Improving the Robustness of Subspace Learning Techniques for Facial Expression Recognition

Dimitris Bolis, Anastasios Maronidis, Anastasios Tefas and Ioannis Pitas

Aristotle University of Thessaloniki, Department of Informatics
Box 451, 54124 Thessaloniki, Greece
email: {mpolis, amaronidis, tefas, pitas}@aiia.csd.auth.gr *

Abstract. In this paper, the robustness of appearance-based, subspace learning techniques for facial expression recognition in geometrical transformations is explored. A plethora of facial expression recognition algorithms is presented and tested using three well-known facial expression databases. Although, it is common-knowledge that appearance based methods are sensitive to image registration errors, there is no systematic experiment reported in the literature and the problem is considered, a priori, solved. However, when it comes to automatic real-world applications, inaccuracies are expected, and a systematic preprocessing is needed. After a series of experiments we observed a strong correlation between the performance and the bounding box position. The mere investigation of the bounding box's optimal characteristics is insufficient, due to the inherent constraints a real-world application imposes, and an alternative approach is demanded. Based on systematic experiments, the database enrichment with translated, scaled and rotated images is proposed for confronting the low robustness of subspace techniques for facial expression recognition.

Key words: Facial Expression Recognition, Appearance Based Techniques, Subspace Learning Methods.

1 Introduction

Visual communication plays a central role in human communication and interaction. Verbal information does not consist the total information used in human communication. Facial expressions and gestures are also of great importance in everyday life, conveying information about emotion, mood and ideas. Consequently, the successful recognition of facial expressions will significantly facilitate the human-computer interaction.

Research in psychology [1] has indicated that at least six emotions (anger, disgust, fear, happiness, sadness and surprise) are universally associated with

* This work has been funded by the Collaborative European Project MOBISERV FP7-248434 (http://www.mobiserv.eu), An Integrated Intelligent Home Environment for the Provision of Health, Nutrition and Mobility Services to the Elderly.

distinct facial expressions. According to this approach these are the basic emotional states which are inherently registered in our brain and recognized globally. Several other facial expressions corresponding to certain emotions have been proposed but remain unconfirmed as universally discernible. In this paper we focus on the facial expressions deriving from these particular emotions and the neutral emotional state.

A transparent way of monitoring emotional state is by using a video camera, which automatically detects human face and captures the facial expressions. Following this approach the data used for input to the expression analyst tool would be a video stream, namely successive luminance images. Many techniques have been proposed in the literature for facial expression recognition [2]. Among them, appearance based methods followed by subspace learning methods are the most popular approach. In subspace techniques the initial image is decomposed in a 1-D vector by row-wise scanning and bases that optimize a given criterion are calculated. Then, the high dimensionality of the initial image space is reduced into a lower one. A simple distance measure is usually applied at the new space in order to perform classification. Various criteria have been employed in order to find the bases of the low dimensional spaces. Some of them have been defined in order to find projections that express the population in an optimal way without using the information about the way the data are separated to different classes, (e.g., Principal Component Analysis (PCA) [3], Non-Negative Matrix Factorization (NMF) [4]). While, other criteria deal directly with the discrimination between classes, e.g. Discriminant NMF (DNMF) [5], Linear Discriminant Analysis (LDA) [6].

The appearance based methods disadvantage is their sensitivity to image registration errors. However, for all the cases, the problem of image registration prior to recognition is considered solved and isn't discussed. As a result, the preprocessing steps are not clearly described, implying that only small displacements of the bounding box may occur, which cannot result in considerable lower performance. This is not the case in automatic real-world applications, which, often, significantly fail to calculate the optimal geometrical characteristics of the bounding box, when even slight distortions could lead in great differences regarding the performance.

The aim of this paper is two fold. Firstly, to illustrate the sensitivity of appearance based subspace learning methods when the registration of the face prior to recognition fails, even for one pixel. Secondly, to propose a training set enrichment approach and the corresponding subspace learning methods for improving significantly the performance of these techniques in the facial expression recognition problem.

In the analysis done in this paper, the 2-D images have been decomposed into 1-D vectors in order to be used as inputs in the subspace techniques. The remainder of the paper is organized as follows: Section 2 is devoted to Subspace Learning Techniques. It is divided into three subsections. PCA, LDA and DNMF are presented in each of them respectively. In Section 3 K-Nearest Neighbor (KNN), Nearest Centroid (NC) and Support Vector Machines (SVM) classifiers are concisely described while in Section 4 a number of experimental results on BU, JAFFE and KOHN-KANADE databases are presented. Finally, in Section 5 the conclusion is drawn.

2 Subspace Techniques

2.1 Principal Component Analysis

Principal Component Analysis (PCA) is an unsupervised subspace learning technique. Let $\mathbf{x} \in \mathbb{R}^M$ be a random vector. The objective in PCA is to find projection vectors $\mathbf{w}_i$ that maximize the variance of the projected samples $\mathbf{z}_i = \mathbf{w}^T\mathbf{x}$. Assuming that the expected value of $\mathbf{x}$ is zero, the problem of finding the projections $\mathbf{w}_i$ is an eigenanalysis problem of the covariance matrix $\mathbf{C} = E[\mathbf{x}\mathbf{x}^T]$. The transformation matrix $\mathbf{W} = [\mathbf{w}_1\mathbf{w}_2 \cdots \mathbf{w}_N]$ comprises by the eigenvectors of $\mathbf{C}$ that correspond to the M' maximum eigenvalues of $\mathbf{C}$. Any data point (vector) $\mathbf{x}$ from the initial space can now be approximated by a linear combination of the M' first eigenvectors to produce a new M'-dimensional vector. This approach achieves projection of the data from the initial space to a new feature space with a predefined dimensionality. In PCA someone has to decide beforehand on the new dimensionality M' or alternatively the new dimensionality may be defined by the percentage of the total sum of the eigenvalues that should be retained after the projection. This percentage essentially indicates the proportion of the information to be retained. The main property of PCA is that it generates uncorrelated variables from initial possibly correlated ones. The disadvantage of PCA is that it might lose much discriminative information of the data, since it does take into account the class labels of the data.

2.2 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) in contrast to PCA is a supervised method for dimensionality reduction. It tries to find a transform to a low-dimensional space such that when $\mathbf{x}$ is projected, classes are well separated. Let us denote by C the total number of classes, by $\boldsymbol{\mu}_i$ the mean vector of class i, by $\boldsymbol{\mu}$ the mean vector of the whole data set and by N_i the number of samples belonging to class i. The objective of LDA is to find $\mathbf{w}$ that maximizes

$$J\left(\mathbf{W}\right) = \frac{tr[\mathbf{W}^T\mathbf{S}_B\mathbf{W}]}{tr[\mathbf{W}^T\mathbf{S}_W\mathbf{W}]},$$

where $tr[\cdot]$ denotes the trace of a matrix and

$$\mathbf{S}_B = \sum_{i=1}^{C}\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}\right)\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}\right)^T$$

is the between-class scatter and

$$\mathbf{S}_W = \sum_{i=1}^{C}\sum_{k=1}^{N_i}\left(\mathbf{x}_k^i - \boldsymbol{\mu}_i\right)\left(\mathbf{x}_k^i - \boldsymbol{\mu}_i\right)^T$$

is the within-class scatter. That is, LDA tries to maximize the distance between the mean vectors of the classes, while minimizing the variance inside each class. The solution of this problem is given by the generalized eigenvalue decomposition of $\mathbf{S}_W^{-1}\mathbf{S}_B$ keeping again the largest eigenvalues. LDA in contrast to PCA, takes into consideration both the within-class scatter and the between-class scatter carrying more discriminant information of the data. LDA is capable of retaining up to $C-1$ dimensions, where C is the total number of classes.

2.3 Discriminant Non-negative Matrix Factorization

The DNMF is a supervised NMF based method that decomposes the feature vectors into parts enhancing the class separability at the same time. The 2-D image of F pixels is row-wise scanned resulting in the vector $\mathbf{x} = [x_1, x_2, \cdots, x_F]^T$. The NMF then tries to approximate the vector $\mathbf{x}$ with a linear combination of the columns of the vector $\mathbf{h}$ such that $\mathbf{x} \simeq \mathbf{Zh}$, where $\mathbf{h} \in \mathbb{R}_+^M$. In general, $M < F$, namely the NMF produce a vector of a lower dimension, compared to the initial vector $\mathbf{x}$. The matrix $\mathbf{Z} \in \mathbb{R}_+^{F \times M}$ is a non negative matrix, whose columns sum to one. The approximation $\mathbf{x} \simeq \mathbf{Zh}$ imposes a certain error, whose value is calculated using the Kullback- Leibler divergence $KL\left(\mathbf{x}\|\mathbf{Zh}\right)$ [7]. The decomposition cost is the sum of the KL divergences for the total number of the feature vectors. This way the following metric can be calculated:

$$D\left(\mathbf{X}\|\mathbf{ZH}\right) = \sum_j KL\left(\mathbf{x}_j\|\mathbf{Zh}_j\right) =$$

$$= \sum_{i,j}\left(x_{i,j}\ln\left(\frac{x_{i,j}}{\sum_k z_{i,k}h_{k,j}}\right) + \sum_k z_{i,k}h_{k,j} - x_{i,j}\right)$$

as the measure of the cost for approximating $\mathbf{X}$ with $\mathbf{ZH}$ [7]. The NMF is the outcome of the following optimization problem:

$$\min_{\mathbf{Z},\mathbf{H}} D\left(\mathbf{X}\|\mathbf{ZH}\right) \quad \text{subject to}$$

$$z_{i,k} \geq 0, h_{k,j} \geq 0, \sum_i z_{i,j} = 1, \forall j.$$

All the elements of $\mathbf{Z}$ and $\mathbf{H}$ should be non negative real numbers. This way, the vector $\mathbf{h}_j$ represents the weight vector and the $\mathbf{Z}$ matrices the M basis images, whose linear combination result in the initial image, permitting only additions between the different basis images.

The DNMF algorithm can be considered as an alternative to NMF plus LDA method [5]. In the case of DNMF, discriminant constraints are incorporated inside the cost of NMF. This form of decomposition leads to the creation of basis images that correspond to discreet parts of the face (e.g., mouth, eyes).

The modified divergence is constructed deriving from the minimization of the Fisher criterion using the new cost function given by:

$$D_d\left(\mathbf{X}\|\mathbf{ZH}\right) = D\left(\mathbf{X}\|\mathbf{ZH}\right) + \gamma tr[\mathbf{S}_w] - \delta tr[\mathbf{S}_b],$$

where γ and δ are constants and $tr[\cdot]$ is the trace of its argument. The minimization of this function is done by finding the minimum for the $tr[\mathbf{S}_w]$ term, and the maximum for the $tr[\mathbf{S}_b]$ one.

The vector $\mathbf{h}_j$ that corresponds to the j-th column of the matrix $\mathbf{H}$, is the coefficient vector for the ρ-th facial image of the r-th class and will be denoted as $\mathbf{h}_\rho^{(r)} = [h_{\rho,1}^{(r)}, h_{\rho,2}^{(r)}, \cdots, h_{\rho,M}^{(r)}]^T$. The mean vector of the vectors $\mathbf{h}_\rho^{(r)}$ for the rth class is denoted as $\boldsymbol{\mu}_\rho^{(r)} = [\mu_1^{(r)}, \mu_2^{(r)}, \cdots, \mu_M^{(r)}]^T$ and the mean of all the classes as $\boldsymbol{\mu} = [\mu_1, \mu_2, \cdots, \mu_M]^T$. Then, the within scatter for the coefficient vectors $\mathbf{h}_j$ is defined as:

$$\mathbf{S}_w = \sum_{r=1}^{K} \sum_{\rho=1}^{N_r} \left(\mathbf{h}_\rho^{(r)} - \boldsymbol{\mu}^{(r)}\right) \left(\mathbf{h}_\rho^{(r)} - \boldsymbol{\mu}^{(r)}\right)^T,$$

whereas the between scatter matrix is defined as:

$$\mathbf{S}_b = \sum_{r=1}^{K} N_r \left(\boldsymbol{\mu}^{(r)} - \boldsymbol{\mu}\right) \left(\boldsymbol{\mu}^{(r)} - \boldsymbol{\mu}\right)^T.$$

The matrix $\mathbf{S}_w$ defines the scatter of the sample vector coefficients around their class mean and a convenient measure for the dispersion of the samples is the trace of $\mathbf{S}_w$. While, the matrix $\mathbf{S}_b$ denotes the between-class scatter matrix and defines the scatter of the mean vectors of all classes around the global mean $\boldsymbol{\mu}$. At this point become obvious why by minimizing and maximizing the traces of $\mathbf{S}_w$ and $\mathbf{S}_b$ respectively, we shrink the classes and increase the separability among them.

This class-specific decomposition is intuitively motivated by the theory that humans use specific discriminant features of the human face for memorizing and recognizing them [8].

All the above methods, aim at projecting the initial high-dimensional datapoints to a feature space with low dimensionality. In that new space, the datapoints are likely to be classified in a more efficient way. In our study, we use three well-known in the literature classifiers, the K-Nearest Neighbor (KNN), the Nearest Centroid (NC) and the Support Vector Machines (SVMs). They are all concisely described in the following paragraph.

3 Classifiers

The K-Nearest Neighbour is a non-linear voting classifier. A datapoint is assigned to the most common class among its K nearest neighbours. In NC the centroids of the several classes are calculated and the datapoint is assigned to the class with the nearest centroid to it.

A support vector machine tries to calculate the optimal hyperplane or set of hyperplanes in a high dimensional space. Intuitively, a good separation is achieved by the hyperplane that maximizes the functional margin, since, in general, the larger the margin the lower the generalization error of the classifier. The SVMs used for our experiments were proposed in [9], and use a modified method to calculate the maximum functional margin, inspired by the Fisher's discriminant ratio. The SVMs are successively applied for a 2-class problem each time. The winning class is then compared with one of the remaining classes following the same method and the procedure is repeated until the prevailing class for each test sample is found.

4 Experimental Results

For our experiments we used the BU [10], JAFFE [11] and COHN- KANADE [12] databases for facial expression recognition. BU contains images from 100 subjects, captured in four facial expressions intensities for each of the six, universally recognized, emotions (anger, disgust, happiness, fear, sadness, surprise) and one neutral pose for each person, namely a total of 2500 images. It contains subjects (56% female, 44% male), ranging from 18 years to 70 years old, with a variety of ethnic/racial ancestries, including White, Black, East-Asian, Middle-east Asian, Indian and Hispanic Latino. We used the most expressive of each facial expression. Thus, the final database we utilized consisted of 700 images.

The JAFFE database contains 213 images of the 7 aforementioned facial expressions, posed by 10 Japanese female models. Each image has been rated on these emotion adjectives by 60 Japanese subjects. Finally, from the COHN-KANADE we used 407 images from 100 subjects. Subjects, in this case, range in age from 18 to 30 years. Sixty-five percent were female; 15 percent were African-American and three percent Asian or Latino. In Figures 1 and 2 typical examples of the six expressions and the neutral case for the JAFFE and COHN-KANADE database are illustrated, respectively.

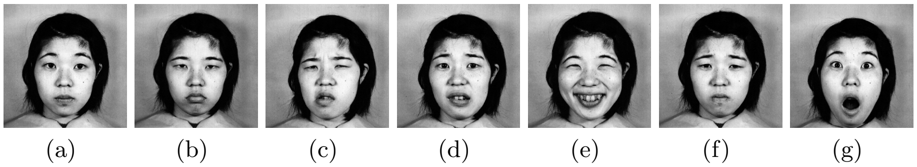

Fig. 1. The JAFFE Facial Expression Database (a) neutral, (b) angry, (c) disgusted, (d) feared, (e) happy, (f) sad, (g) surprised.

We preprocessed the images manually in order to have the eyes in fixed predefined positions in the frame. Firstly, we gathered the coordinates of the eyes in the initial images. The initial distance between the eyes was calculated and the

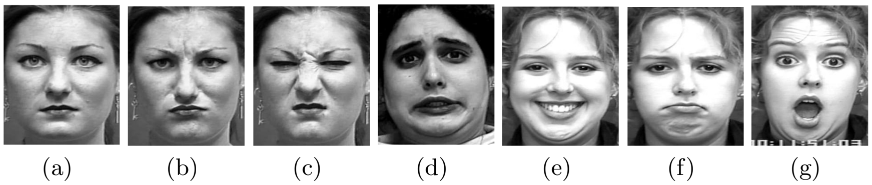

Fig. 2. The Cohn-Kanade Facial Expression Database (a) neutral, (b) angry, (c) disgusted, (d) feared, (e) happy, (f) sad, (g) surprised.

image was down-scaled in an isotropic way, in order to succeed a 16-pixel eyes distance. In the final step we cropped the image to the size of 40×30 producing a bounding box centered to the subject's face. The image cropping was based on the eyes due to their inherent attribute of maintaining fixed position, independently to the various facial expressions. Other features of the face (e.g., mouth, eye-brows) have the tendency to be shifted in other positions, regarding certain expressions. For example, in the case of surprise the eye-brows appear in higher position in comparison with the neutral expression. Thus, manual cropping based on other features, apart from the eyes, could produce discriminant information on itself leading to overestimation of the performance of the classification.

Under this perspective, we constructed two versions of enriched databases. For the first one (enriched database), the above mentioned centered images were shifted one pixel in the four basic directions (left, right, up and down). In the second case (fully enriched database), one cross of five possible positions for each eye was considered (original position, one-pixel left, right, up and down), resulting into 25 different possible pairs of eyes. The position of these pairs of eyes were then used for the production of the final centered following the above mentioned procedure for centering the images, resulting in translated, rotated and scaled images.

In a second level, we implemented a number of combinations of subspace learning techniques exploiting the PCA, LDA and DNMF algorithms and the three well known classifiers NCC, KNN and SVMs in order to examine their effectiveness in classifying the aforementioned facial expressions along with the neutral emotional state. For this purpose we conducted a five-fold cross-validation. Regarding the PCA and LDA outputs we used the Nearest Centroid algorithm, while, the SVM method was applied at the DNMF algorithm outputs.

We conducted three series of experiments. In the first one, the centered images were used to form both the training and the testing set. Secondly, the centered images were used for the training set, while the left-shifted images were used for the testing set, in order to examine the sensitivity of the performance in displacements of the bounding box. In the last series of experiments, the training was formed from the whole set of the images (both centered and shifted images), while the centered images alone constituted the testing set.

The comparative results, for the KANADE, JAFFE and BU database, are depicted in the Tables 1, 2 and 3 respectively. In the first two columns of the tables the various methods utilized are given, both for reducing the dimensionality and for classifying the samples.

KNN was used for $K = 1$ and $K = 3$. The second case is presented, due to its better results. As far as the presented method of PCA plus LDA is concerned, PCA used for maintaining the 95% of the covariance matrix energy while the LDA reduced the resulted vector to the dimension of 6. The cases of maintaining other percentages of the covariance matrix energy were tested as well, without leading to better results. Regarding the DNMF followed by SVM approach, the dimension of the feature vector was reduced from 1200 to 120 by the DNMF and then the SVM realized the classification using a RBF kernel. Other type of kernels were, also, used producing similar results.

In the third column of the tables there are the success rates, in the case of the centered images, for both the training and testing set. The next column shows the performance when misplaced images are used for the testing set (1-pixel misplacement on the horizontal axis in this case). In the fifth column, the performance of the enriched database, exploiting, merely, the translated images, is depicted. Finally in the last column, the performance of the fully enriched database is appeared, where the 25 transformed versions of the original database were used.

On one hand, it can be, easily observed, that even a slight divergence from the centered images (one pixel in the case of our experiments) lead in significant lower performance (up to 8%). On the other hand, after the enrichment with transformed images, a clear improvement in the performance is observed in the vast majority of the cases for both the two versions of the database enrichment (up to 15.9% for the enrichment with the translated images and 22.3% for the fully enriched version). Both the sensitivity in small translations of the bounding box and the robustness when enriching the training set are systematically observed in our experiments. Additionally, it was observed that the more transformations are used the greater the improvement of the accuracy becomes.

Table 1. KANADE 5-fold cross validation accuracy rates

Classifier	Approach	Centered(%)	Misplaced(%)	Enriched(%)	Fully Enriched(%)
NC	PCA	36.4	36.0	36.5	39.7
	LDA	62.5	55.0	72.4	74.9
	PCA+LDA	67.0	65.1	68.8	73.7
KNN	PCA	39.0	39.2	39.7	38.5
	LDA	63.3	55.7	71.6	75.7
	PCA+LDA	67.3	65.8	67.6	69.4
SVM	DNMF	56.4	49.4	67.6	69.2

Table 2. JAFFE 5-fold cross validation accuracy rates

Classifier	Approach	Centered(%)	Misplaced(%)	Enriched(%)	Fully Enriched(%)
NC	PCA	29.0	26.0	27.5	34.6
	LDA	53.5	45.5	51.5	62.9
	PCA+LDA	54.5	46.5	63.5	62.4
KNN	PCA	31.5	31.0	26.0	40.0
	LDA	52.5	44.5	51.5	62.0
	PCA+LDA	57.0	48.5	58.5	64.9
SVM	DNMF	41.6	34.6	57.5	63.9

Table 3. BU 5-fold cross validation accuracy rates

Classifier	Approach	Centered(%)	Misplaced(%)	Enriched(%)
NC	PCA	34.6	34.0	34.9
	LDA	56.0	54.4	62.3
	PCA+LDA	63.3	62.3	64.9
KNN	PCA	33.1	33.0	32.7
	LDA	56.6	53.7	61.3
	PCA+LDA	60.4	60.0	62.1
SVM	DNMF	55.4	53.0	61.4

5 Conclusion

Facial expressions consist an integral part of the human communication. Efficient methods for recognizing human emotions, exploiting the facial expressions, are expected to revolutionize the scientific field of human-machine interaction. Subspace learning techniques followed by well known classifiers are among the most used methods for human facial expression recognition. However, after a series of experiments we observed a great sensitivity of this kind of algorithms to geometrical translation of the images, even for the case of one pixel. Real-world applications carry an inherent difficulty regarding the precise detection of the facial characteristics' position, resulting in inaccurate image registering. The experiments, we conducted, show that the systematic enrichment of a database with geometrically transformed (translated, scaled and rotated) images results in significant improvement in the performance in the majority of the cases. By using more sophisticated transformations for enriching the initial databases, in the future, further improvement in the performance could is expected.

References

1. Ekman, P., Friesen, W.V.: Constants across cultures in the face and emotion. Journal of Personality and Social Psychology **17**(2) (1971) 124–129
2. Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S.: A survey of affect recognition methods: audio, visual and spontaneous expressions. In: ICMI '07: Proceedings of

the 9th international conference on Multimodal interfaces, New York, NY, USA, ACM (2007) 126–133

3. Jolliffe, I.: Principal Component Analysis. Springer Verlag (1986)
4. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature **401** (1999) 788–791
5. Zafeiriou, S., Tefas, A., Buciu, I., Pitas, I.: Exploiting discriminant information in non negative matrix factorization with application to frontal face verification. **17**(4) (2007) 395–416
6. Belhumeur, P.N., Hespanha, J.P., Kriegman, D.J.: Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. IEEE Transactions on Pattern Analysis and Machine Intelligence **19**(7) (August 1997) 711–720
7. D.D.Lee, H.S.Seung: Algorithms for non-negative matrix factorization. NIPS (2000) 556–562
8. Chellappa, R., Wilson, C.L., Sirohey, S.: Human and machine recognition of faces: a survey. Proceedings of the IEEE **83**(5) (May 1995) 705–741
9. Tefas, A., Kotropoulos, C., Pitas, I.: Using support vector machines to enhance the performance of elastic graph matching for frontal face authentication. IEEE Transactions on Pattern Analysis and Machine Intelligence **23** (2001) 735–746
10. Yin, L., Wei, X., Sun, Y., Wang, J., Rosato, M.J.: A 3d facial expression database for facial behavior research. In: FGR '06: Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition, Washington, DC, USA, IEEE Computer Society (2006) 211–216
11. Lyons, M., Akamatsu, S., Kamachi, M., Gyoba, J.: Coding facial expressions with gabor wavelets. (1998) 200–205
12. Kanade, T., Cohn, J.F., Tian, Y.: Comprehensive database for facial expression analysis. Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition, Grenoble, France (2000) 46–53