. The mean image $\overline{I}$ of the training set, is taken into account.

$$S_X = \frac{1}{N}\sum_{i=1}^{N}[(I_i - \overline{I})X][(I_i - \overline{I})X]^T \tag{2}$$

The maximization of the total scatter of projections is chosen as the criterion to select the vector X^{2DPCA}. The total scatter of the projected samples is characterized by the trace of the covariance matrix of the projected feature vectors. It has been considered the optimal projection axis X^{2DPCA} as the is the unitary vector that maximizes $tr(S_X)$, which corresponds to the eigenvector of largest associated eigenvalue of the image covariance matrix S, defined as a $w \times w$ nonnegative matrix that can be directly evaluated using the training samples,

$$S = \frac{1}{N}\sum_{i=1}^{N}[(I_i - \overline{I})^T(I_i - \overline{I})]. \tag{3}$$

2.2 Two-Dimensional Linear Discriminant Analysis

The idea presented as 2DPCA, has been upgraded to consider the class information [9,10]. Suppose there are L known pattern clases having M samples for each class, $N = L \cdot M$. The idea is to project each image as in (1), but to obtain X^{2DLDA} with the information provided by the classes. The covariance over the set of images can be decomposed into between-class and within-class. The mean image as in 2DPCA, as well as the mean image of the class $\overline{I^j}$, $j = 1, \dots, L$, are taken into account.

$$S_{XB} = \sum_{j=1}^{L} M[(\overline{I^j} - \overline{I})X][(\overline{I^j} - \overline{I})X]^T;\ S_{XW} = \sum_{j=1}^{L}\sum_{i=1}^{M}[(I_i^j - \overline{I^j})X][(I_i^j - \overline{I^j})X]^T \tag{4}$$

The objective function maximized in this case to select X^{2DLDA} is considered a class specific linear projection criterion, and can be expressed as a quotient of the traces: $tr(S_{XB})/tr(S_{XW})$. The total between and within covariances are defined as $w \times w$ nonnegative matrices and can be directly evaluated.

$$S_B = \sum_{j=1}^{L} M[(\overline{I^j} - \overline{I})][(\overline{I^j} - \overline{I})]^T; \quad S_W = \sum_{j=1}^{L}\sum_{i=1}^{M}[(I_i^j - \overline{I^j})][(I_i^j - \overline{I^j})]^T \tag{5}$$

Both matrices are formally identical to the corresponding traditional LDA, and by maximizing the traces quotient, the within-class scatter is minimized, whereas the between-class scatter is maximized, giving as result the maximization of discriminating information. The optimal projection axis X^{2DLDA} corresponds to the eigenvector of $S_B \cdot S_W^{-1}$, of largest associated eigenvalue.

2.3 Coupled Subspace Analysis

Recently a new approach has been presented to reconstruct the original image matrices with two low dimensional coupled subspaces, in the sense of least square error [11]. These two subspaces encode the row and column information of the image matrices.

Let us denote Y_i, of height h' and width w', as the lower dimensional matrix representation of sample I_i, $i = 1, \dots, N$, derived from two projection matrices $B_{h \times h'}$ and $C_{w \times w'}$,

$$Y_{h' \times w'} = B^T_{h' \times h} \cdot I_{h \times w} \cdot C_{w \times w'}. \tag{6}$$

The matrices B and C are chosen as those that best reconstruct the original images from the projections, in the sense of least square error satisfying the following optimal matrix reconstruction criterion,

$$(B^*, C^*) = \operatorname*{argmin}_{B,C} \sum_i \|B \cdot Y_i \cdot C^T - I_i\|_F^2. \tag{7}$$

Being $\| \cdot \|_F$ the Frobenius norm of a matrix.

The objective function has no closed form and to obtain a local optimal solution, an iterative procedure has been presented. The whole procedure is called Coupled Subspace Analysis or Generalized Low Rank Approximation [11,12]. As it has been shown, this procedure is connected with traditional PCA and 2DPCA. Principal Component Analysis is a special case of CSA algorithm with $w = 1$ and 2DPCA is a special case of CSA algorithm with fixed $B = Id$.

3 Projection and Reconstruction

In 2DPCA and 2DLDA, as in traditional PCA, a proportion of retained variance can be fixed, $\sum_1^d \lambda_i / \sum_1^w \lambda_i$, where $\lambda_1 > \lambda_2 > \cdots > \lambda_w$ are the eigenvalues and $X_1, X_2, \ldots, X_d$ are the eigenvectors corresponding to the d largest eigenvalues.

Once d is fixed, $X_1, X_2, \ldots, X_d$ are the ortonormal axes used to perform the feature extraction. Let $V = [Y_1, Y_2, \ldots, Y_d]$ and $U = [X_1, X_2, \ldots, X_d]$, then

$$V_{h\times d} = I_{h\times w} \cdot U_{w\times d}. \tag{8}$$

A set of projected vectors, $Y_1, Y_2, \ldots, Y_d$, are obtained for both methods. Each projection over an optimal projection vector is a vector, instead of a scalar as in traditional PCA. A feature matrix $V_{h\times d}$ for each considered dimension reduction method is produced, containing either the most amount of variance, or the most discriminating features of image I.

In CSA the projection is performed through the optimal projection matrices $B^*_{h\times h'}$ and $C^*_{w\times w'}$ as in (6). As result, the extracted features form a lower dimensional matrix of height h' and width w'. In these dimension reduction methods, a reconstruction of the images from the features is possible (Fig. 1). It is possible to obtain an approximation of the original image with the retained information determined by d for PCA, 2DPCA and 2DLDA, $\widetilde{I}_{h\times w} = V_{h\times d} \cdot U^T_{d\times w}$, or by (h', w') for CSA, $\widetilde{I}_{h\times w} = B_{h\times h'} \cdot Y_{h'\times w'} \cdot C^T_{w'\times w}$.

4 Facial Verification Using SVM

SVM is a method of learning and separating binary classes [13]; it is superior in classification performance and is a widely used technique in pattern recognition and especially in face verification tasks [3,6].

Given a set of features $y_1, y_2, \ldots, y_N$ where $y_i \in \mathbb{R}^n$, and each feature vector associated to a corresponding label $l_1, l_2, \ldots, l_N$ where $l_i \in \{-1, +1\}$, the aim of a SVM is to separate the class label of each feature vector by forming a hyperplane. The optimal separating hyperplane is determined by giving the largest margin of separation between different classes. This hyperplane is obtained through a minimization process subjected to certain constrains. Theoretical work has solved the existing difficulties of using SVM in practical application [14].

As SVM is a binary classifier, a *one vs. all* scheme is used. For each class, each subject, a binary classifier is generated with positive label associated to feature vectors that correspond to the class, and with negative label associated to all the other classes.

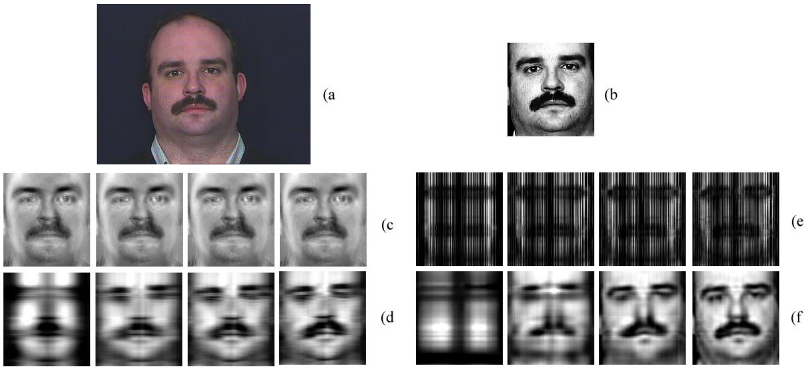

Fig. 1. a) One of the original frontal images in our database. b) Automatically selected window containing the facial expression of the subject in equalized gray scale. From left to right: c) reconstructed images, for $d = 30, 90, 150, 210$, from PCA projection. d) and e) reconstructed images, for $d = 2, 4, 6, 8$, from 2DPCA and 2DLDA projections respectively. f) reconstructed images, for $h' = w' = 3, 9, 15, 21$, from CSA projection.

In our experiments and in order to fit in the small sample size problem [5], the same two frontal and neutral images of every subject are selected as the training set for every experiment. A disjoint group of images, all of them affected by the same perturbation, is selected as the test set. The training set is used in the feature extraction process through PCA, 2DPCA, 2DLDA and CSA. Then, the training images are projected onto the new ortonormal axes and the feature vector (PCA), vectors (2DPCA,2DLDA), or low dimensional matrix (CSA) are obtained. The required SVMs are trained for each subject.

Several strategies have been used to train and combine the SVMs. When training and classifying PCA features, each image generates one feature vector $Y_{d\times 1}$ and one SVM is trained for each subject, with its feature vectors labelled as $+1$ and all the other feature vectors as -1.

For feature vectors obtained from 2DPCA and 2DLDA, each image generates a set of projected vectors, $V_{h\times d} = [Y_1, Y_2, \ldots, Y_d]$, which are considered under three different strategies. The first strategy, considered as a fusion at extraction level, generates a unique feature vector through a concatenation of d projected vectors, and then one SVM is trained for each subject as in PCA. The second and third approaches, fusion at matching score level, consider d projected vectors and, consequently, d SVMs are trained for each subject. These d outputs are then combined to produce a final classification output, first through an arithmetic mean, and secondly through a weighted mean.

On the other hand, applying CSA produces a low dimensional matrix $Y_{h'\times w'}$ for every image. This feature matrix is then transformed into a vector, $Y_{h'\cdot w'\times 1}$, and as in PCA one SVM is trained for each subject.

Once the SVMs are trained, each image from the test set is projected obtaining the corresponding features for each dimension reduction method (1,6).

The features of the test set are classified through the SVMs to measure the performance of the generated system.

For the SVM obtained from the PCA, from the concatenation strategy of 2DPCA, 2DLDA and the CSA feature vectors, the output is compared with the known label of every test image. However, for the ensemble of SVMs obtained from the 2DPCA and 2DLDA feature vectors, the d outputs are combined whether through an arithmetic or a weighted mean. Arithmetic approach combines the d outputs through an arithmetic mean. At weighted approach, every output is weighted with the amount of variance explained by its dimension, which means that each output will be taken into account proportionally to the value of the eigenvalue associated to the corresponding eigenvector: $\lambda_i / \sum_{j=1}^{d} \lambda_j$ is the weight for the $i-$SVM, $i = 1, 2, \ldots, d$.

To measure the system performance, a cross validation procedure is carried out. The Equal Error Rate, EER, that is the value for which false positive rate (FP) is equal to false negative rate (FN) is a valuable reference of the performance of the system.

4.1 Combining Results

To combine the results obtained by the three novel methods, 2DPCA, 2DLDA and CSA, a fusion at matching score level has been performed for the best approaches of each method. The scores obtained by the SVMs of the three methods have been linearly combined.

$$\text{Final Score} = \alpha \cdot \text{2DPCA} + \beta \cdot \text{2DLDA} + \gamma \cdot \text{CSA} \tag{9}$$

Being α, β and $\gamma \in [0, 1]$, and verifying $\alpha + \beta + \gamma = 1$. To reach this combination two sources have been considered: firstly the scores obtained have been directly combined and secondly, a previous standardization of the scores has been performed, before combining the individual scores. The aim of this process is not only to achieve better results by combining the different Spatial Dimension Reduction Methods, but to weight the contribution of each of them towards the best performance in the verification process.

5 Design of Experiment

Although a large number of very complete facial databases acquired under very good conditions are available for the scientific community, our research group has found valuable to collected FRAV2D, a quite complete set of facial images including 109 subjects. This particular database has been created under extremely detailed procedures with a deep control of each of the acquiring conditions. All the images have been taken under 8 different conditions varying pose, expression and illumination and only varying one interest variable for each of them. 32 images of each subject were taken, being 12 frontal, 4 preforming a 15° rotation, 4 performing a 30° rotation, 4 with zenithal instead of diffuse illumination, 4 with expression changes and 4 occluding parts of the face. This database as well as other facial databases are freely available upon request at the research group home page[1].

Each experiment has been performed for the 109 subjects from our database. In all the experiments, the train set for the extraction of features and for the classifiers training is formed by the same 2 frontal and neutral images of each subject, in order to fit in the small sample size problem. Then, the classifiers have been tested over 5 different groups of images. Firstly, the 10 remaining frontal and neutral images for each subject have been used to perform the cross validation process. In a second experiment, the 4 images obtained with zenithal illumination have formed the test set. The 4 15° turn images have been selected to measure the performance of the system to pose variations. In the fourth experiment 4 images with expressions changes have been used. And finally, 4 occluded images for each subject have formed the test set.

6 Results

Tests, varying the dimensions of the different feature spaces, have been carried out for the four dimension reduction methods. For PCA, experiments have been performed for values $d = 30, 60, 90, 120, 150, 180, 210$. For 2DLDA and 2DPCA, values $d = 1, 2, 3, \ldots, 10$ are used under the three different classification strategies (Concatenation, Arithmetic and Weighted). And for CSA, low dimensional square matrices of sizes $h' = w' = 3, 6, 9, 12, 15, 18$ have been considered.

The lowest EER values in percentage, corresponding to each experiment and to each dimension reduction method under the different strategies, are presented in Table 1, as well as the values of d or $(h' \times w')$ for which they were achieved.

The optimum strategies for each method have been combined as proposed (9), and the results for the standardized option are shown to be the best performance (Table 1).

7 Conclusions

Better results for spatial dimension reduction methods (2DPCA, 2DLDA and CSA) than for traditional PCA are evidently achieved as shown in the EER values presented in Table 1.

The fusion at the matching score level used (Arithmetic and Weighted) never improves the fusion at the feature extraction level (Concatenation strategy).

Better results are achieved in all the experiments by means of the combination of the three scores, even thought in some of them the improvements are not significant. The coefficients of the optimal linear combination weighs the contribution of each method to the best solution.

2DLDA is theoretically the best approach as it distinguishes between within-class and between-class scatter and the criterion minimizes the first, while the second is maximized. As can be observed in the results of the 5 experiments

[1] http://frav.escet.urjc.es

Table 1. First, best EER (in percentage), values obtained for each dimension reduction method in each experiment: 1) Frontal, 2) Illumination, 3) 15° Rotated, 4) Expression, 5) Occluded. In brackets: the dimension for which they are achieved. Second, EER for the best combination (9) of the standardized scores of every method and the coefficients α for 2DPCA, β for 2DLDA and γ for CSA in which it occur.

Exp.		2DPCA			2DLDA		
	PCA	Conc.	Arith.	Weigh.	Conc.	Artith.	Weigh.
1)	3.0 (210)	1.3 (9)	1.9 (6)	1.9 (10)	**1.2 (6)**	1.9 (6)	1.9 (9)
2)	6.5 (210)	2.9 (7)	3.6 (6)	3.7 (9)	**2.0 (9)**	3.0 (7)	3.2 (10)
3)	21.5 (120)	**13.9 (1)**	13.9 (1)	13.9 (1)	16.9 (2)	17.5 (2)	20.6 (2)
4)	12.7 (210)	9.4 (10)	11.0 (5)	11.5 (9)	**7.2 (6)**	9.0 (5)	8.5 (9)
5)	30.7 (120)	25.2 (10)	24.8 (10)	24.7 (10)	24.6 (9)	26.5 (6)	27.3 (6)

Exp.	CSA	Fusion:	EER	α	β	γ
1)	1.3 (12×12)		1.2	0	0.4	0.6
2)	2.7 (18×18)		2.0	0	1	0
3)	20.6 (24×24)		13.6	0.3	0.3	0.4
4)	8.5 (12×12)		7.0	0	0.8	0.2
5)	**22.6 ($\mathbf{24 \times 24}$)**		22.0	0	0.4	0.6

carried out (Table 1), 2DLDA achieves the best results for 3 of them (frontal, illumination and expression). On the other hand 2DPCA (for 15° rotated) and CSA (for occluded) achieve the best results for 2 experiments with images suffering from very strong difficulties.

The results of these experiments show how difficult is to provide guidelines and information for practitioners to select not only the best performing method, but also an adequate choice in the dimension d or $(h' \times w')$. What we can clearly state as a main conclusion is that in none of the 5 experiments that have been exhaustively carried out the result of the fusion at the matching score level following (9) worsen the best result achieved individually by the best performing method for each experiment. On the contrary, results are equal or slightly improved, though no significantly, by the fusion strategy.

Therefore, the main guideline proposed in our work is the use of fusion at the matching score level schemes with different dimension reduction methods as the ones used in this work. We can consider that there is not a specific best approach to face the variations to appear in a face verification problem. Our present work shows, under the small sample size problem assumptions, that the fusion approach can obtain as good results as the obtained by the best individual approach.

Deeper work has to be done to combine and use in an optimum way all the information provided by the dimension reduction methods. The inclusion of recently presented variations and extensions, as Non-iterative Generalized Low Rank Approximations [15], face new challenges on the way of combining the different scores.

References

1. Turk, M., Pentland, A.: Eigenfaces for recognition. Journal of Cognitive Neurosicience **3** (1999) 71–86
2. Belhumeur, P., Hespanha, J., Kriegman, D.: Eigenfaces vs. fisherfaces: recognition using class specific linear projection. IEEE Transactions on Pattern Analysis and Machine Intelligence **19** (1997) 711–720
3. Fortuna, J., Capson, D.: Improved support vector classification using PCA and ICA feature space modiffication. Pattern Recognition **37** (2004) 1117–1129
4. Ross, A., Jain, A.: Information fusion in biometrics. Pattern Recognition Letters **24** (2003) 2115–2125
5. Guo, G., Dyer, C.: Learning from examples in the small sample case: face expression recognition. IEEE Transactions on Systems, Man, and Cybernetics-Part B **35** (2005) 477–488
6. Pang, S., Kim, D., Bang, S.Y.: Memebership authentication in the dynamic group by face classification using SVM ensemble. Pattern Recognition Letters **24** (2003) 215–225
7. Yang, J., Yang, J.: From image vector to matrix: a straightforward image projection technique–IMPCA vs. PCA. Pattern Recognition **35** (2002) 1997–1999
8. Yang, J., Zhang, D., Frangi, F., Yang, J.: Two-dimmensional PCA: A new approach to apperance-based face representation and recognition. IEEE Transacctions on Pattern Analysis and Machine Intelligence **26** (2004) 131–137
9. Chen, S., Zhu, Y., Zhang, D., Yang, J.: Feature extraction approaches based on matrix pattern: MatPCA and MatFLDA. Pattern Recognition Letters **26** (2005) 1157–1167
10. Li, M., Yuan, B.: A novel statistical linear discriminant analysis for image matrix: two-dimensional fisherfaces. In: Proceedings of the International Conference on Signal Processing. (2004) 1419–1422
11. Xu, D., Yan, S., Zhang, L., Liu, Z., Zhang, H.: Coupled subspace analysis. Technical Report MSR-TR-2004-106, Microsof Research (2004)
12. Ye, J.: Generalized low rank approximations of matrices. Machine Learning **61** (2005) 167–191
13. Cortes, C., Vapnik, V.: Support vector network. Machine Learning **20** (1995) 273–297
14. Joachims, T.: Making large scale support vector machine learning practical. In: Advances in Kernel Methods: Support Vector Machines. MIT Press, Cambridge, MA (1998)
15. Liu, J., Chen, S.: Non-iterative generalized low rank approximation of matrices. Pattern Recognition Letters **27** (2006) 1002–1008