Neural Networks 24 (2011) 814–823

Contents lists available at SciVerse ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

2011 Special Issue

# Improving subspace learning for facial expression recognition using person dependent and geometrically enriched training sets

Anastasios Maronidis, Dimitris Bolis, Anastasios Tefas *, Ioannis Pitas

*Aristotle University of Thessaloniki, Department of Informatics, Box 451, 54124 Thessaloniki, Greece*

ARTICLE INFO

*Keywords:*
Facial expression recognition
Appearance based techniques
Subspace learning methods

ABSTRACT

In this paper, the robustness of appearance-based subspace learning techniques in geometrical transformations of the images is explored. A number of such techniques are presented and tested using four facial expression databases. A strong correlation between the recognition accuracy and the image registration error has been observed. Although it is common-knowledge that appearance-based methods are sensitive to image registration errors, there is no systematic experiment reported in the literature. As a result of these experiments, the training set enrichment with translated, scaled and rotated images is proposed for confronting the low robustness of these techniques in facial expression recognition. Moreover, person dependent training is proven to be much more accurate for facial expression recognition than generic learning.

© 2011 Elsevier Ltd. All rights reserved.

## 1. Introduction

Facial expressions and gestures complement verbal communication in everyday life, conveying information about emotion, mood and ideas (Zeng, Pantic, Roisman, & Huang, 2009). The facial expressions play central role in an everyday conversation. Even the voice intonation present lower impact on efficient communication than the facial expressions do (Mehrabian, 1968). It has been noted in the literature that the ideal model in human–computer communication, would be the human–human communication paradigm (Bruce, 1993; Hara & Kobayashi, 1996; Takeuchi et al., 1993). Consequently, a successful automatic facial expression recognition system is expected to significantly facilitate the human–computer interaction. Furthermore, it could be integrated in many technologies of this kind, bordering behavioral science and medicine, (e.g., assisted living) (Pantic & Rothkrantz, 2000).

Research in psychology Ekman and Friesen (1971) has indicated that at least six emotions (anger, disgust, fear, happiness, sadness and surprise) are universally associated with distinct facial expressions. According to this approach, these are the basic emotional states which are inherently registered in human brain and are universally recognized. Several other facial expressions corresponding to certain emotions have been proposed, but remain unconfirmed as universally discernible (Ekman & Friesen, 1971). In this paper, we focus on the facial expressions deriving from these particular emotions and the neutral emotional state. In the next paragraph a brief outline of a real world system, used for facial expression recognition, will be presented.

A transparent way of monitoring the human emotional state is by using a video camera, which automatically detects human face and captures the facial expressions. Following this approach, the data used as input to the expression analysis tool would be a video stream, namely successive luminance or color image frames. Many techniques have been proposed in the literature for implementing this tool. Some of them use static images, while others work with image sequences. Furthermore, the image representations used for expression recognition are local or global ones. Local (or landmark-based) techniques employ fiducial image points or point grids (e.g., the CANDIDE model) and their deformations for facial expression recognition (Kotsia, Zafeiriou, & Pitas, 2007). Global techniques use image features derived from the entire facial image region of interest (ROI) (Kyperountas, Tefas, & Pitas, 2010). The classification techniques operating on these image representations, have been categorized into template-based, also known as appearance-based methods, (fuzzy) rule-based, ANN-based, HMM-based and Bayesian (Pantic & Rothkrantz, 2003). In the following two paragraphs the subspace learning methods, which are commonly used in appearance-based approaches, will be introduced.

Subspace learning methods are based on principles originally used for statistical pattern recognition and have been successfully implemented in many computer vision problems, such as facial expression classification (Kyperountas et al., 2010), human face recognition (Kyperountas, Tefas, & Pitas, 2008) and object recognition (Leibe & Schiele, 2003). The problem that emerges, when it comes to appearance-based methods, is that usually initial images

* Corresponding author. Tel.: +30 2310991932.
*E-mail addresses:* tefas@aiia.csd.auth.gr (A. Tefas), pitas@aiia.csd.auth.gr (I. Pitas).

0893-6080/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2011.05.015

lie on a high dimensional space. The main goal of subspace learning methods is to reduce the data dimensionality, maintaining the meaningful information. These techniques simplify the problem of dimensionality reduction to a simple multiplication between a matrix (projection matrix) and a vector (initial image). In other words, both information extraction and computational load decrease are achieved by applying such algorithms.

In subspace learning techniques, the initial image is decomposed in a 1D vector by row-wise scanning and bases that optimize a given criterion are derived. Then, the high dimensionality of the initial image space is reduced into a lower one. Several 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 represent the data in an optimal way, without using the information about the way the data are separated to different classes, e.g., Principal Component Analysis (PCA) (Jolliffe, 1986) and Non-Negative Matrix Factorization (NMF) (Lee & Seung, 1999). Other criteria deal directly with the discrimination between classes, e.g., Discriminant Non-Negative Matrix Factorization (DNMF) (Zafeiriou, Tefas, Buciu, & Pitas, 2006), Linear Discriminant Analysis (LDA) (Belhumeur, Hespanha, & Kriegman, 1997) and Clustering Discriminant Analysis (CDA) (Chen & Huang, 2003). Subspace learning methods are usually combined with a classifier, like *k*-Nearest Neighbor (KNN), Nearest Centroid (NC), Nearest Cluster Centroid (NCC) or Support Vector Machine (SVM) in order to classify the data in the new low-dimensional space.

Among the various subspace learning methods LDA is the most popular when the objective is classification. However, LDA confronts some fundamental problems. One of them is the small sample size problem, where the number of samples is smaller than their dimensionality. In Kyperountas, Tefas, and Pitas (2007), a method for overcoming this problem in face verification has been proposed. Another problem with LDA is that it is capable of retaining as many projections as the number of classes minus one. This is a very strict limitation, especially when dealing with two-class problems, where the maximum number of projections is one. In Goudelis, Zafeiriou, Tefas, and Pitas (2007), a class-specific approach for face verification has been proposed in order to overcome this limitation.

Another limitation of LDA that emerges from Bayesian theory is the assumption that the classes have multi-variate Gaussian distribution. However, usually the data within a class are not normally distributed. For instance, a class might consist of a mixture of Gaussians. Clustering Discriminant Analysis (CDA) (Chen & Huang, 2003) is a subspace learning method that has been developed in order to handle such cases. Specifically, CDA introduces a different kind of labeling which relies on the clustering of the data samples. Thus, it attempts to exploit the potential subclass structure of the classes of the data.

As has mentioned, the first crucial step toward automatic facial expression recognition is face detection. The output of this procedure is a bounding box (facial region of interest, facial ROI), which is ideally placed around the facial area. The image information within this bounding box is subsequently used as input to the classification algorithm. When it comes to theoretical analysis on the classification performance of all the aforementioned algorithms, the problem of image registration prior to recognition is considered solved. However, this is not the case in most of the real-world applications. Although, it is common knowledge that appearance-based methods are sensitive to image registration errors, there is no systematic experimental study, resulting in a feasible solution, reported in the literature. In general, the preprocessing steps are usually not clearly described and the bounding box, used for recognition, is arbitrarily selected, implying that only small displacements of the bounding box may occur. However, when it comes to automatic real-world applications, inaccuracies regarding the face detection are expected and a systematic preprocessing is needed. An experimental analysis on quantifying the misclassification probability due to registration error has been done in Rentzeperis, Stergiou, Pnevmatikakis, and Polymenakos (0000). However, the authors do not propose a solution for improving the overall performance of a real-world application.

An additional major source of inaccuracies could be attributed to the difficulty of creating a single model that could operate optimally in cases of different people. It is common-knowledge that there is a great variation in the way several facial expressions are performed by distinct persons, due to personality or cultural background variations. This fact creates difficulties in developing a generic facial expression recognition algorithm. However, there are cases, where the expressors are, a priori, known. For instance, in cognitive robotics for assisted living, the persons that interact with the robot are typically known, are few (in many cases just one person) and do not change over a long period of time. In this case, attempting to model the way that the facial expressions are performed by the specific persons is more reasonable rather than using a generic approach.

The motivation of our work was to create a facial expression recognition system that would be fast and would operate in realistic assisted living environments involving few persons (e.g., one elderly person living independently) (Nani, Caleb-Solly, Dogramadgi, Fear, & van den Heuvel, 0000). Our experiments reveal that Landmark-based systems tend to be slow and error-prone, since facial landmark detection and tracking frequently fails. Therefore, we abandoned the idea of using grid-based facial expression recognition. The solution we followed was the one based on subspace techniques. However, we had to study the robustness of such techniques to the frequently occurring face detection and tracking inaccuracies.

The aim of this paper is three-fold. First, to illustrate the sensitivity of subspace learning methods when the registration of the facial ROI prior to recognition fails, even slightly ($\approx$6% on the distance between the eyes). For instance, the eye perturbation that results using a standard face detection scheme is more than 7% for human scan faces decimated to 10 pixels eye distance (Rentzeperis et al., 0000). Additionally, we would like to illustrate that the inter-database recognition performance is much worse than the intra-database performance that is usually reported in the literature. Second, to propose a training set enrichment approach for improving significantly the performance of subspace learning techniques in the facial expression recognition problem. Moreover, to highlight that even perfect manual face alignment in high resolution can be improved by the proposed training set enrichment. Third, to indicate the contribution of enriching the training set with images of a tested person, in order to create person specific recognizers, thus, improving the subspace learning and the recognition performance.

The remainder of this paper is organized as follows. In Section 2, the subspace learning techniques and classifiers that have been utilized for producing this paper's results, are presented. Additionally, the Spectral Clustering method that has been used in order to produce the subclass labels required by the CDA algorithm, is described in detail. In Section 3, the whole facial expression recognition procedure that has been followed is described. In Section 4, the proposed approach for solving the image registration problem, along with the person dependent training approach are explained. Conclusions are drawn in Section 5.

## 2. Subspace learning techniques

In the following analysis, as mentioned above, the 2D facial image ROIs have been decomposed into 1D vectors by row-wise

scanning in order to be used as inputs in the subspace learning techniques. From now on, we consider a set $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ of $n$ such vectors (called data samples), corresponding to one image each. We denote by $\mathbf{y} = \mathbf{V}^T\mathbf{x}$ the projection of $\mathbf{x}$ to the new, low dimensional space using $\mathbf{V}$ as the projection matrix. The initial dimensionality of the data is denoted by $m$, while the dimensionality of the projection space is denoted by $m'$.

### 2.1. Principal Component Analysis

Principal Component Analysis (PCA) (Jolliffe, 1986) is an unsupervised subspace learning technique. Assuming that the mean vector of $\mathbf{x}_q$ is zero, the problem of finding the projection matrix $\mathbf{V}$ is an eigenanalysis problem of the sample covariance matrix

$$\mathbf{C} = \frac{1}{n}\sum_{q=1}^{n}\mathbf{x}_q\mathbf{x}_q^T. \tag{1}$$

The transformation matrix $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{m'}]$ consists of the eigenvectors of $\mathbf{C}$ that correspond to the $m'$ largest eigenvalues of $\mathbf{C}$. Any data sample $\mathbf{x}$ from the initial space can be approximated by a linear combination of the $m'$ first eigenvectors to produce a new $m'$-dimensional vector. In PCA, someone has to decide directly 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 retained information.

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 discriminant information of the data, since it does not take into account the class labels of the data.

### 2.2. Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) (Belhumeur et al., 1997), in contrast to PCA, is a supervised method for dimensionality reduction. It tries to find a projection to a low-dimensional space such that, in this subspace, the classes are well discriminated. Let us denote the total number of classes by $c$, the mean vector of the whole dataset by $\boldsymbol{\mu}$ and the number of samples, the $q$-th sample and the mean vector of the $i$-th class, by $n_i$, $\mathbf{x}_q^{(i)}$ and $\boldsymbol{\mu}^{(i)}$, respectively. The objective of LDA is to find the transformation matrix $\mathbf{V}$ that maximizes

$$J(\mathbf{V}) = \frac{\mathrm{tr}[\mathbf{V}^T\mathbf{S}_B^{\mathrm{LDA}}\mathbf{V}]}{\mathrm{tr}[\mathbf{V}^T\mathbf{S}_W^{\mathrm{LDA}}\mathbf{V}]}, \tag{2}$$

where $\mathrm{tr}[\cdot]$ denotes the trace of a matrix,

$$\mathbf{S}_B^{\mathrm{LDA}} = \sum_{i=1}^{c}\left(\boldsymbol{\mu}^{(i)} - \boldsymbol{\mu}\right)\left(\boldsymbol{\mu}^{(i)} - \boldsymbol{\mu}\right)^T \tag{3}$$

is the between-class scatter and

$$\mathbf{S}_W^{\mathrm{LDA}} = \sum_{i=1}^{c}\sum_{q=1}^{n_i}\left(\mathbf{x}_q^{(i)} - \boldsymbol{\mu}^{(i)}\right)\left(\mathbf{x}_q^{(i)} - \boldsymbol{\mu}^{(i)}\right)^T \tag{4}$$

is the within-class scatter matrix. $\mathbf{V}$ is found by solving the generalized eigenvalue decomposition problem

$$\mathbf{S}_B^{\mathrm{LDA}}\mathbf{v} = \lambda\mathbf{S}_W^{\mathrm{LDA}}\mathbf{v}, \tag{5}$$

while retaining the largest eigenvalues and putting the corresponding eigenvectors in $\mathbf{V}$.

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, since the rank of $\mathbf{S}_B^{\mathrm{LDA}}$ is at most $c - 1$ (Belhumeur et al., 1997).

### 2.3. Clustering Discriminant Analysis

Clustering Discriminant Analysis (CDA) (Chen & Huang, 2003), like LDA, looks for a transform $\mathbf{V}$, such that the projections $\mathbf{y} = \mathbf{V}^T\mathbf{x}$ for each class are well discriminated. The difference from LDA is that the classes might contain many clusters (subclasses). Let us denote the total number of clusters inside the $i$-th class by $d_i$, the number of samples of the $j$-th cluster of the $i$-th class by $n_{ij}$, its $q$-th sample by $\mathbf{x}_q^{(i,j)}$ and its mean vector by $\boldsymbol{\mu}^{(i,j)}$. CDA attempts to maximize

$$J(\mathbf{V}) = \frac{\mathrm{tr}[\mathbf{V}^T\mathbf{S}_B^{\mathrm{CDA}}\mathbf{V}]}{\mathrm{tr}[\mathbf{V}^T\mathbf{S}_W^{\mathrm{CDA}}\mathbf{V}]}, \tag{6}$$

where

$$\mathbf{S}_B^{\mathrm{CDA}} = \sum_{i=1}^{c-1}\sum_{l=i+1}^{c}\sum_{j=1}^{d_i}\sum_{h=1}^{d_l}\left(\boldsymbol{\mu}^{(i,j)} - \boldsymbol{\mu}^{(l,h)}\right)\left(\boldsymbol{\mu}^{(i,j)} - \boldsymbol{\mu}^{(l,h)}\right)^T \tag{7}$$

is the between-cluster scatter and

$$\mathbf{S}_W^{\mathrm{CDA}} = \sum_{i=1}^{c}\sum_{j=1}^{d_i}\sum_{q=1}^{n_{ij}}\left(\mathbf{x}_q^{(i,j)} - \boldsymbol{\mu}^{(i,j)}\right)\left(\mathbf{x}_q^{(i,j)} - \boldsymbol{\mu}^{(i,j)}\right)^T \tag{8}$$

is the within-cluster scatter matrix. In a few words, CDA tries to discriminate subclasses belonging to different classes, while minimizing the scatter within every subclass. Also it puts no constraints on subclasses of the same class. The solution is provided by Eq. (5) using $\mathbf{S}_B^{\mathrm{CDA}}$ and $\mathbf{S}_W^{\mathrm{CDA}}$.

As already has been mentioned, the main advantage of CDA against LDA is that CDA exploits the potential subclass information to discriminate the classes. One more advantage is that CDA is capable of retaining $d - 1$ dimensions, where $d$ is the total number of subclasses of the data. Of course, $d - 1$ is greater than or at least equal to $c - 1$, which is the maximum retained dimensionality by LDA. It is worth noting that if no clusters are found in the data classes, then CDA is identical to LDA. In our study, a Spectral Clustering approach has been utilized for extracting the subclass structure of the data.

Two important mathematical tools for Spectral Clustering are the similarity graph and the affinity matrix. Consider a metric $d\left(\mathbf{x}_q, \mathbf{x}_p\right)$ and some parametric monotonically decreasing function $w_{q,p}(\sigma) = w\left(d\left(\mathbf{x}_q, \mathbf{x}_p\right), \sigma\right)$, which measures the similarity between every pair of such data samples. We define the similarity graph as the graph $(\mathcal{X}, \mathcal{E})$, where $\mathcal{X}$ is the set of the data samples (graph nodes) and $\mathcal{E}$ is the set of the edges between the data samples. The weights of the edges calculated by the similarity function $w$ constitute a matrix $\mathbf{W}$, which has at position $(q, p)$ the weight $w_{q,p}(\sigma)$ between the $q$, $p$ nodes. Of course, $\mathbf{W}$ has to be a symmetric matrix.

The affinity matrix $\mathbf{P}$ is an $n \times n$ matrix, which contains the whole node connectivity information. There are several ways to define the affinity matrix. Here we have used the random walk approach:

$$\mathbf{P} = \mathbf{D}^{-1}\mathbf{W}, \tag{9}$$

where $\mathbf{D}$ is the diagonal degree matrix, with $D_{q,q} = \sum_{p=1}^{n} w_{q,p}$.

Given the number $K$ of clusters, Spectral Clustering algorithm firstly computes the $K$ largest eigenvalues of $\mathbf{P}$. Then constructs an $n \times K$ matrix who has as columns the $K$ corresponding eigenvectors. It has been shown in von Luxburg (2007) that the rows of this matrix could be used as a new representation of the initial data, which is more useful from a clustering perspective. Thus, on this new data representation any common clustering algorithm should be employed in a more efficient way. Here, for our needs, we have employed the $K$-means algorithm (Theodoridis & Koutroumbas, 2006).

An issue that arises from the above discussion is how to estimate the 'correct' number of clusters. In our study, we have used the eigengap heuristic approach:

- Perform eigenanalysis on the affinity matrix $\mathbf{P}$.
- Rank the eigenvalues in descending order: $(\lambda_1, \lambda_2, \ldots, \lambda_n)$.
- Find the maximum gap $\delta$ between consecutive eigenvalues $(\lambda_q, \lambda_{q+1})$.
- Use the index $q$ as an estimation of the total number of clusters.
- Use this eigengap $\delta$ as a plausibility measure, where $\delta$ takes values between 0 and 1.

In Azran and Ghahramani (0000), the authors extended this heuristic. They have shown that by letting the random walk take multiple steps, different scales of partitioning are explored. In the case where the number of steps is $M$, the transition matrix is given by multiplying $\mathbf{P}$ with itself $M$ times and is then called the $M$-th order transition matrix. This matrix contains the probabilities of the random walk to transit from one state to another in $M$ steps. The idea behind this approach is to use the eigengap heuristic on these $M$-th order transition matrices for several values of $M$. It can be easily shown that the set of the eigenvalues of $\mathbf{P}^M$ is $\left(\lambda_1^M, \lambda_2^M, \ldots, \lambda_n^M\right)$. Using the eigengap heuristic on these sets for diverse values of $M$ $(1 \leq M \leq M_{\max})$, results in a set of eigengaps $\{\delta(M)\}_M$. The local maxima of this set are estimations of different scales of partitioning with plausibility measured by the corresponding $\delta$.

In the experiments, wherever we have employed the CDA method, the clustering was done by utilizing the above multi-scale approach, retaining the most plausible partition. Specifically, we have used the Euclidean metric

$$d\left(\mathbf{x_q}, \mathbf{x_p}\right) = \sqrt{\sum_{s=1}^{m} \left(x_{s,q} - x_{s,p}\right)^2}, \tag{10}$$

where $x_{s,q}$ is the $s$-th component of $\mathbf{x}_q$ and as similarity function the Gaussian similarity function which is defined as

$$f_{q,p}(\sigma) = \exp\left(-\frac{d\left(\mathbf{x_q}, \mathbf{x_p}\right)}{\sigma^2}\right). \tag{11}$$

The parameter $\sigma^2$ plays the role of the variance and determines the scale of the neighborhood of every data sample. Our empirical study, has shown that $\sigma = 0.25\hat{E}[d\left(\mathbf{x}_q, \mathbf{x}_p\right)]$ is a value which offers intuitively satisfactory results. $\hat{E}$ denotes the sample mean. Thus, we have fixed $\sigma$ to that value.

### 2.4. Discriminant non-negative matrix factorization

Non-negative matrix factorization (NMF) (Lee & Seung, 1999) tries to approximate the vector $\mathbf{x}$ with a linear combination of the columns of a lower dimensional vector $\mathbf{h}$ such that $\mathbf{x} \simeq \mathbf{Zh}$, where $\mathbf{h} \in \mathbb{R}_+^{m'}$. The matrix $\mathbf{Z} \in \mathbb{R}_+^{m \times m'}$ is a non negative matrix, whose columns sum to one. The approximation error of $\mathbf{x} \simeq \mathbf{Zh}$ is calculated using the Kullback–Leibler divergence $KL(\mathbf{x} \parallel \mathbf{Zh})$ (Lee & Seung, 2000). The decomposition cost is the sum of the KL divergences for the total number of the feature vectors:

$$D(\mathbf{X} \parallel \mathbf{ZH}) = \sum_q KL\left(\mathbf{x}_q \parallel \mathbf{Zh}_q\right) \tag{12}$$

$$= \sum_{s,q} \left(x_{s,q} \ln\left(\frac{x_{s,q}}{\sum_t z_{s,t} h_{t,q}}\right) + \sum_t z_{s,t} h_{t,q} - x_{s,q}\right), \tag{13}$$

where $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) = \left[x_{s,q}\right]$, $\mathbf{H} = (\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_n) = \left[h_{t,q}\right]$ and $\mathbf{Z} = \left[z_{s,t}\right]$.

Discriminant Non-Negative Matrix Factorization (DNMF) (Zafeiriou et al., 2006) is a supervised NMF-based method that produces discriminant non-negative feature vectors. In DNMF, a modified divergence is constructed deriving from the minimization of the Fisher criterion using the new cost function given by

$$D_d(\mathbf{X} \parallel \mathbf{ZH}) = D(\mathbf{X} \parallel \mathbf{ZH}) + \gamma \mathrm{tr}[\mathbf{S}_W^{\mathrm{DNMF}}] - \delta \mathrm{tr}[\mathbf{S}_B^{\mathrm{DNMF}}], \tag{14}$$

where $\gamma$ and $\delta$ are constants.

The vector $\mathbf{h}_\rho$ that corresponds to the $\rho$-th column of the matrix $\mathbf{H}$, is the coefficient vector for the $q$-th facial image of the $i$-th class and will be denoted by $\mathbf{h}_q^{(i)} = [h_{1,q}^{(i)}, h_{2,q}^{(i)}, \ldots, h_{m',q}^{(i)}]^T$. The mean vector of the vectors $\mathbf{h}_q^{(i)}$ for the $i$-th class is denoted as $\boldsymbol{\mu}^{(i)} = [\mu_1^{(i)}, \mu_2^{(i)}, \ldots, \mu_{m'}^{(i)}]^T$ and the mean of all the classes as $\boldsymbol{\mu} = [\mu_1, \mu_2, \ldots, \mu_{m'}]^T$. Then, the within-class scatter matrix $\mathbf{S}_W^{\mathrm{DNMF}}$ and the between-class scatter matrix $\mathbf{S}_B^{\mathrm{DNMF}}$ are defined as

$$\mathbf{S}_W^{\mathrm{DNMF}} = \sum_{i=1}^{c} \sum_{q=1}^{n_i} \left(\mathbf{h}_q^{(i)} - \boldsymbol{\mu}^{(i)}\right)\left(\mathbf{h}_q^{(i)} - \boldsymbol{\mu}^{(i)}\right)^T, \tag{15}$$

$$\mathbf{S}_B^{\mathrm{DNMF}} = \sum_{i=1}^{c} n_i \left(\boldsymbol{\mu}^{(i)} - \boldsymbol{\mu}\right)\left(\boldsymbol{\mu}^{(i)} - \boldsymbol{\mu}\right)^T. \tag{16}$$

The solution is provided by the following minimization problem:

$$\min_{\mathbf{Z},\mathbf{H}} D_d(\mathbf{X} \parallel \mathbf{ZH}) \quad \text{under the constraints} \tag{17}$$

$$z_{s,t} \geq 0, \qquad h_{t,q} \geq 0, \qquad \sum_{\omega} z_{q,\omega} = 1, \quad \forall(s,t,q). \tag{18}$$

The resulting reduced-dimensionality vectors $\mathbf{h}$ form the discriminant facial image representation to be used in facial expression recognition. 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 (Chellappa, Wilson, & Sirohey, 1995). In Kotsia et al. (2007), DNMF has been used in combination with the Support Vector Machine (SVM) classifier (Burges, 1998; Chapelle, Haffner, & Vapnik, 1999) in facial image characterization problems.

## 3. Facial expression recognition procedure

### 3.1. Preprocessing

Initially, we preprocessed the images manually in order to have perfect alignment and the eyes in fixed pre-defined positions in the facial image ROI. To do so, we worked with the high resolution images and the eye coordinates were gathered in the initial images by two individuals. The initial distance between the eyes in high resolution was calculated and the images were down-scaled in an isotropic way, in order to have a 16-pixel distance between the two eyes. In the final step we cropped the images to the size of 40 × 30 pixels, producing a bounding box centered to the subject's face. These images are considered perfectly aligned and are referred as"centered" dataset in the rest of the paper. Image cropping was based on the eye region centers, due to their invariance to the various facial expressions. The position of other facial features (e.g., mouth, eye-brows) tend to shift during certain expressions. For example, in the case of surprise, the eyebrows shift upwards vertically, in comparison with the neutral expression. Thus, manual image cropping based on other facial features, apart from the eyes, could produce discriminant information on itself, leading to an overestimation of the performance of the classification. Furthermore, the detection and tracking accuracy of such features is limited, as we have found experimentally.


**Fig. 1.** The Cohn–Kanade Facial Expression Database. (a) neutral, (b) anger, (c) disgust, (d) fear, (e) happiness, (f) sadness, (g) surprise.

### 3.2. Training

Every preprocessed image was mapped from a $40 \times 30$ matrix to a 1200 dimensional vector. The training was performed by using one of the subspace learning techniques that have been presented in Section 2. According to the division of the facial feature extraction stage, that has been presented in Pantic and Rothkrantz (2003), in our approach, the features are extracted in an automatic way from still images. Thus, temporal information is not used. These features are holistic and view-based.

As far as the presented methods of PCA + LDA and PCA + CDA are concerned, PCA was used for maintaining the 95% of the covariance matrix energy. When PCA is not mentioned, it is implicitly considered that a 100% of the total variance was retained. Thus, in this case the zero eigenvalues of the covariance matrix of the data were rejected. PCA was used in order to overcome the undersampling problem, where the number of the samples is less than the dimensionality of the data. The cases of maintaining other percentages of the covariance matrix energy were tested as well, without leading to better results.

On the one hand, LDA reduced data vector dimensionality to 6 dimensions. On the other hand, CDA, as has been discussed in Section 2, was capable of reducing the data vector dimensionality to more than 6 dimensions. In the following analysis, the results for the number of dimensions that gave the best results are presented. Regarding DNMF, the dimensionality of the feature vector was reduced from 1200 to 120 dimensions.

### 3.3. Testing

All the above methods, aim at projecting the initial high-dimensional data samples to a feature space with low dimensionality. In that new space, the data samples are likely to be classified in a more efficient way. Thus, a tested sample was projected to the reduced feature space, using one of the above-mentioned approaches and a classifier was then applied on the projected sample. In our study, we have used the well-known classifiers, KNN, NC and SVMs in the literature. We have also used the NCC (Maronidis, Tefas, & Pitas, 2010), which accompanies the CDA algorithm. In NCC, the cluster centroids are calculated and the testing sample is assigned to the class in which the nearest cluster centroid belongs to.

A support vector machine tries to calculate the optimal hyperplane or a 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 SVM used for our experiments was proposed in Tefas, Kotropoulos, and Pitas (2001). It employs a modified method to calculate the maximum functional margin, inspired by the Fisher's discriminant ratio. The SVM is 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 testing sample is found. In our study, SVM realized the classification on the feature vectors extracted by DNMF, using an RBF kernel.

**Table 1**
Cross validation versus inter-database performance rates (%) in BU, JAFFE and Kanade.

| Classifier | DR method | JK-B | BU | BK-J | JAFFE | BJ-K | KANADE |
|---|---|---|---|---|---|---|---|
| NC | PCA + LDA | 29.9 | 63.3 | 23.0 | 54.5 | 31.0 | 67.0 |
| NCC | PCA + CDA | 31.9 | 63.4 | 28.6 | 57.0 | 33.2 | 69.6 |
| SVM | DNMF | 29.0 | 55.4 | 31.0 | 41.6 | 31.0 | 56.4 |

## 4. Proposed approach and results

A series of experiments on facial expression recognition has been conducted using the frontal images of three standard databases. The databases, that have been used for the experiments are described below.

The Cohn–Kanade AU-Coded Facial Expression Database (Kanade, Cohn, & Tian, 2000) affords a test bed for research in automatic facial image analysis and is available for use by the research community. Image data consist of approximately 500 image sequences from 100 subjects in 7 different universal emotional states (anger, disgust, fear, happiness, sadness, surprise and neutral state). Subjects range in age from 18 to 30 years. Sixty-five percent were female; 15% were African-American and three percent Asian or Latino ones. One of the cameras was located directly in front of the subject, and the other was positioned 30° to the subject's right. Sample images of the database are presented in Fig. 1.

BU database (Yin, Wei, Sun, Wang, & Rosato, 2006) contains 100 subjects. Each subject performed the seven aforementioned facial expressions in front of a 3D face scanner. With the exception of the neutral expression, each of the six basic expressions (happiness, disgust, fear, angry, surprise and sadness) include four levels of intensity. Therefore, there are 25 3D expression models for each subject, resulting in a total of 2500 3D facial expression models. Each expression shape model is associated to a corresponding facial texture image captured at two views (about $+45$ and $-45°$ versus the frontal view). As a result, the database consists of 2500 two-view texture images and geometric shape models.

The JAFFE database (Lyons, Akamatsu, Kamachi, & Gyoba, 0000) contains 213 images of the 7 aforementioned facial expressions, posed by 10 Japanese female models. Each image has been rated on these emotion labels by 60 Japanese subjects. Typical examples are illustrated in Fig. 2. In all the above databases, the used images are grayscale and have been rescaled to $40 \times 30$ pixels.

### 4.1. Inter-database experiments

In this section the behavior of subspace learning techniques in person independent experiments in both the case of cross validation within the same database and inter-database validation is examined. A number of facial expression recognition algorithms are tested using the three aforementioned facial expression recognition databases (BU, JAFFE and Kanade). Three typical examples are depicted in Table 1. The first and second columns depict the classifier and the subspace learning method used for dimensionality reduction (DR method), respectively. The best performing classifier for each dimensionality reduction method is


**Fig. 2.** The JAFFE Facial Expression Database. (a) neutral, (b) anger, (c) disgust, (d) fear, (e) happiness, (f) sadness, (g) surprise.

presented here. DNMF is always combined with SVM as originally proposed in Kotsia et al. (2007) for facial expression recognition. The following three pairs of columns present the intra-database cross-validation versus the inter-database performance rates. e.g., JK-B refers to training using all the samples from the JAFFE and Kanade databases and testing on the BU database. The column with label BU, refers to results using cross validation within the BU database.

The objective of this series of experiments is to highlight the significant drop in the performance, when inter-database experiments are performed. In Table 1, it can be observed that in the majority of the cases, the accuracy of the cross validation experiment is more than the double of the inter-database one. Furthermore, the superiority of CDA combined with NCC classifier against the other two DR methods combined with their respective best classifier is apparent. However, it is obvious that the overall performance rates are rather low. Moreover, to the best of our knowledge, it is the first time that inter-database experiments are performed and highlight the serious problems that come up when someone tries to develop generic person and database independent facial expression recognition algorithms. Indeed, in the Literature only intra-database results are reported whereas the performance in the inter-database case is very low for all the different subspace-based facial expression recognition algorithms. This can be attributed to the specific conditions under which, each database has been acquired. Even the fact that different types of cameras are used for each database has great impact to the results. The aim of this study is to highlight the problem and to propose simple solutions that can improve the performance of subspace learning methods for generic and person specific facial expression recognition. In the following sections, possible solutions to this problem are presented and the comparative results are referred.

### 4.2. Enriching the training set

When it comes to automatic real-world facial expression recognition applications, inaccuracies in the facial image ROI's size and position are expected. Therefore, either a systematic preprocessing involving ROI resizing is needed or alternatively training approaches can be used that robustify the facial expression recognition algorithm against inaccuracies in the test image size and position. In this section, the robustness of appearance-based, subspace learning techniques for facial expression recognition in geometrical transformations is explored. Also, a method for database enrichment is proposed. Although, it is common-knowledge that appearance based methods are sensitive to image registration errors, we have found no systematic experiment quantifying this sensitivity in the literature. After a series of experiments, a strong correlation between the performance and the image registration error has been observed. Even slight geometrical distortions in the facial image ROI could lead in great differences regarding the recognition performance. The mere investigation of the use of the optimal parameters of the facial ROI is inefficient, due to the inherent constraints that a real-world application imposes and an alternative approach is required. Thus, by forming the training set, using the initial database combined with a variety of geometrically transformed facial image ROIs, higher levels of robustness can be obtained improving the overall success rates at the same time. That is, a facial image database enrichment with translated, scaled and rotated images is proposed for confronting the low robustness of subspace techniques for facial expression recognition.

#### 4.2.1. The procedure

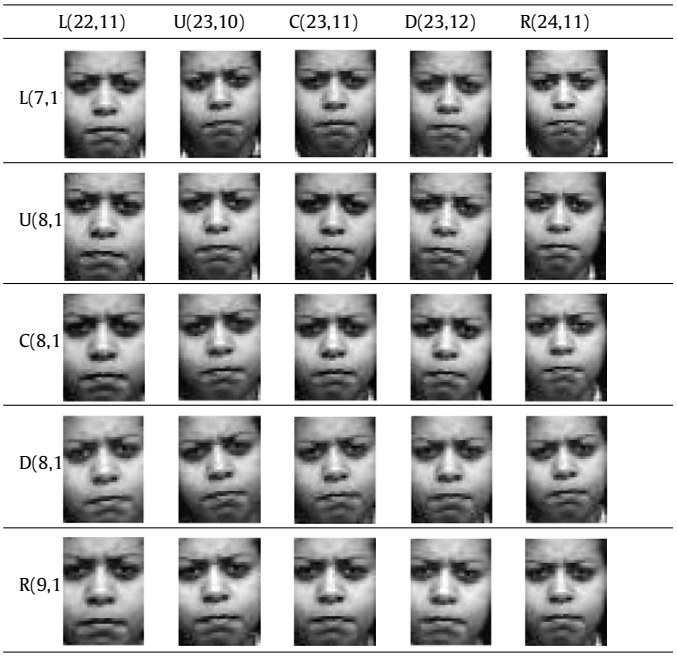
Under this perspective, we constructed two versions of enriched databases. For the first one called "enriched database", the "centered" (i.e., perfectly aligned) dataset was enriched with translated image versions. Specifically, translations to each of the four basic directions (left, right, up and down) by approximately 6% of the between-eyes distance were considered. Thus, this first type of enrichment resulted in 5 times larger database, compared to the initial one.

To construct the second dataset called "fully enriched" hereafter, four hypothetical types of errors regarding the eye detection were preconsidered. To be more specific, for each eye we considered that the eye detector fails to detect the correct position by ≈6% of the between-eyes distance in the high resolution images, in one of the basic four directions (left, right, up and down). The error ≈6% corresponds to one pixel error in the reduced dimensionality images used for subspace analysis and classification. Consequently, including the correct position as well, the eye detector could potentially produce 5 different outputs for a given eye position. That is, the detector either calculates the correct position or one of the four erroneous positions mentioned above. Thus, considering both the left and the right eye, such an eye detector could potentially produce 25 different outputs, given the ground truth for each eye. We then used each of these 25 possible positions as an input to the preprocessing algorithm described in Section 3.1. The result was 25 different versions of each image, consisted of the original perfectly aligned images and translated, rotated, and scaled image versions. To make this procedure clear, two typical examples are given below. First, using the erroneous positions by 6% left for both of the eyes, the preprocessing algorithm produces translated to the right image, since the eyes should be positioned symmetrically to the vertical axis. Second, using the erroneous positions by 6% left for the left eye, and 6% right for the right eye, an isotropically downscaled image by 12% is produced. Following this logic, each position pair results in a different geometrical transformation, when it is used by the preprocessing algorithm.

The set of fully enriched samples produced by one image of the dataset Cohn–Kanade is depicted in Table 2. On the first column and row, the direction, L(left), U(up), C(center), D(down), R(right) of shifting and the exact position in pixel-coordinates of the left and the right eye respectively are referred. For example, on the first column, D(8,12) means that the left eye is shifted (erroneously detected) by one pixel (6%) downwards and its exact coordinates in the image are 8 pixels from the left boundary and 12 pixels from the top boundary. The proposed sampling is a compromise between the resulting dataset size (25 times the initial size) and the possible geometrical transforms that can be used.

Then, we implemented a number of combinations of subspace image representations exploiting PCA, LDA, CDA and DNMF algorithms and NC, NCC, KNN and SVM classifiers, in order to

**Table 2**
Enriched training facial image samples resulting from one image of the Cohn–Kanade database.




examine their effectiveness in classifying the aforementioned facial expressions. For this purpose, we conducted a five-fold cross-validation. Regarding the PCA and LDA outputs, we used the Nearest Centroid algorithm, while, the Nearest Cluster Centroid and SVM methods were applied on the CDA and DNMF algorithm outputs, respectively.

### 4.2.2. Results

We conducted four series of experiments. In the first one, the centered images (CC) were used to form both the training and the testing set. This experiment corresponds to a hypothetical perfect alignment in a real-system. Thus, the performance reported using the "centered" dataset is the upper bound one would expect by using a perfect image alignment system. In the second one, the centered images were used for the training set, while the left-shifted images (LL) were used for the testing set, in order to examine the sensitivity of the recognition performance in displacements of the bounding box by ≈6%. In the third one, the training set was the "enriched" dataset formed from both the centered and shifted images, while the centered images alone constituted the testing set. Finally, in the fourth series of experiments, the "fully enriched" dataset formed the training set, while again the centered images were used for testing. The latter two approaches were conducted in order to explore the improvement of the performance of the several methods, when enriching the training set. The comparative results, for the Kanade, JAFFE and BU databases, are depicted in Tables 3–5, respectively.

In the first two columns of the tables the various utilized methods are given, both for reducing the dimensionality and for classifying the samples. KNN was used for $K = 1$ and $K = 3$. Here, the second case ($K = 3$) has been presented, since it gave better results. In the third column of these tables the success rates are presented, in the case of the centered images, for both the training and test set. The next column shows the performance when misplaced images are used for the test set. In the fifth column, the performance of the enriched database, exploiting merely the translated images, is depicted. Finally in the last column, the performance using the fully enriched database for training, where the 25 transformed versions of the original database were used, is depicted. The bold value in each row indicates the best performance of the corresponding method among the four approaches.

**Table 3**
Kanade 5-fold cross validation performance rates (%).

| Classifier | DR method | Centered | Misplaced | Enriched | Fully enriched |
|---|---|---|---|---|---|
| NC/NCC | PCA | 36.4 | 36.0 | 36.5 | **39.7** |
| | LDA | 62.5 | 55.5 | 72.4 | **74.9** |
| | PCA + LDA | 67.0 | 65.1 | 68.8 | **73.7** |
| | CDA | 66.0 | 56.5 | 68.1 | **69.6** |
| | PCA + CDA | **68.9** | 66.0 | 64.3 | 63.3 |
| 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** |
| | CDA | 64.8 | 57.7 | 70.6 | **70.9** |
| | PCA + CDA | **71.2** | 66.6 | 63.0 | 63.8 |
| SVM | DNMF | 56.4 | 49.4 | 67.6 | **69.2** |

**Table 4**
JAFFE 5-fold cross validation performance rates (%).

| Classifier | DR method | Centered | Misplaced | Enriched | Fully enriched |
|---|---|---|---|---|---|
| NC/NCC | 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 |
| | CDA | 48.1 | 44.9 | 59.5 | **67.3** |
| | PCA + CDA | 49.3 | 50.6 | 59.5 | **62.9** |
| 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** |
| | CDA | 53.1 | 48.8 | 58.0 | **68.8** |
| | PCA + CDA | 54.9 | 48.3 | 58.3 | **66.8** |
| SVM | DNMF | 41.6 | 34.6 | 57.5 | **63.9** |

**Table 5**
BU 5-fold cross validation performance rates (%).

| Classifier | DR method | Centered | Misplaced | Enriched | Fully enriched |
|---|---|---|---|---|---|
| NC/NCC | PCA | 34.6 | 34.0 | **34.9** | 34.3 |
| | LDA | 56.0 | 54.4 | 62.3 | **68.1** |
| | PCA + LDA | 63.3 | 62.3 | **64.9** | 63.3 |
| | CDA | 53.9 | 46.4 | 63.8 | **67.0** |
| | PCA + CDA | **64.3** | 58.0 | 61.6 | 60.7 |
| KNN | PCA | 33.1 | 33.0 | 32.7 | **37.3** |
| | LDA | 56.6 | 53.7 | 61.3 | **65.0** |
| | PCA + LDA | 60.4 | 60.0 | **62.1** | 59.7 |
| | CDA | 52.0 | 45.9 | 62.7 | **66.6** |
| | PCA + CDA | 60.8 | 55.4 | 59.5 | **62.1** |
| SVM | DNMF | 41.6 | 34.6 | 57.5 | **63.9** |

It can be, easily observed, that whatever method is used, even a slight divergence from the centered images (≈6% in the case of our experiments) causes, in certain cases, a severe drop in performance (up to 8.5%). That is, a small misplacement that can be attributed to small face detection localization errors in a real system will cause a much lower performance than the one that could be obtained with perfect face registration. 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 versions of the database enrichment (up to 15.9% for the enrichment with the translated images and 22.3% for the fully enriched version). The robustness when enriching the training set is systematically observed in our experiments. Additionally, it is observed that the more transformations are used the greater the improvement of the accuracy becomes.

Moreover, even with manual alignment the occurred registration errors lead to worse performance compared to the enriched training sets. This can be observed by comparing the performance between the centered and the enriched dataset in Tables 3–5.

**Fig. 3.** Frontal FER-AIIA images. From upper left to lower right: neutral, anger, disgust, fear, happiness, sadness, surprise.

Another observation can be made for the performance of PCA. That is, PCA tends to filter out outliers like misaligned images and thus generalizes better. However, when PCA is used as pre-processing before dimensionality reduction, in some cases causes performance loss. This can be attributed to the fact that PCA discards eigenvectors that correspond to smaller eigenvalues and these eigenvectors may contain discriminant information for the enriched and fully enriched case leading to bad modeling of the classes and subclasses of the data. Additionally, the enrichment may generate samples with variable geometric characteristics which are more probable to match with the test sample. Therefore, such an enrichment of the training dataset is expected to robustify the subspace facial expression recognition systems versus facial image ROI detection/localization and tracking errors. It is important to highlight that when the enriched and fully enriched datasets are used, a clear performance improvement occurs both in comparison with the misplaced and the centered datasets. Thus, this scheme is proposed not merely for compensating the registration errors that indeed occur in real world applications (Rentzeperis et al., 0000; Whitehill, Littlewort, Fasel, Bartlett, & Movellan, 2009), but also for improving the performance even when the optimal registration has been performed.

### 4.3. Person dependent training

In this section the use of person dependent training procedures, in order to obtain more effective facial expression learning, is explained. The motivation for these experiments was the fact that facial expressions tend to be person specific, as explained in Section 1 and in certain settings (e.g., assisted living) training and testing for facial expressions is ideally performed for few (sometimes only one) persons. The expectation is that by enriching the training set with images of the persons that lie in the test set, a more efficient learning is obtained.

The three aforementioned databases are not appropriate for studying person dependent/independent performance. The reason is that they do not provide facial images of the same persons at different sessions. For this purpose, the new database FER-AIIA for facial expression recognition has been created by the AIIA laboratory, Aristotle University of Thessaloniki. It contains 600 videos and 1200 2D luminance images captured using a Logitech C-200 camera. Cameras of this type can typically be used in low-budget human-centered interfaces. The series of the facial images depict the seven universally recognized facial expressions (angry, disgust, fear, happiness, sadness, surprise and neutral state), posed by five subjects (2 females and 3 males) in 4 distinct days (sessions). Each session consists of five recordings. Consequently, certain factor variations, as the illumination variation from day to day are taken into account. However, in every case the illumination was selected to be within a sensible range regarding the everyday life in a normal house. Both physical (sun) and artificial light (conventional lamps) were used. Thus, the database is consistent with everyday life conditions, as it is the case for real-world applications. The identity of the depicted person is also provided for every image. Some typical samples of this database are illustrated in Fig. 3. The camera's output was firstly grayscaled and rescaled to 40 × 30 pixel images as a result of the pre-processing step.

**Table 6**
FER-AIIA Leave-One-Person-Out performance rates (%).

| DR method | NC/NCC | 1-NN | 3-NN |
|---|---|---|---|
| PCA | 27.3 | 44.8 | 45.7 |
| LDA | 54.6 | 54.9 | 54.8 |
| CDA | 55.0 | 57.3 | 53.5 |
| PCA(95%) + LDA | 47.6 | 57.2 | 57.5 |
| PCA(95%) + CDA | 54.8 | 60.5 | **60.9** |
| PCA(90%) + LDA | 43.6 | 46.0 | 48.0 |
| PCA(90%) + CDA | 53.3 | 56.8 | 55.3 |

Then, first, we conducted a series of Leave-One-Person-Out Experiments on the FER-AIIA database. That is, all the images of a specific person constituted the test set, while the rest images formed the training set. These experiments were performed for every person providing a person specific recognition accuracy and the mean value of these accuracies across persons was calculated. The results are presented in Table 6. The first column depicts the utilized pattern recognition method. The following three columns contain the recognition rates for the Nearest Centroid, the 1-Nearest Neighbor and the 3-Nearest Neighbor classifiers, respectively. As expected, the recognition rates are similar to those that we have obtained from the intra-database experiments, shown in Table 1, on the other three datasets (BU, JAFFE and Kanade). Specifically, the maximum recognition rate is 60.9% achieved by the PCA(95%) + CDA method using the 3-NN classifier.

Second, a series of Leave-One-Session-Out Experiments were performed. In this case, the training was performed for all the subjects of the database for images recorded on 3 out of 4 available sessions and the data of the session left out were used for the testing. That is, a 4-fold cross validation approach was followed. The results are presented in Table 7. The maximum performance rate is now 95.6%, which is much higher than 60.9% reported for the Leave-One-Person-Out experiments. It has been obtained by the CDA method using the 3-NN classifier. This superior performance can be explained by the fact that the system has been trained to recognize the expressions of the subjects used in testing. Of course, it should be stressed that in none of the experiments the same image was used in both the training and testing set.

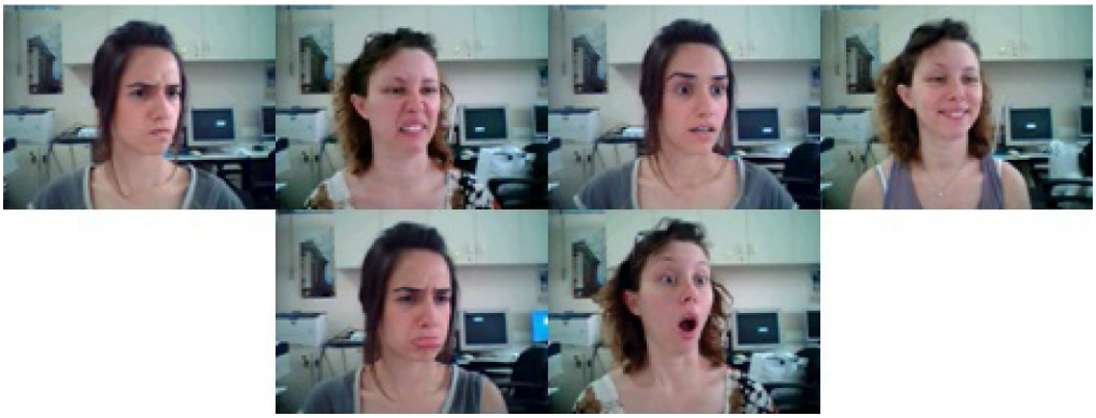


**Fig. 4.** 20° rotated FER-AIIA images. From upper left to lower right: anger, disgust, fear, happiness, sadness, surprise.

**Table 7**
FER-AIIA Leave-One-Session-Out performance rates (%).

| DR method | NC/NCC | 1-NN | 3-NN |
|---|---|---|---|
| PCA | 31.7 | 71.8 | 74.5 |
| LDA | 95.2 | 94.8 | 94.9 |
| CDA | 95.5 | 95.0 | **95.6** |
| PCA(95%) + LDA | 86.2 | 89.3 | 91.4 |
| PCA(95%) + CDA | 91.7 | 91.4 | 92.3 |
| PCA(90%) + LDA | 68.5 | 77.6 | 78.8 |
| PCA(90%) + CDA | 77.6 | 84.1 | 84.6 |

**Table 8**
FER-AIIA person specific accuracy rates (%).

| Approach | NC | 1-NN | 3-NN |
|---|---|---|---|
| PCA | 90.4 | 80.8 | 86.7 |
| LDA | **97.5** | 97.1 | 97.1 |
| CDA | 96.7 | 95.4 | 96.7 |
| PCA(95%) + LDA | 93.3 | 95.0 | 95.0 |
| PCA(95%) + CDA | 94.2 | 94.2 | 93.8 |
| PCA(90%) + LDA | 92.5 | 95.4 | 95.8 |
| PCA(90%) + CDA | 93.3 | 92.5 | 92.5 |

Third, a set of experiments was conducted in order to explore the combined improvement in the performance when both person dependent training is performed and fully enriched training dataset is used. The results show that in the vast variety of the cases the performance rate increase is even higher. To be more specific, for these types of experiments, the improvement was found in the range of 1%–5%.

Additionally, person-specific experiments were performed to simulate the case of a learning system that has been exclusively trained and tested for each subject independently. Table 8 depicts the accuracy rates for this category of experiments. By comparing the results of Tables 6 and 8, we can see that training and testing on data of the same person produces far superior results (accuracy in the range of 80.8%–97.5%) than when we train on some persons and test on others (accuracy in the range of 27.3%–60.9%). This performance drops, rather mildly, as can be seen in Table 7 (accuracy in the range of 31.7%–95.6%), when other person images are included in the training set.

Finally, experiments for defining the system's operational limits, due to face rotations, were conducted. Every person was asked to perform each facial expression with her/his head turned by 10 and 20° to the right, left, upwards or downwards with respect to the camera position. Thus, four videos for every expression were recorded. Experiments for facial expression recognition were performed using the subspace learning method trained using the first and the last frame from each video sequence included in the FER-AIIA LAB database. The average recognition rate has dropped by approximately 3% for the 10° experiments, remaining, thus, quite close to the non-rotation ones. On the contrary, the achieved expression recognition rate dropped radically in average by 23% for the 20° experiments. Fig. 4 shows characteristic instances from this dataset. This dramatic reduction in the recognition accuracy rate is realistic if we consider that the 20° head rotation has a severe impact regarding the visible facial features, since as we have observed in many cases the eye detector could not produce a correct localization result.

Summarizing the results of this section, it can be noted that when the facial expression recognition system is meant to be used for a specific person, very high performance can be achieved using person-dependent training. This usecase makes sense, e.g., in cognitive robotics for assisted living, where the person that interacts with the robot is known or in film/games postproduction, where the film actor identities are known in advance.

## 5. Conclusions

Facial expressions consist an integral part of human non-verbal communication. Subspace learning methods have become a frequently used tool to perform facial expression recognition. However, in this paper, the great sensitivity of these kinds of algorithms to geometrical distortion of the images, even for the case of one pixel, has been highlighted. Real-world applications carry an inherent difficulty regarding the precise face and facial feature localization, resulting in inaccurate image registration. The systematic enrichment of a database with geometrically transformed (translated, scaled and rotated) images, which has been proposed in this paper, has shown to give significant improvement in the recognition performance in the majority of the cases. Moreover, it has been shown that facial expression recognition has rather low performance in its generic form, i.e., when no training images are available for the test person. Person dependent training has also been proposed for certain applications, that involve a single user (e.g., assisted living). The experiments have shown that a major improvement can be achieved when using subspace learning for facial expression recognition combined with person-dependent training.

## Acknowledgments

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 248434 (MOBISERV).

## References

Azran, A., & Ghahramani, Z. Spectral methods for automatic multiscale data clustering. In: CVPR (pp. I: 190–197).

Belhumeur, P., Hespanha, J., & Kriegman, D. (1997). Eigenfaces vs. fisherfaces: recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19*, 711–720.

Bruce, V. (1993). What the human face tells the human mind: some challenges for the robot-human interface. *Advanced Robotics*, *8*, 341–355.

Burges, C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*,.

Chapelle, O., Haffner, P., & Vapnik, V. (1999). Support vector machines for histogram-based image classification. *IEEE Transactions on Neural Networks*, *10*, 1055–1064.

Chellappa, R., Wilson, C. L., & Sirohey, S. (1995). Human and machine recognition of faces: a survey. *Proceedings of the IEEE*, *83*, 705–741.

Chen, X. W., & Huang, T. S. (2003). Facial expression recognition: a clustering-based approach. *Pattern Recognition Letters*, *24*, 1295–1302.

Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, *17*, 124–129.

Goudelis, G., Zafeiriou, S., Tefas, A., & Pitas, I. (2007). Class-specific kernel-discriminant analysis for face verification. *IEEE Transactions on Information Forensics and Security*, *2*, 570–587.

Hara, F., & Kobayashi, H. (1996). State-of-the art in component technology for an animated face robot-its component technology development for interactive communication with humans. *Advanced Robotics*, *11*, 585–604.

Jolliffe, I. (1986). *Principal component analysis*. Springer Verlag.

Kanade, T., Cohn, J.F., & Tian, Y. (2000). Comprehensive database for facial expression analysis. In *Proceedings of the fourth IEEE international conference on automatic face and gesture recognition. Grenoble, France* (pp. 46–53).

Kotsia, I., Zafeiriou, S., & Pitas, I. (2007). A novel discriminant non-negative matrix factorization algorithm with applications to facial image characterization problems. *IEEE Transactions on Information Forensics and Security*, *2*, 588–595.

Kyperountas, M., Tefas, A., & Pitas, I. (2010). Salient feature and reliable classifier selection for facial expression classification. *Pattern Recognition*, *43*, 972–986.

Kyperountas, M., Tefas, A., & Pitas, I. (2008). Dynamic training using multistage clustering for face recognition. *Pattern Recognition*, *41*, 894–905.

Kyperountas, M., Tefas, A., & Pitas, I. (2007). Weighted piecewise lda for solving the small sample size problem in face verification. *IEEE Transactions on Neural Networks*, *18*, 506–519.

Lee, D., & Seung, H. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*, 788–791.

Lee, D., & Seung, H. (2000). Algorithms for non-negative matrix factorization. In *NIPS* (pp. 556–562).

Leibe, B., & Schiele, B. (2003). Analyzing appearance and contour based methods for object categorization.. In *IEEE computer vision and pattern recognition (CVPR), Vol. 2* (pp. 409–415). IEEE Computer Society.

Lyons, M., Akamatsu, S., Kamachi, M., & Gyoba, J. Coding facial expressions with gabor wavelets (pp. 200–205).

Maronidis, A., Tefas, A., & Pitas, I. (2010). Frontal view recognition using spectral clustering and subspace learning methods. In *Lecture notes in computer science*: *Vol. 6352. International conference on artificial neural networks* (pp. 460–469). Springer.

Mehrabian, A. (1968). Communication without words. *Psychology Today*, *2*, 53–56.

Nani, M., Caleb-Solly, P., Dogramadgi, S., Fear, C., & van den Heuvel, H. MOBISERV: an integrated intelligent home environment for the provision of health, nutrition and mobility services to the elderly. In *4th companion Robotics workshop. Brussels, Belgium*.

Pantic, M., & Rothkrantz, L. J. M. (2000). Automatic analysis of facial expressions: the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, *22*, 1424–1445.

Pantic, M., & Rothkrantz, L. J. M. (2003). Toward an affect-sensitive multimodal human–computer interaction. *Proceedings of the IEEE*, *91*, 1370–1390.

Rentzeperis, E., Stergiou, A., Pnevmatikakis, A., & Polymenakos, L. Impact of face registration errors on recognition. In *3rd IFIP conference on artificial intelligence applications and innovations AIAI. Athens, Greece* (pp. 187–194).

Takeuchi, A., & Nagao, K. (1993). Communicative facial displays as a new conversational modality. In S. Ashlund, K. Mullet, A. Henderson, E. Hollnagel, & T. N. White (Eds.), *INTERCHI* (pp. 187–193). ACM.

Tefas, A., Kotropoulos, C., & Pitas, I. (2001). Using support vector machines to enhance the performancee of elastic graph matching for frontal face authentication. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*, 735–746.

Theodoridis, S., & Koutroumbas, K. (2006). *Pattern recognition* (3rd ed.) Academic Press.

von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, *17*, 395–416.

Whitehill, J., Littlewort, G., Fasel, I., Bartlett, M., & Movellan, J. (2009). Toward practical smile detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*, 2106–2111.

Yin, L., Wei, X., Sun, Y., Wang, J., & Rosato, M. J. (2006). A 3D facial expression database for facial behavior research. In *FGR'06: proceedings of the 7th international conference on automatic face and gesture recognition* (pp. 211–216). Washington, DC, USA: IEEE Computer Society.

Zafeiriou, S., Tefas, A., Buciu, I., & Pitas, I. (2006). Exploiting discriminant information in non negative matrix factorization with application to frontal face verification. *IEEE Transactions on Neural Networks*, *17*, 683–695.

Zeng, Z., Pantic, M., Roisman, G. I., & Huang, T. S. (2009). A survey of affect recognition methods: audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, *31*, 39–58.