Comparison of ICA approaches for facial expression recognition

I. Buciu [1,2] C. Kotropoulos [1] I. Pitas [1]

[1] Department of Informatics, Aristotle University of Thessaloniki

GR-541 24, Thessaloniki, Box 451, Greece

costas,pitas@aiia.csd.auth.gr

[2] Electronics Department

Faculty of Electrical Engineering and Information Technology

University of Oradea 410087, Universitatii 1, Romania

ibuciu@uoradea.ro

Abstract

Independent Component Analysis (ICA) and Gabor wavelets extract the most discriminating features for facial action unit classification by employing either a cosine similarity measure (CSM) classifier or support vector machines (SVMs). So far, only the ICA approach, which is based on the InfoMax principle, has been tested for facial expression recognition. In this paper, in addition to the InfoMax approach, another five ICA approaches extract features from two facial expression databases. In particular, the Extended InfoMax ICA, the undercomplete ICA, and the nonlinear kernel-ICA approaches are exploited for facial expression representation for the first time. When applied to images, ICA treats the images as being mixtures of independent sources and decomposes them into an independent basis and the corresponding mixture coefficients. Two architectures for representing the images can be employed yielding either independent and sparse basis images or independent and sparse distributions of image representation coefficients. After feature extraction, facial expression classification is performed with the help of either a CSM classifier or an SVM classifier. A detailed comparative study is made with respect to the accuracy offered by each classifier. The correlation between the accuracy and the mutual information of independent components or the kurtosis is evaluated. Statistically significant correlations between the aforementioned quantities are identified. Several issues are addressed in the paper: (i) whether features having super- and sub-Gaussian distribution facilitate facial expression classification; (ii) whether a nonlinear mixture of independent sources improves the classification accuracy; and (iii) whether an increased "amount" of sparseness yields more accurate facial expression recognition. In addition, performance enhancements by employing leave-one-set of expressions-out and subspace selection are studied. Statistically significant differences in accuracy between classifiers using several feature extraction methods are also indicated.

Keywords

Independent component analysis (ICA), super-Gaussian distribution, sub-Gaussian distribution, nonlinear mixtures of independent sources, cosine similarity measure classifier, support vector machine classifier, facial expression recognition, mutual information, kurtosis, correlation, statistical significance.

I. Introduction

Human facial expression analysis has captured an increasing attention from psychologists, anthropologists, and computer scientists [1]. Surveys on automatic facial expression analysis can be found in [2,3,4]. Generally speaking, facial expression recognition methods can be classified into *appearance-based* methods and *geometry-based* ones. In the first category, fiducial points of the face are selected either manually [5] or automatically [6]. The facial images are convolved with Gabor filters and the extracted Gabor filter responses at

the fiducial points form vectors that are further used for facial expression classification. Alternatively, Gabor filters can be applied to the entire facial image instead of specific facial regions. Regarding the geometry-based methods, the coordinates of the fiducial points form a feature vector that represents facial geometry. Although the appearance-based methods yield a reasonable facial expression recognition accuracy, the highest recognition rate has been obtained when both the responses of Gabor wavelets and the coordinates of fiducial points are combined [5,7,8]. The analysis can be performed either on still images [5] or image sequences, where temporal information is considered [9]. Gabor and Independent Component Analysis (ICA) representations were used for the recognition of 6 single upper facial action units (AUs) and 6 lower face AUs in [10]. The AUs correspond roughly to the movement of the individual 44 facial muscles. The best recognition rates were achieved by both Gabor wavelets and ICA representations [10]. The local properties of ICA representation were found to be important for identity recognition [11]. Identity and facial expression recognition performance were also investigated by directly comparing ICA versus Principal Component Analysis (PCA) in [12], where it was found that ICA outperformed PCA. On the contrary, insignificant performance differences between ICA and PCA were reported on the same database in [13]. Guo and Dyer addressed facial expression classification, when a small number of training samples was only available [14]. In particular, a new linear programming-based technique was developed for both feature extraction and classification and a pairwise framework for feature selection was designed instead of considering all classes simultaneously. Gabor filters were used to extract facial features and large margin classifiers such as support vector machines (SVMs) and AdaBoost were employed for facial expression classification. Susskind et al. studied the nature of emotional space [1], where evidence is presented justifying that emotion categories are not entirely discrete and independent, but they vary along underlying continuous dimensions.

The facial expression recognition accuracy reported by Donato et al. in [10] was obtained by applying the InfoMax approach [15]. Taking this ICA approach as baseline for feature extraction, five additional ICA approaches, namely, the extended-InfoMax [16], the Joint Approximate Diagonalization of Eigen-matrices (JADE) [17], the fastICA [18] ,

the undercomplete ICA (uICA) [19], and the nonlinear kernel-ICA [20] are investigated in this paper. By employing the aforementioned ICA approaches, we extend Donato's work. Additional issues are addressed, such as whether sub-Gaussian facial feature extraction through the extended-InfoMax facilitates facial expression classification; whether a nonlinear mixture of independent components through the nonlinear kernel-ICA influences the classification accuracy. Moreover, we assess the effect of sparseness on the classification accuracy. It is worth mentioning that the results reported in [10] refer to the recognition of facial actions derived from the Facial Action Coding System (FACS), while, in this paper, we are interested in the classification of facial expressions that are combinations of facial action units. Each ICA approach has its own advantages over the others. For example, the original InfoMax approach [15] is *not* able to recover signals having a sub-Gaussian distribution. To alleviate this deficiency, the Extended InfoMax approach has been developed that can extract the sub-Gaussian sources [16]. The strength of each ICA approach is investigated with respect to the facial expression classification accuracy, when the extracted features feed either a CSM classifier or an SVM.

The rest of the paper is organized as follows. ICA is viewed as a feature extraction method in Section II. Section III briefly summarizes each ICA approach investigated in the paper and the rationale for its use in feature extraction within the framework of facial expression recognition. Section IV describes two ICA architectures for feature extraction. The facial expression image databases used in the experiments are introduced in Section V. Section VI presents two classifiers applied to the feature vectors obtained by ICA for facial expression classification. Experimental results are reported in Section VII, where issues such as independent basis images or independent coefficients, their mutual information, their sparseness, and how these characteristics are correlated with the classification accuracy are addressed. In addition, performance enhancements by employing leave-one-set of expressions-out and subspace selection are studied. Statistically significant differences in accuracy between classifiers using several feature extraction methods are also indicated. Finally, conclusions are drawn in Section VIII.

II. ICA as a feature extraction method

In pattern classification, feature extraction represents an important processing step, because one looks for features that incorporate sufficient class information and possess reliable discriminating power in order to obtain a satisfactory classification accuracy. Frequently, dimensionality reduction is performed as a first step, aiming at removing any redundant information, whilst preserving the information which contributes more to the classification accuracy. One of the most popular techniques for dimensionality reduction is PCA. This technique is based on second-order statistics of the data and performs dimensionality reduction by retaining components that correspond to the largest eigenvalues of the covariance matrix, while discarding components that have insignificant contribution to the trace of the covariance matrix. In principle, PCA yields uncorrelated components. When the data have a Gaussian distribution, the uncorrelated components are independent as well. However, if the data are mixtures of non-Gaussian components, PCA fails to extract components having a non-Gaussian distribution. On the contrary, ICA takes into account the higher-order statistics of the data in an attempt to recover non-Gaussian components. For completeness, we mention that under certain conditions, non-Gaussian components could be recovered by applying *Exploratory Projection Pursuit* (EPP) [21] as well.

From a statistical point of view, the least interesting structure is the Gaussian one. In one dimension, two moments, the mean and the variance, completely define the probability density function (pdf). Moreover, the Gaussian distribution has the highest entropy among all distributions with a given covariance matrix [22]. Taking the Gaussian distribution as a reference, any quantity that measures the level of 'interestingness' of the data, is a quantity that measures the non-Gaussian structure of the data. A principled measure of nongaussianity is the negentropy. The negentropy of a standardized random variable (i.e. one that has zero-mean and unit variance) can be approximated by the third-order moment and the fourth-order cumulant (i.e. the kurtosis) in a computationally simple manner. Therefore, we need moments and cumulants of order higher than 2 to capture the non-Gaussian structure of data [22]. Seeking non-Gaussian components is related to looking for statistical independence [22]. A measure of non-Gaussianity of a random

variable (RV) s is its normalized kurtosis estimated as:

$$\mathrm{kurt}(\mathsf{s}) = \frac{\sum_i (s_i - \overline{s})^4}{\left[\sum_i (s_i - \overline{s})^2\right]^2} - 3 \tag{1}$$

where s_i are observations of s and $\overline{s}$ denotes the sample mean of s. The normalized kurtosis of a Gaussian RV is zero. Super-Gaussian RVs have a positive kurtosis. A typical super-Gaussian RV is the Laplacian one. Sub-Gaussian RVs have a negative kurtosis with a typical example being a uniform RV in the interval $[-\alpha, \alpha] \in \mathbb{R}$.

ICA can be formulated by considering the following statistical model:

$$\mathbf{x} = \mathbf{A}\,\mathbf{s} \tag{2}$$

where $\mathbf{s} = (s_1, s_2, \ldots, s_n)^T$ is a latent random vector with independent components that are combined via a $p \times n$ mixing matrix $\mathbf{A}$ to form a zero-mean observation vector $\mathbf{x} = (x_1, x_2, \ldots, x_p)^T$. ICA estimates a demixing matrix $\mathbf{W}$ of dimensions $n \times p$ that will recover the original components of $\mathbf{s}$ as:

$$\mathbf{u} = \mathbf{W}\,\mathbf{x} = \mathbf{W}\,\mathbf{A}\,\mathbf{s} \tag{3}$$

where $\mathbf{u} = [u_1, u_2, \ldots, u_i, \ldots, u_n]$ is an estimate of $\mathbf{s}$. Given a batch of m observation data $\mathbf{x}_j$, $j = 1, \ldots, m$ we can form $\mathbf{X}$ whose columns are $\mathbf{x}_j$. Then (3) becomes:

$$\mathbf{U} = \mathbf{W}\,\mathbf{X} = \mathbf{W}\,\mathbf{A}\,\mathbf{S} \tag{4}$$

where $\mathbf{X}$ and $\mathbf{U}$ are $p \times m$ and $n \times m$ matrices, respectively. Usually, we call the columns of $\mathbf{U}$ (and implicitly the columns of $\mathbf{S}$) *independent sources*. The columns of $\mathbf{X}$ are measurements from a number of sensors that capture the sources. Usually, the number of observed components is equal to the number of independent components ($p = n$). There are ICA methods that cope with cases $p < n$ or $p > n$, called *overcomplete* or *undercomplete* ICA, respectively. Basically, the ICA algorithms attempt to obtain an estimate of $\mathbf{W}$ by using an objective (contrast) function that must be maximized or minimized, depending on the formulation.

III. ICA approaches

Let $p = n$. The *InfoMax* approach performs ICA based on the information maximization approach proposed by Bell and Sejnowski [15]. This approach relies on the maximization

of the entropy of the joint distribution $f(\mathbf{u})$. The demixing matrix $\mathbf{W}$ is updated through an iterative process. At iteration $k+1$, $\mathbf{W}$ is updated according to:

$$\mathbf{W}_{k+1} = \mathbf{W}_k + \eta[\mathbf{I} + (\mathbf{1} - 2\mathbf{z}_k)\,\mathbf{u}_k^T]\,\mathbf{W}_k, \tag{5}$$

where η is the learning rate controlling the convergence speed of (5), $\mathbf{1}$ is a $n \times 1$ vector of ones, $\mathbf{I}$ is the $n \times n$ identity matrix, and $\mathbf{z}$ is a $n \times 1$ vector having elements:

$$z_i = g(u_i) \qquad i = 1, \ldots, n \tag{6}$$

with $g(.)$ being a component-wise nonlinearity applied to all elements of the demixer output $\mathbf{u}$, at each iteration k. The form of the nonlinearity must be chosen to match the cumulative distribution function of the input. In the InfoMax algorithm [15], this non-linearity is approximated by the logistic transfer function:

$$g(u_i) = 1/(1 + e^{-u_i}) \qquad i = 1, \ldots, n. \tag{7}$$

The just described approximation works well when it comes to recover super-Gaussian components, but fails to extract the components having a sub-Gaussian distribution if such components exist in the mixture of non-Gaussians. To remedy this drawback, Lee et al. have extended the InfoMax approach to the *Extended InfoMax* approach by employing a new learning rule that is able to separate both sub- and super-Gaussian distributions [16]. The learning rule, that is able to switch between these distributions, iteratively updates the demixing matrix as follows:

$$\mathbf{W}_{k+1} = \mathbf{W}_k + \eta[\mathbf{I} - \mathbf{\Xi}\ \tanh(\mathbf{u}_k)\mathbf{u}_k^T - \mathbf{u}_k\mathbf{u}_k^T]\mathbf{W}_k, \tag{8}$$

where $\mathbf{\Xi}$ is an $n \times n$ diagonal matrix whose ii-th element, ξ_{ii}, takes the value 1 for a super-Gaussian source and the value -1 for a sub-Gaussian one, and tanh() denotes the hyperbolic tangent function that is applied to the elements of $\mathbf{u}_k$ in a component-wise fashion. The adaptation of ξ_{ii} is given by:

$$\xi_{ii} = \mathrm{sign}(E\{\mathrm{sech}^2(u_{ki})\}E\{u_{ki}^2\} - E\{[\tanh(u_{ki})]u_{ki}\}), \tag{9}$$

where $i = 1, \ldots, n$, u_{ki} is the i-th element of $\mathbf{u}_k$, and sign() and sech() denote the sign and hyperbolic secant functions, respectively.

Another approach for separating sources, the so called *Joint Approximate Diagonalization of Eigen-matrices* (JADE) was proposed by Cardoso and Souloumiac [17]. The main advantage of JADE is the fact that it does not employ any learning step. Its drawback is the relatively small number of components that can be extracted making it inadequate for a large number of mixture components.

The fourth approach is *fastICA* developed by Hyvarinen [18], which maximizes negentropy. The fastICA steps for estimating several independent components with deflationary orthogonalization are the following [22]:

1. Center the data to zero their mean.
2. Choose the number n of independent components to be estimated. Set $p = 1$. Whiten the data to obtain $\mathbf{z} = \mathbf{V}\mathbf{x} = \mathbf{V}\mathbf{A}\mathbf{s}$.
3. Choose randomly an initial vector of unit norm for $\mathbf{w}_p$.
4. Let $\widehat{\mathbf{w}}_{p,k+1} = E\{\mathbf{z}_k g(\mathbf{w}_{p,k}^T \mathbf{z}_k)\} - E\{g^{'}(\mathbf{w}_{p,k}^T \mathbf{z}_k)\}\mathbf{w}_{p,k}$, where $g(\xi) = (1/a)\log(\cosh(a\xi))$ is the contrast function and its derivative is given by $g^{'}(\xi) = \tanh(a\xi)$.
5. Do the following orthogonalization $\widetilde{\mathbf{w}}_{p,k+1} = \widehat{\mathbf{w}}_{p,k+1} - \sum_{j=1}^{p-1}(\widehat{\mathbf{w}}_{p,k+1}^T \mathbf{w}_j)\mathbf{w}_j$.
6. Let $\mathbf{w}_{p,k+1} = \frac{\widetilde{\mathbf{w}}_{p,k+1}}{\|\widetilde{\mathbf{w}}_{p,k+1}\|}$.
7. If $\mathbf{w}_p$ has not converged, go back to step 4.
8. Set $p \longleftarrow p + 1$. If $p \leq n$, go back to step 3.

A major advantage of fastICA is its speed, making it even 100 times faster than the previously described approaches.

For all ICA approaches described so far, it has been assumed that the number of components equals the number of sensors. If the number of sources is very large, the application of ICA is limited by memory constraints. Therefore, the preprocessing PCA step is not only intended to decorrelate the data, but also to lower their dimension. By keeping only $l < p$ appropriately chosen dimensions the demixing matrix $\mathbf{W}$ becomes of size $l \times l$. When discarding the $(p - l)$ dimensional subspace with the smallest variance, there is a risk to throw away the independent components (ICs) that might be contained in this subspace, since there is no guarantee that ICs exist only in the l dimensional subspace defined by the principal components (PCs) with the largest eigenvalues. For instance, an IC with a very small variance was found to be associated with the form of the "on-off" experimental

protocol when analyzing fMRI data [23]. To address the weakness of the previously described ICA approaches, Stone and Porrill have developed the *undercomplete Independent Component Analysis* (uICA) for preserving the information that might be lost during PCA and established the following contrast function for maximizing the entropy [19]:

$$h(\mathbf{W}) = \frac{1}{2}\log|\mathbf{W}\mathbf{D}_x\mathbf{W}^T| + E\left\{\sum_{i=1}^{n}\log\left(\frac{\partial z_i}{\partial u_i}\right)\right\}, \tag{10}$$

allowing to have a non-square $n \times p$ demixing matrix without applying PCA for data dimensionality reduction. $\mathbf{D}_x$ is the sample covariance matrix of the input data $\mathbf{x}$. If $z_i = g(u_i) = \tanh(u_i)$, (10) can be maximized using, for example, the Broyden-Fletcher-Goldfarb-Shanno (BFGS) quasi-Newton method. The derivative of (10) is given by:

$$\frac{\partial h}{\partial \mathbf{W}} = \mathbf{W}^{\#T} - 2E\{\mathbf{u}\mathbf{x}^T\}, \tag{11}$$

where $\mathbf{W}^{\#} = (\mathbf{D}_x\mathbf{W}^T)(\mathbf{W}\mathbf{D}_x\mathbf{W}^T)^{-1}$ is the pseudoinverse of $\mathbf{W}$ with respect to the positive definite sample covariance matrix $\mathbf{D}_x$. However, when considering the whitened data, the covariance matrix equals the identity matrix, simplifying the first term of (10) to $\frac{1}{2}\log|\mathbf{W}\mathbf{W}^T|$ and $\mathbf{W}^{\#}$ to $\mathbf{W}^T(\mathbf{W}\mathbf{W}^T)^{-1}$.

All the aforementioned approaches treat the mixture $\mathbf{X}$ of independent components $\mathbf{S}$ as a linear one. It may happen to have components that are mixed through nonlinear functions. A kernel Hilbert space is used by Bach and Jordan to come up with the so called *kernel-ICA* approach in order to extract such nonlinearly mixed sources [20]. Two contrast functions that rely on canonical correlations in this reproducing space have been defined namely the *kernel ICA-KCCA* (where KCCA stands for Kernel Canonical Correlation Analysis) and the *ICA-KGV* (where KGV stands for Kernel Generalized Variance). Kernel ICA-KCCA minimizes the first kernel canonical correlation that depends on the data $\mathbf{x}_j$, $j = 1, \ldots, m$ only through the centered Gram matrices for l ICs. Kernel ICA-KGV minimizes the kernel generalized variance. The interested reader may consult [20] for more details.

IV. Two architectures for performing ICA on facial expression images

Donato suggests that ICA features contain suitable and powerful discriminative information for classifying facial action units [10]. Facial expressions are combinations of such

facial action units. Hence, ICA features may also be suitable for facial expression classification. In this paper, ICA is applied to facial images for feature extraction towards facial expression classification. We have m images containing human facial expressions, each image being of size $r \times c$ pixels, vectorized into a $p = rc$-dimensional vector by lexicographic ordering. ICA can be applied to facial images for expression classification in two ways known as Architectures I and II, respectively [24].

A. Architecture I

The observation matrix $\mathbf{X}$ is formed by treating the facial images as row vectors. Thus $\mathbf{X}$ is an $m \times p$ matrix. By doing so, ICA recovers m independent images.

First, PCA is applied. Let $\mathbf{D}_x$ be the covariance matrix of the original images, $\mathbf{D}_x = \frac{1}{m}\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}} = \mathbf{P}\boldsymbol{\Lambda}\mathbf{P}^T$, where $\widetilde{\mathbf{X}}^T = [\mathbf{x}_1 - \boldsymbol{\psi}| \ldots |\mathbf{x}_m - \boldsymbol{\psi}]$ with $\boldsymbol{\psi} = \frac{1}{m}\sum_{k=1}^{m}\mathbf{x}_k$. Let us choose $l < p$ eigenvectors of $\mathbf{D}_x$ (those with the largest eigenvalues) and form $\mathbf{P}_l \in \mathbb{R}^{p\times l}$ whose columns are the eigenvectors. Each training face image $\mathbf{x}_k$ can be projected to the eigenvectors (called here eigenfaces) and be represented by $\mathbf{y}_k = \mathbf{P}_l^T(\mathbf{x}_k - \boldsymbol{\psi})$. Let us construct $\mathbf{Y}^T = [\mathbf{y}_1| \ldots |\mathbf{y}_m]$. Then $\mathbf{Y} = \widetilde{\mathbf{X}}\mathbf{P}_l$. The original images can be reconstructed as linear combinations of the basis images $\mathbf{P}_l$ as $\mathbf{X}_{recPCA} = \mathbf{Y}\mathbf{P}_l^T$. In the following, we assume that $\boldsymbol{\psi} = 0$ and accordingly $\widetilde{\mathbf{X}} = \mathbf{X}$.

Next, ICA is applied to $\mathbf{P}_l^T$. A number of l ICs can be recovered into the rows of basis $\mathbf{U}$:

$$\mathbf{U} = \mathbf{W}\mathbf{P}_l^T. \tag{12}$$

Hence, we have $\mathbf{P}_l^T = \mathbf{W}^{-1}\mathbf{U}$, provided that $\mathbf{W}$ is invertible and the ICA reconstruction of $\mathbf{X}$ is given by the approximation:

$$\mathbf{X}_{recICA} = \mathbf{Y}\mathbf{P}_l^T = \mathbf{Y}(\mathbf{W}^{-1}\mathbf{U}) = (\mathbf{X}\mathbf{P}_l\mathbf{W}^{-1})\mathbf{U}. \tag{13}$$

The rows of $\mathbf{B} = \mathbf{X}\mathbf{P}_l\mathbf{W}^{-1}$ contain the ICA coefficients of the linear combination of rows (basis vectors) in $\mathbf{U}$, where the training images are represented by the matrix $\mathbf{X}$. The rows of $\mathbf{B}$ are used further for classification. The ICA coefficients of a zero-mean test image $\mathbf{x}_{test}$ are obtained as:

$$\mathbf{b}_{test}^T = \mathbf{x}_{test}^T\mathbf{P}_l\mathbf{W}^{-1}. \tag{14}$$

To perform classification based on distances or angles between $\mathbf{b}_{test}^T$ given by (14) and the rows of $\mathbf{B}$, the basis vectors should be orthonormal. Accordingly, $\mathbf{U}$ should be a row orthonormal matrix (i.e. $\mathbf{U}\mathbf{U}^T = \mathbf{I}$) or equivalently $\mathbf{W}\mathbf{W}^T = \mathbf{I}$. This implies that $\mathbf{W}$ is a rotation matrix. An ICA approach that returns a rotation matrix for $\mathbf{W}$ is fastICA. Accordingly, the ICA coefficients (14) do not offer more information than those derived by PCA coefficients for the aforementioned classifiers as it was also pointed out in [25]. To address this point, we perform ICA subspace selection. If $\mathbf{U}$ is not orthonormal, then it should undergo a Gram-Schmidt orthogonalization (i.e. a QR decomposition) before classification that is based on distances or angles.

B. Architecture II

Now consider $\mathbf{X}^T$. In this case, the pixels are assumed to be independent [24]. The columns of $\mathbf{X}$ are linear combinations of basis vectors obtained from the columns of matrix $\mathbf{W}$. In Architecture II, ICA is performed on the projected data $\mathbf{Y}^T = \mathbf{P}_l^T\mathbf{X}^T$. Therefore, the basis images obtained by performing PCA and ICA can be represented as $\mathbf{P}_l\mathbf{W}^{-1}$ and the coefficients needed for ICA reconstruction are expressed by the columns of $\mathbf{U} = \mathbf{W}\mathbf{Y}^T$. The reconstructed images are:

$$\mathbf{X}_{recICA}^T = (\mathbf{P}_l\mathbf{W}^{-1})(\mathbf{W}\mathbf{Y}^T). \tag{15}$$

A zero-mean test image is represented as:

$$\mathbf{u}_{test} = \mathbf{W}\mathbf{P}_l^T\mathbf{x}_{test}. \tag{16}$$

To perform classification based on distances or angles between $\mathbf{u}_{test}$ given by (16) and the columns of $\mathbf{U}$, the basis images should be orthonormal. This implies that $\mathbf{W}$ should be an orthonormal matrix, i.e. $\mathbf{W}^T\mathbf{W} = \mathbf{I}$. In such a case, the ICA coefficients (16) do not offer any additional information than the coefficients derived by PCA in this case as well.

V. Data description

The experiments have been performed using two databases of facial expression images. The first database has been derived from the Cohn-Kanade (C-K) AU-coded facial expression database [26] that contains single or combined action units. Facial action units

Fig. 1. An example of one expresser from the JAFFE database posing 7 facial expressions (first row) and another one from the Cohn-Kanade database posing 6 facial expressions (second row).

have been converted to emotions according to [27]. Thirteen persons (expressers) who are able to express the six basic emotions create the database. Each subject from C-K database delivers an expression over time starting from a neutral pose and ending with a very intense expression, thus having several frames with different expression intensities. We picked up three poses with low (close to neutral), medium, and high (close to the maximum) intensity of facial expression, respectively. By doing so, the statistical variability of facial emotions is roughly captured. Therefore, the total number of images is 234 in the first database. The second database contains 213 images of Japanese female facial expressions (JAFFE) [28]. Ten expressers produced 3 or 4 examples for each of the 6 basic facial expressions (anger, disgust, fear, happiness, sadness, surprise) plus a neutral pose, thus producing a total of 213 images of facial expressions. Let us enumerate the 7 facial expressions in JAFEE by $j = 1, \ldots, 7$. In the case of the C-K database, we have only 6 expressions, therefore the enumeration ends at 6. Table I summarizes the details for the two databases.

Each raw image $\mathbf{x}$ has been manually aligned with respect to the upper left face corner. The registration was performed by clicking the eyes - thus retrieving the eyes coordinates, followed by rotating the image to horizontally align the face according to eyes, cropping the face to remove the image borders and, finally, downsampling the image to a final size of 60×45 pixels for computational purposes. Figure 1 presents samples of facial expressions of one person from the JAFFE database posing 7 facial expressions and another person from the C-K database posing 6 facial expressions.

TABLE I

DETAILS OF THE TWO DATABASES.

Details	C-K database	JAFFE database
Expressers	13	10
Emotions	6	7
Instances per emotion	3	2,3 or 4
Total number of facial images	234	213
Number of training images	164	150
Number of test images	70	63

VI. CLASSIFIERS

In the experiments, two different classifiers are employed. We used the *Cosine Similarity Measure* (CSM) classifier, since such a classifier was reported to yield a good classification performance [10]. The classification method is based on the nearest neighbor rule and uses the angle between a test vector $\mathbf{b}_{test}$ and the facial expression class center $\mathbf{b}_j$ as a similarity measure:

$$d_j = \frac{\mathbf{b}_{test}^T \mathbf{b}_j}{\|\mathbf{b}_{test}\| \|\mathbf{b}_j\|} \qquad j = 1, \ldots, N_e, \tag{17}$$

where $N_e = 7$ for JAFFE ($N_e = 6$ for C-K database) and chooses the class that corresponds to the maximal cosine similarity

$$\arg\max_{j=1,\ldots,N_e} \{d_j\}. \tag{18}$$

In the case of Architecture II, $\mathbf{b}$ is replaced by $\mathbf{u}$. From (17) it is seen that CSM is an 1-nearest neighbor classifier for normalized feature vectors.

SVMs [29] were employed for facial expression recognition, too. The sequential minimal optimization technique developed by Platt [30] was used to train SVMs having $\mathbf{b}$ and $\mathbf{u}$ as input, respectively. Since classical SVM theory was intended to solve a two class classification problem, we chose the Decision Directed Acyclic Graph (DDAG) learning architecture proposed by Platt et al. to cope with the multi-class classification [31]. It is worth noting that CSM and SVMs are the most popular classifiers for facial expression recognition, as they have been extensively used in [9], [10], [12].

The classifier accuracy, defined as the percentage of the correctly classified test images, is used to assess the performance of the facial expression recognition systems that employ the six ICA approaches in order to extract features, which subsequently feed the aforementioned classifiers.

VII. ICA assessment

The six ICA approaches were applied to create feature vectors $\mathbf{b}_j, \mathbf{b}_{test}$ or $\mathbf{u}_j, \mathbf{u}_{test}$. We split the data into disjoint training and test sets. We used 164 and 150 images for training and we left out 70 and 63 images for testing in the C-K and JAFFE database, respectively. Both training and test set images were chosen randomly from the database. However, we ensured that both training and test data sets contain samples from all expressers and expressions. In the case of SVMs, five kernels were used namely the linear kernel, the polynomial kernel of degree 2,3, and 4, and the radial basis function (RBF). For all SVMs the penalizing parameter was set to 10 and the width of RBF kernel was set to 0.005. Among the five kernels only the two kernels, which yield the highest accuracy, are retained. However, for the JAFFE database and Architecture II, three kernels are retained, because the linear kernel has performed equally well to the polynomial kernel of degree 3.

The first objective is to find which ICA image representation performs best with respect to the classifier accuracy. Experiments were conducted by varying the number of principal components (PCs) from 5 to 160 (for the C-K database) and from 5 to 145 (for the JAFFE database) accounting from 24% to 99.8% of the trace of the covariance matrix. Due to the limited memory capacity and the algorithmic complexity, we were able to extract up to a maximum of 80 components in the JADE and the kernel-ICA approaches.

In order to see if the accuracy differences are statistically significant, we apply the approximate analysis described in [32]. We have examined if accuracy differences are statistically significant for pairs of the same classifier fed by features extracted by two different ICA approaches as well as for pairs of different classifiers fed by the best performing ICA approaches. The analysis is repeated for each database and architecture. Let us assume that the accuracies p_1 and p_2 are binomially distributed random variables. Let $\hat{p}_1, \hat{p}_2$ denote the empirical accuracies, and $\overline{p} = \frac{\hat{p}_1+\hat{p}_2}{2}$. The hypothesis $H_0 : p_1 = p_2 = \overline{p}$ is tested at 95% level of significance. The accuracy difference has variance $\beta = var(p_1 - p_2) = 2\frac{\overline{p}(1-\overline{p})}{N}$,

where N is the number of test facial expression images. If

$$\hat{p}_1 - \hat{p}_2 \geq 1.65\ \sqrt{\beta} \tag{19}$$

we reject H_0 with risk 5% of being wrong. Then, we may claim that the accuracy difference is statistically significant at 95% level of significance.

The second issue investigated in the paper is related to the variation of recognition accuracy with respect to the mutual information of the basis images or their coefficients. The statistical dependencies of facial expression representations were measured by computing the average mutual information between pairs of basis images that yield the maximum recognition accuracy. The mutual information of two RVs u_1, u_2 is given by:

$$I(\mathsf{u}_1, \mathsf{u}_2) = H(\mathsf{u}_1) + H(\mathsf{u}_2) - H(\mathsf{u}_1, \mathsf{u}_2) \tag{20}$$

where H(u) is the differential entropy of the RV u [24]. The average mutual information calculated over all possible pairs of basis images is a good measure of the independence of basis images.

The nature of independent components (ICs) and the influence of discarded PCs in the recognition accuracy are investigated as well. The super- and sub-Gaussian nature of basis images was tested by measuring their normalized kurtosis (1). Furthermore, non-linear mixtures of independent components were also investigated.

To obtain a better quantitative insight on how well the accuracy is correlated to the mutual information and the kurtosis over the number of PCs, we have computed the correlation coefficient and the corresponding p-value. Mutual information, kurtosis, and accuracy were computed for various numbers of components from the set $\{5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120, 130, 140,\ 150, 160\}$ for the C-K database and $\{5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120, 130, 140\}$ for the JAFFE database. Accordingly, we have 17 (15) values of the aforementioned quantities (mutual information, kurtosis, accuracy) for varying numbers of components that are stored in three 17(15)-dimensional vectors. The correlation was then calculated between the elements of the vector comprising the mutual information values and the vector comprising the accuracy values as well as between the vector having as elements the kurtosis values and the vector of accuracies.

A. Cohn-Kanade database

A.1 Architecture I

The experimental results are presented in Table II. The number of PCs varies between 5 and 160 and admits the values in the set $\{5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120, 130, 140, 150, 160\}$. For each number of PCs, features were extracted by the several ICA approaches and the classifier accuracy is measured over the test facial expression images. The highest accuracy obtained along with the corresponding number of PCs are listed in columns numbered by "1" and "2". For both the CSM classifier and the SVM with a polynomial kernel of degree 3, a small number of PCs yields a close to the best classification accuracy. The classification accuracy obtained by the CSM classifier, when it employs features extracted by the InfoMax, the JADE, the fastICA, and the kernel-ICA was found to be identical. A decrease of approximately 3 % in accuracy was found, when features extracted by the Extended InfoMax and the uICA. Overall, the best recognition accuracy was 82.9 % and was obtained by the linear SVM with fastICA, when 110 PCs were used. While such a large number of PCs is needed for the linear SVM in order to achieve the highest accuracy, 30 PCs are adequate for the SVM with a polynomial kernel of degree 3 in order to attain an accuracy of 81.43 %, which is reasonable compromise between accuracy and dimensionality reduction. In Table II, the highest accuracy appears in bold.

For each classifier, the accuracy differences due to different ICA approaches are not statistically significant at 95 % level of significance. The accuracy differences between the several pairs of classifiers that employ the best performing ICA approaches, such as (CSM & fastICA, SVM linear & fastICA), (SVM linear & fastICA, SVM cubic & extended ICA) etc., are not statistically significant at 95 % level of significance as well.

One merit of ICA is that it produces independent and sparse basis images or coefficients depending on the architecture employed. For Architecture I, the basis images are expected to be independent and sparse. Their independence is measured by the average mutual information listed in the third column of Table II.

The presence of a super- or a sub-Gaussian distribution in the basis images is tested in columns "4" and "5" of Table II. These columns show the average positive and negative kurtosis of the basis images indicating a super-Gaussian and a sub-Gaussian

TABLE II

EXPERIMENTAL RESULTS FOR THE C-K DATABASE AND ARCHITECTURE I. THE COLUMNS NUMBERED FROM 1 TO 9 REPRESENT: 1) CLASSIFICATION ACCURACY (%), 2) NUMBER OF PCS, 3) AVERAGE BASIS IMAGE MUTUAL INFORMATION, 4) AND 5) NORMALIZED AVERAGE POSITIVE AND NEGATIVE KURTOSIS OF THE BASIS IMAGES, 6) AND 7) CORRELATION COEFFICIENT BETWEEN THE CLASSIFICATION ACCURACY AND THE MUTUAL INFORMATION WITH ITS CORRESPONDING P-VALUE, 8) AND 9) CORRELATION COEFFICIENT BETWEEN THE CLASSIFICATION ACCURACY AND THE POSITIVE KURTOSIS WITH ITS CORRESPONDING P-VALUE.

Classifier	Approach	1 (%)	2	3	4	5	6	7	8	9
CSM	InfoMax	74.3	10	0.0758	4.1	NA	-0.03	0.91	0.01	0.95
	Extended InfoMax	71.4	10	0.0794	3.4	-0.8	-0.44	0.14	0.42	0.16
	JADE	74.3	30	0.0332	14.1	NA	-0.44	0.22	0.27	0.47
	fastICA	74.3	30	0.0341	13.8	-0.5	-0.44	0.14	0.36	0.24
	uICA	71.4	50	0.0082	32.9	-0.7	-0.31	0.38	0.27	0.31
	kernel-ICA	74.3	30	0.0628	1.38	NA	-0.55	0.12	0.82	0.006
SVM linear	InfoMax	80	110	0.0013	34.8	NA	-0.97	0	0.84	0.0006
	Extended InfoMax	81.4	130	0.0014	46.3	-1.5	-0.98	0	0.80	0.0001
	JADE	78.6	70	0.0067	27.6	NA	-0.99	0	0.92	0.0003
	fastICA	**82.9**	110	0.0023	49.9	0	-0.97	0	0.78	0.0002
	uICA	82.7	140	0.0318	1.2	NA	-0.78	0.0002	0.68	0.002
	kernel-ICA	78.6	70	0.0440	1.4	NA	-0.80	0.007	0.67	0.012
SVM polynomial ($q = 3$)	InfoMax	80	20	0.0480	8.4	NA	-0.56	0.053	0.34	0.27
	Extended InfoMax	81.4	30	0.0353	13.6	-0.9	-0.63	0.026	0.40	0.19
	JADE	80	20	0.0505	9.2	NA	-0.60	0.020	0.56	0.28
	fastICA	80	20	0.0480	8.6	-0.7	-0.47	0.12	0.26	0.39
	uICA	78.3	100	0.0430	1.0	NA	-0.49	0.10	0.38	0.21
	kernel-ICA	80	20	0.0743	1.1	NA	-0.50	0.28	0.52	0.30

distribution, respectively, and constitute a measure of sparseness of the basis images. "NA" in the column "5" stands for "Not Available", i.e. when a sub-Gaussian distribution of basis images is not detected. The average negative kurtosis listed in column "5" shows that the presence of sub-Gaussian components does not necessarily enhance the classifier performance.

Ten basis images extracted from the C-K database during training with each method in the case of Architecture I are depicted in Figure 2. As one can notice from Figure 2, the

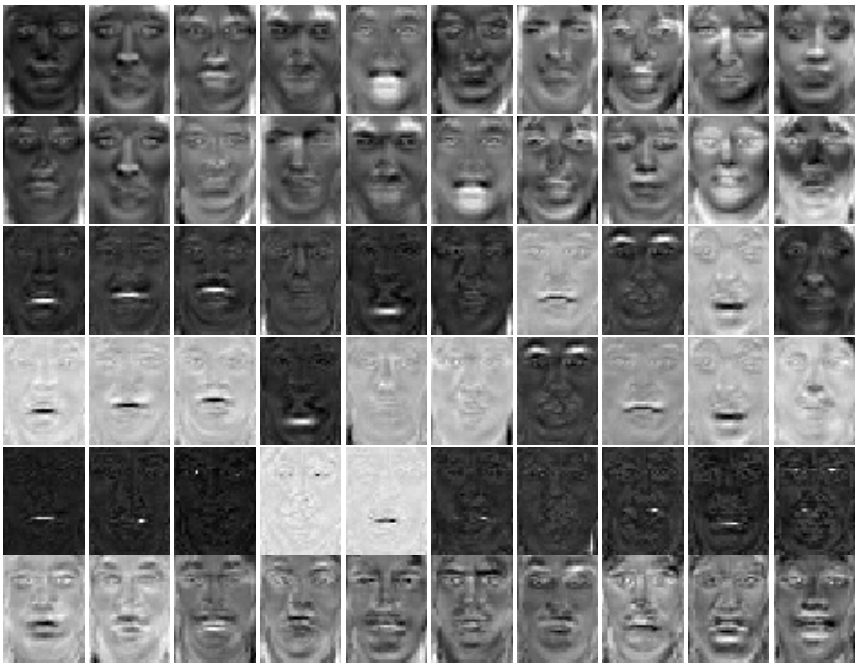

Fig. 2. First ten basis images for Architecture I obtained by InfoMax (1st row), Extended InfoMax (2nd row), JADE (3rd row), fastICA (4th row), undercomplete ICA (5th row), and kernel-ICA (6th row). The images are depicted in decreasing order of normalized kurtosis.

basis images for JADE, fastICA, and uICA are more sparse than the basis images derived by the remaining methods.

Columns "6" and "7" in Table II record the correlation coefficient between the accuracy and the average mutual information over all possible pairs of basis images extracted for each number of PCs (mutual information for short, hereafter) and the corresponding p-value. The last two columns list the correlation coefficient between the classification accuracy and the average positive kurtosis of the basis images (positive kurtosis for short,

hereafter). The strongest correlation between accuracy and mutual information was found for the linear SVM. The minus sign achieved for all classifiers indicates a negative correlation, meaning that a decrease in mutual information (hence greater independence) correlates with an increase of the classifier accuracy. The correlation is weak in the case of the CSM classifier and the SVM with a polynomial kernel of degree 3. Indeed, for the CSM classifier, the p-value exceeds 0.05, a fact that indicates that the correlation coefficient is not statistically significant. For the SVM with a polynomial kernel of degree 3, the Extended InfoMax and the JADE exhibit a correlation coefficient between accuracy and mutual information that is statistically significant. A similar behavior was observed for the correlation between the basis image sparseness and accuracy. For an SVM with a linear kernel, a strong statistically significant correlation between accuracy and the positive kurtosis values is found.

The uICA was used in order to avoid discarding PCs having a small variance, but might contain ICs. The uICA was not able to improve the accuracy by processing the original image data. On the contrary, for the CSM classifier and the SVM with a polynomial kernel of degree 3, applying PCA for input dimensionality reduction is found to be a good practice, since it yields a high accuracy for a small number of PCs.

The linear SVM is the only classifier for which the details count, since a large number of PCs is needed in order to obtain the highest accuracy. However, this is due to the linear separating hyperplane which performs best in high-dimensional spaces.

To assess the descriptive power of nonlinear IC mixtures, Kernel-ICA is applied. It is observed that the nonlinear ICA does not enhance the classification accuracy.

The classification accuracy reported in Table II has been averaged over all test facial expression instances. To obtain a better insight into the performance of the linear SVM, which employs fastICA for feature extraction in the C-K database, we have computed the confusion matrix during testing shown in Table III. The rows of the confusion matrix refer to the actual (correct or ground truth) expression labels and its columns refer to the predicted expression labels by the classifier. Its diagonal entries correspond to the number of facial expressions correctly classified, while its off-diagonal entries record the numbers of misclassified test facial expression instances. It is seen that more errors are committed

when angry and sad facial expressions are processed. This result is not surprising as the two expressions look similar and can be easily confused by humans too. We must note that, having the confusion matrix at disposal, a multiclass ROC analysis might be derived [33].

TABLE III

CONFUSION MATRIX FOR THE TEST IMAGES FROM C-K DATABASE, WHEN A LINEAR SVM IN ARCITECTURE I WITH FASTICA IS EMPLOYED.

	anger	disgust	fear	happiness	sadness	surprise
anger	5	0	0	0	3	0
disgust	2	10	1	0	0	0
fear	1	0	10	0	0	0
happiness	1	0	1	15	0	0
sadness	2	0	0	0	8	0
surprise	0	0	0	0	1	10

A.2 Architecture II

The experimental findings are summarized in Table IV. All ICA approaches with the CSM classifier yield the same accuracy (72.9%), as one can see from column "1". The best accuracy (80%) was obtained by the SVM with an RBF kernel that employs features extracted by the Extended InfoMax. However, the accuracy difference between 80% and 72.9% is not statistically significant for 95% level of significance. Moreover, the pairwise performance differences within each classifier due to different ICA approaches are not statistically significant at the same level of significance. This is also valid for all pairs of classifiers that employ the ICA approach yielding the highest accuracy.

The second architecture derives coefficients that are as independent and sparse as possible. The mutual information and the average positive and negative kurtosis was measured for coefficients, as shown in columns "3"–"5" of Table IV. Ten basis images corresponding to C-K database which are obtained after training each method in Architecture II are depicted in Figure 3. They have a rather holistic appearance compared with the sparse basis images of Figure 2.

TABLE IV

EXPERIMENTAL RESULTS FOR THE C-K DATABASE AND ARCHITECTURE II. THE COLUMNS NUMBERED FROM 1 TO 9 REPRESENT: 1) CLASSIFICATION ACCURACY (%), 2) NUMBER OF PCS, 3) AVERAGE COEFFICIENT MUTUAL INFORMATION, 4) AND 5) NORMALIZED AVERAGE KURTOSIS OF SUPER- AND SUB-GAUSSIAN COEFFICIENTS, 6) AND 7) CORRELATION COEFFICIENT BETWEEN THE CLASSIFICATION ACCURACY AND THE MUTUAL INFORMATION WITH ITS CORRESPONDING P-VALUE, 8) AND 9) CORRELATION COEFFICIENT BETWEEN THE CLASSIFICATION ACCURACY AND THE POSITIVE KURTOSIS WITH ITS CORRESPONDING P-VALUE.

Classifier	Approach	1 (%)	2	3	4	5	6	7	8	9
CSM	InfoMax	72.9	40	0.0260	14.7	NA	-0.70	0.01	0.26	0.41
	Extended InfoMax	72.9	10	0.1363	2.3	-1.3	-0.57	0.049	0.64	0.02
	JADE	72.9	10	0.1311	1.1	NA	-0.49	0.176	0.09	0.08
	fastICA	72.9	10	0.0884	3.5	-1.7	-0.21	0.50	0.08	0.78
	uICA	72.9	60	0.0002	0.1	-1.8	-0.21	0.49	0.32	0.308
	kernel-ICA	72.9	10	0.1311	1.1	-0.5	-0.36	0.337	0.03	0.92
SVM linear	InfoMax	75.7	90	0.0050	38.6	NA	-0.91	0	0.60	0.003
	Extended InfoMax	72.8	110	0.0005	5.2	-1.5	-0.98	0	0.80	0.001
	JADE	72.8	60	0.013	42.1	NA	-0.94	0.0004	-0.06	0.88
	fastICA	75.2	110	0.006	30.2	0	-0.98	0	-0.9	0.005
	uICA	73.3	100	0.008	10.5	-0.5	-0.70	0.1	0.65	0.02
	kernel-ICA	75.7	40	0.020	0.4	-0.8	-0.75	0.1	0.48	0.4
SVM polynomial ($q = 3$)	InfoMax	71.4	20	0.0049	8.9	NA	-0.11	0.73	0.71	0.008
	Extended InfoMax	74.3	10	0.1363	2.3	-1.3	-0.08	0.79	0.03	0.91
	JADE	75.7	20	0.0432	0.8	NA	-0.10	0.80	0.40	0.09
	fastICA	75.7	20	0.0001	8.5	-0.3	0.27	0.38	0.76	0.004
	uICA	75.7	90	0.0013	9.1	-0.3	-0.23	0.46	0.76	0.47
	kernel-ICA	75.7	20	0.0440	0.8	-0.5	-0.20	0.3	0.45	0.10
SVM RBF	InfoMax	74.3	30	0.0192	12.1	NA	-0.54	0.069	0.10	0.75
	Extended InfoMax	**80**	120	0.0038	6.8	-1.4	-0.96	0	0.88	0
	JADE	75.7	70	0.0534	51.8	NA	-0.78	0.008	0.74	0.009
	fastICA	78.6	100	0.0659	76.3	0	-0.99	0	0.74	0.005
	uICA	71.8	60	0.0002	0.1	-1.8	-0.17	0.59	0.65	0.019
	kernel-ICA	75.7	70	0.0070	1.7	-0.3	-0.41	0.3	0.57	0.09

Fig. 3. First ten basis images for Architecture II obtained by InfoMax (1st row), Extended InfoMax (2nd row), JADE (3rd row), fastICA (4th row), undercomplete ICA (5th row), and kernel-ICA (6th row). The images are depicted in decreasing order of normalized kurtosis.

As for Architecture I, a weak correlation between the CSM classifier accuracy and mutual information was found. Only InfoMax and Extended InfoMax yield a statistically significant correlation. In contrast, strong statistically significant correlations between the accuracy of the SVM classifier with an RBF kernel and mutual information were measured. In this case, 3 out of the 6 ICA approaches yield statistically significant correlations and the best performing classifier (i.e., SVM-RBF with Extended InfoMax) shows the second highest correlation. The linear SVM shows a strong correlation between mutual information and accuracy at least for 4 out of the 6 ICA approaches (i.e., Informax, Extended InfoMax, JADE, fastICA) consistently in Tables II - V. This suggests that independence is associated with a more linearly separated feature space.

Overall, the Architecture II yields a smaller classification accuracy than the Architecture I.

B. JAFFE database

B.1 Architecture I

The experimental results are summarized in Table V. The facial expressions in JAFFE database are a little bit harder to be recognized than those recorded in the C-K database due to the fact that the human expressers in the former database were less expressive than those in the latter database. As a consequence, a larger number of PCs had to be retained in order to obtain the maximum recognition rate of 66.67% for the CSM classifier. This rate was obtained by all ICA approaches with Architecture I. However, the accuracy differences between all possible classifier pairs employing different ICA approaches are not statistically significant at 95 % level of significance.

In JAFFE database, a statistically significant correlation coefficient between mutual information and the accuracy of the CSM classifier for all ICA approaches was found except uICA. Moreover, the correlation coefficient between the accuracy of the CSM classifier and kurtosis was found to be statistically significant for all ICA approaches. This was not the case for the correlation coefficient between the accuracy of the CSM classifier and either mutual information or kurtosis for the C-K database. The linear SVM classifier yields the highest accuracy 79.4 %, when the extended InfoMax and the fastICA approaches are employed. From the inspection of Table V, it is seen that very strong statistically significant correlations between the classification accuracy and the mutual information of basis images as well as the classification accuracy and the positive kurtosis of the basis images are measured for the best performing ICA approaches with the linear SVM. Table VI depicts the confusion matrix for test images from JAFFE when a linear SVM in Architecture-I with fastICA is employed. It is seen that "fear" is the most difficult expression to be recognized, which is confused 2 times with "neutral", another 2 times with "sadness", once with "anger" and another time with "disgust". We note that the expressers from the JAFFE database are less expressive compared to those from the C-K database.

TABLE V

EXPERIMENTAL RESULTS FOR THE JAFFE DATABASE AND ARCHITECTURE I. THE COLUMNS NUMBERED FROM 1 TO 9 REPRESENT: 1) CLASSIFICATION ACCURACY (%), 2) NUMBER OF PCS, 3) AVERAGE BASIS IMAGE MUTUAL INFORMATION, 4) AND 5) NORMALIZED AVERAGE POSITIVE AND NEGATIVE KURTOSIS OF THE BASIS IMAGES, 6) AND 7) CORRELATION COEFFICIENT BETWEEN THE CLASSIFICATION ACCURACY AND THE MUTUAL INFORMATION WITH ITS CORRESPONDING P-VALUE, 8) AND 9) CORRELATION COEFFICIENT BETWEEN THE CLASSIFICATION ACCURACY AND THE POSITIVE KURTOSIS WITH ITS CORRESPONDING P-VALUE.

Classifier	Approach	1 (%)	2	3	4	5	6	7	8	9
CSM	InfoMax	66.7	40	0.0077	15.5	NA	-0.75	0.004	0.62	0.030
	Extended InfoMax	66.7	50	0.0043	16.4	NA	-0.85	0.0004	0.66	0.017
	JADE	66.6	50	0.0014	19.8	NA	-0.81	0.007	0.68	0.040
	fastICA	66.7	50	0.0041	17.0	-0.5	-0.88	0	0.70	0.010
	uICA	66.7	60	0.0066	5.6	-0.2	-0.41	0.183	0.69	0.011
	kernel-ICA	66.7	50	0.0179	2.2	NA	-0.84	0.003	0.72	0.027
SVM linear	InfoMax	76.2	60	0.0013	19.6	NA	-0.98	0	0.92	0
	Extended InfoMax	**79.4**	110	0.0095	29.5	NA	-0.99	0	0.92	0
	JADE	73.2	80	0.0089	31.8	NA	-0.77	0.008	0.74	0.09
	fastICA	**79.4**	110	0.0095	27.4	NA	-0.97	0.001	0.91	0
	uICA	77.2	110	0.0113	1.4	NA	-0.83	0.001	0.62	0.009
	kernel-ICA	76.2	80	0.0097	2.3	NA	-0.60	0.3	0.26	0.2
SVM RBF	InfoMax	71.4	70	0.0028	22.6	NA	-0.92	0	0.83	0.007
	Extended InfoMax	60.3	20	0.0289	7.4	NA	-0.51	0.08	0.74	0.005
	JADE	63.4	20	0.0263	8.7	NA	-0.62	0.36	0.71	0.09
	fastICA	63.4	20	0.0266	8.1	NA	-0.42	0.17	0.14	0.65
	uICA	62.5	40	0.0122	22.9	-0.2	-0.39	0.20	0.21	0.50
	kernel-ICA	63.5	20	0.0396	1.9	NA	-0.45	0.09	0.40	0.19

TABLE VI

CONFUSION MATRIX FOR THE TEST IMAGES FROM THE JAFFE DATABASE WHEN A LINEAR SVM IN ARCITECTURE I WITH FASTICA IS EMPLOYED.

	anger	disgust	fear	happiness	neutral	sadness	surprise
anger	7	0	0	0	0	0	0
disgust	0	8	0	0	0	0	0
fear	1	1	6	0	2	2	0
happiness	0	0	0	9	1	2	0
neutral	0	0	0	0	7	1	0
sadness	1	0	0	0	0	7	0
surprise	0	0	0	0	2	0	6

B.2 Architecture II

The highest accuracy of 79.4 % was obtained with the linear SVM and fastICA. For the SVM-RBF classifier, it is worth mentioning that the accuracy difference when Extended InfoMax is employed instead of uICA is statistically significant at the 95% level of significance. All other pairwise accuracy differences either within the same classifier due to different ICA approaches employed or across different classifiers are statistically insignificant at the same level of significance.

In the case of the SVM with a linear kernel, a statistically significant strong correlation between the classification accuracy and the mutual information was found for features extracted by InfoMax, Extended InfoMax, JADE, and fastICA. The negative correlation between the accuracy of the linear SVM and mutual information indicates again that performance increases as mutual information decreases.

C. Performance enhancement using leave-one-set of expressions-out

One possible way of improving accuracy is by exploiting maximally the available data set. To do so, we repeated the experiments by employing the leave-one-set of expressions-out (leave-one-out for short, [LVO]) strategy. That is, one set of expressions was left out for test in a cyclic fashion. During one rotation, the number of training images is 228 and the number of test images is 6 and by performing 39 rotations overall 234 test images are

produced for the C-K database. In a similar way, the rotations yield 214 test images for the JAFFE database.

For both databases, the accuracy of all classifiers employing different ICA approaches was increased substantially, as can be seen in Table VII. For example, an impressive performance enhancement was noticed for the kernel-ICA with the linear SVM in Architecture I applied to the C-K database. Its accuracy was raised from 78.6 % to 86.6% with LVO.

The statistical significance of accuracy differences at 95% level of significance was studied for each Architecture and each database: (i) within the same classifier for all possible pairs due to different ICA approaches; (ii) across different classifiers employing the best performing ICA approaches.

For the C-K database and Architecture I, the only statistically significant accuracy difference is that between the accuracy of the CSM classifier that employs InfoMax (81.4 %)and the SVM with a cubic kernel that employs fastICA (87.6 %). For the C-K database and Architecture II, the use of fastICA instead of uICA within the SVM classifier with an RBF kernel yields statistically significant performance improvement. The reader can verify that the accuracy differences between 84% and 77.3% as well as between 84% and 77% are also statistically significant.

For the JAFFE database and Architecture I, it can easily be checked that the accuracy differences between the CSM classifier and the SVM linear classifier are statistically significant irrespective of the ICA approach employed for feature extraction. Similarly, the InfoMax within the SVM classifier employing an RBF kernel yields a statistically significant performance than the other ICA approaches. The accuracy differences between the CSM classifier and the SVM classifier with an RBF kernel, when InfoMax is used, are also statistically significant. However, there is no statistically significant performance difference between the SVM classifier with a linear kernel that employs fastICA and the SVM classifier with an RBF kernel that employs InfoMax. For the JAFFE database and Architecture II, the use of fastICA instead of uICA within the SVM classifier with a linear kernel yields a statistically significant accuracy difference. Similarly, statistically significant gains exist between the SVM classifier with a liner kernel and fastICA (or the SVM

TABLE VII

AVERAGED ACCURACY OBTAINED WITH LEAVE-ONE-OUT. (NA STANDS FOR ACCURACY RESULTS THAT ARE NOT AVAILABLE).

Classifier	Approach	C-K database		JAFFE database	
		Architecture I	Architecture II	Architecture I	Architecture II
CSM	InfoMax	81.4	77.3	69.6	72.6
	Extended InfoMax	82	80	69.6	70
	JADE	79	81.5	69.6	71
	fastICA	81.3	81.5	69.6	71
	uICA	80.1	81.5	69.6	68.3
	kernel-ICA	81.1	80	69.6	67.8
SVM linear	InfoMax	81.3	NA	80.3	77.5
	Extended InfoMax	81.3	NA	83.5	80
	JADE	82.3	NA	82.5	78
	fastICA	84.6	NA	**84**	**81**
	uICA	83.3	NA	82.6	66
	kernel-ICA	86.6	NA	82.1	78
SVM polynomial ($q = 3$)	InfoMax	83.7	80	NA	NA
	Extended InfoMax	84.6	77	NA	NA
	JADE	82.4	80	NA	NA
	fastICA	**87.6**	80	NA	NA
	uICA	83.3	77.3	NA	NA
	kernel-ICA	85.7	78.2	NA	NA
SVM RBF	InfoMax	NA	81.5	79	79
	Extended InfoMax	NA	83.8	64.7	**81**
	JADE	NA	80	68.3	77.5
	fastICA	NA	**84**	69.3	74
	uICA	NA	70	65.2	72.5
	kernel-ICA	NA	79	68.3	77

classifier with an RBF kernel and Extended InfoMax) and the CSM classifier irrespective of the ICA approach that feeds the latter classifier.

D. Comparisons with PCA

To assess the removal of higher-order correlation captured by ICA, the CSM classifier was directly applied to the eigenimages extracted by PCA. The experiments were ran only for the CSM classifier on the test set and the results are listed in Table VIII. The LVO method, which is detailed in Section VII-C, was also used with PCA and the corresponding results are included in Table VIII.

TABLE VIII

EXPERIMENTAL RESULTS FOR THE C-K AND JAFFE DATABASES WHEN PCA IS USED FOR FEATURE EXTRACTION AND THE CSM CLASSIFIER IS APPLIED TO THE SAME TRAINING AND TEST SETS. THE ACCURACY ESTIMATED BY USING THE LEAVE-ONE-OUT METHOD (LVO) IS ALSO RECORDED.

C-K database												
No. of PCs	5	10	20	30	40	50	60	70	80	90	100	110
Accuracy (test set)	57.1	**75.7**	70	71.4	71.4	72.8	71.4	71.4	72.8	71.4	71.4	68.5
Accuracy (LVO)	67.7	80.7	80.7	78.6	81.2	81.6	81.6	81.6	**82**	81.6	81.6	81.6
JAFFE database												
No. of PCs	5	10	20	30	40	50	60	70	80	90	100	110
Accuracy (test set)	38.1	60.3	65	66.7	66.7	66.7	66.7	66.7	**69.8**	69.8	69.8	69.8
Accuracy (LVO)	50	67.4	72.7	71.4	71	70.5	71.8	71.8	72.7	**73.2**	73.2	73.2

In the C-K database, the highest accuracy on the test set (i.e. 75.71%) is achieved with only 10 PCs. For CSM, the accuracy difference due to PCA instead of the best performing ICA approach in Architecture I (74.3%), as is recorded in Table II, is not statistically

significant at 95% level of significance. Nor is statistically significant at the same level of significance, the accuracy difference due to PCA instead of the best performing ICA approach for CSM in Architecture II (72.9%), as is recorded in Table IV. When PCA is used for feature extraction, the accuracy of CSM is improved by LVO reaching 82%. For the CSM classifier, Table VII reveals that its highest accuracy obtained with ICA approaches in either Architecture I or Architecture II is 81.4% or 81.5%, respectively. Obviously, the accuracy differences are not statistically significant at 95% level of significance.

In the JAFFE database, the same maximum accuracy for the CSM classifier (69.8 %) was obtained by both PCA and ICA with Architecture II in the test set. ICA in Architecture I within the CSM classifier yields accuracy 66.7% in the test set. It can easily be verified that the accuracy differences on the test set are not statistically significant. When PCA is used for feature extraction, the accuracy of CSM is improved by LVO reaching 73.2%. For the CSM classifier, Table VII reveals that its highest accuracy obtained with ICA approaches in either Architecture I or Architecture II is 69.6% or 72.6%, respectively. The accuracy differences are not statistically significant at 95% level of significance, in this case as well.

E. Subspace selection

Unlike PCA, there is no inherent ordering into the independent components [10]. An ordering parameter could be the class discriminability of each component [24] defined as the ratio

$$r = \frac{\sigma_{between}(k)}{\sigma_{within}(k)} \tag{21}$$

where

$$\sigma_{between}(k) = \sum_j (\overline{b}_k^j - \overline{b}_k)^2 \tag{22}$$

$$\sigma_{within}(k) = \sum_j \sum_i (b_k^{ij} - \overline{b}_k^i)^2 \tag{23}$$

with $\overline{b}_k$ denoting the gross mean of coefficient b_k, $\overline{b}_k^j$ being the jth facial expression class mean of coefficient b_k, and b_k^{ij} standing for the kth coefficient of the ith training image in the jth facial expression class.

It has been found that, by ordering the independent components with respect (21), ICA can outperform the PCA approach [10]. We have repeated the experiments with the CSM classifier in Architecture I, when feature selection is done according to (21) and compared the resulted accuracy with that reported previously (i.e. without subspace selection). We conducted the experiments for the maximum number of components and then we selected as many independent components according to (21), so that the maximum accuracy was obtained. The results are summarized in Table IX. By comparing the results in Table IX and those in Table V, one can see that, in JAFFE database, the accuracy obtained by each ICA approach after subspace selection is higher than that reported without subspace selection with the extended ICA being an exception. By cross-examining Tables IX and II, this observation is roughly valid for the accuracy obtained by each ICA approach with the exception of kernel-ICA in C-K database. However, accuracy differences are not statistically significant for neither the C-K database nor the JAFFE one.

TABLE IX

ACCURACY (%) FOR THE CSM CLASSIFIER IN ARCHITECTURE I ON BOTH DATABASES ALONG WITH THE NUMBER OF COMPONENTS CORRESPONDING TO THE MAXIMUM ACCURACY (IN PARENTHESIS AND ITALICS), RETRIEVED BY EMPLOYING SUBSPACE SELECTION.

Database	Approach					
	InfoMax	Extended InfoMax	JADE	fastICA	uICA	kernel-ICA
C-K	77.1 (*80*)	77.1 (*90*)	74.2 (*40*)	78.5 (*110*)	72.5 (*80*)	70 (*30*)
JAFFE	69.8 (*70*)	66.6 (*80*)	68.2 (*50*)	69.8 (*130*)	67.7 (*40*)	68.2 (*50*)

We should also mention that a supervised ICA technique, the so called ICA-FX [34], was developed in order to obtain features that are not only independent from each other, but also convey class information, contrary to the other ICA approaches studied in this paper, which are unsupervised (i.e. they do not utilize the class information). Unlike the method described in [24], ICA-FX allows an intrinsic class information embedding. To examine to what extent the classification performance is affected by incorporating the class information inside the training procedure, we ran the ICA-FX approach on the C-K

database and compared it with the classical ICA approach previously exploited. Due to the fact that the Architecture I does not allow us to make a comparison against ICA-FX, since ICA is performed on the PCA projection matrix implying loss of the class label, we chose ICA Architecture II, where class label is preserved. Table X shows that the CSM classifier yields a higher accuracy when it is fed by features extracted by ICA-FX than those extracted by the other six ICA approaches. The difference in accuracies is found to be statistically significant at 95 % confidence level.

TABLE X

Accuracy results by employing subspace selection with the help of the ICA-FX approach. The results are shown for the Architecture II on C-K database using the CSM and the SVM classifiers.

C-K database, Architecture II		
Classifier	CSM	SVM RBF
Accuracy	84.28	78.8

VIII. Discussion and Conclusions

A systematic comparative study of six ICA approaches was performed for facial expression classification in order to select the one which provides the best classification accuracy using two databases, two facial feature extraction architectures, and two classifiers. Regarding the classification performance, overall, the fastICA combined with SVMs yields a reasonable compromise between accuracy and fast run time for feature extraction. In our study we addressed the following issues:

1. *Performance variation with the number of PCs*: We found that a small number of PCs can produce a reasonable recognition performance for a CSM classifier. Although the present paper exhibits many common issues with the work described in [10], we must notice that the present study differs in too many aspects with that in [10] that does not allow for a fair comparison between the results reported here and in [10].

2. *Implications of applying PCA prior to ICA to reduce data dimensionality*: We found that the use of uICA does not yield a higher classification accuracy than preprocessing

observations by PCA.

3. *Features having super- and sub-Gaussian distribution did not improve facial expression classification accuracy.*

4. *Independent features obtained by non-linear unmixing of observations using kernel-ICA, do not improve the classification performance.* This fact indicates that either there is no such a non-linear mixture in the data, or, if any non-linear mixture exists, its contribution to the classification performance is minimal.

5. The main conclusion drawn from the experiments is that, overall, as can be seen from Tables II- V, *there is a strong correlation between the average mutual information of independent components and accuracy. A similar finding was obtained for sparseness.* For the linear SVM classifier, this relationship is consistently statistically significant, when InfoMax, Extended InfoMax, or fastICA is used for feature extraction. However, the degree of the correlation varies with the classifier and database involved.

ICA yields an efficient coding by performing a sparse image representation and removing the higher order correlations. Whether this is necessary for efficient image representation and pattern recognition purposes, it is still an open problem. It seems (and this is known to the scientific community) that SVMs are more affected by the outliers and noise which is the case of holistic representation. The outliers and "noise" are characterized by those parts of the face that are not essential for facial expression recognition and are present in a holistic representation that has a low degree of sparseness. As more localized features are obtained by ICA by employing more PCs and reducing the mutual information, thus increasing the degree of sparseness, the "noise" is eliminated and the performance of SVM improves. In many cases, we found that obtaining more sparse basis images (or coefficient) does not necessarily lead to a more accurate facial expression classification. These results can be related to the work conducted by Petrov and Li [35]. They investigated local correlation and information redundancy in natural images and they found that the removal of higher-order correlations between the image pixels increased the efficiency of image representation insignificantly. Accordingly, their results suggest that the reduction of higher-order redundancies than the second-order ones is not the main cause of receptive field properties of neurons in V1. It is worth mentioning that there are other sparse image

representations such as Sparse Component Analysis [36], Non-negative Matrix Factorization [37] and Local Non-negative Matrix Factorization [38], for example, which, unlike ICA do not assume component independence.

Although we do not deny the role of sparse image representations in visual cortex, we argue that a more important characteristic of an efficient image representation is feature orientation. Thus, a sparse representation alone does not seem to be sufficient in achieving the maximum recognition performance. This observation comes from [39], where ICA and Gabor filter representation applied to facial expression recognition were compared. Both ICA and Gabor filters approaches gave sparse representations and a highly kurtotic (non-Gaussian) feature distribution. However, the Gabor images that contain important spatially oriented features led to a higher accuracy than the ICA features. Another important aspect is normalization. Brady and Field showed that the entropy of Gabor responses to natural scenes does increase when a V1 response normalization model is applied [40]. This normalization model decreases the high-order dependencies between the Gabor responses in natural scenes, as also shown by Wainwright et al [41], where relationships between Gabor filters, ICA, and sparse coding are investigated. However, while Gabor filters and ICA may be highly related, Gabor filters have an advantage over ICA when the amount of training data is limited. Contrary to ICA, where the features are learned through learning algorithms involving large set of training samples whose size influences the results, Gabor filtering does not actually involve a training procedure [39].

It is worth noting that Vicente et al. recently investigated PCA and ICA [25] by comparing their performances for face recognition when simple classifiers (such as 1-NN classifiers) are involved. Their main conclusion does not contradict ours evidence: no significant performance gains exist between ICA and PCA when no feature selection is performed prior to classification for these classifiers. However, our work differs in many aspects from [25] that deals with face recognition only. First, they compared only InfoMax against fastICA, while we compared several linear and non-linear ICA approaches. Second, we used SVM, which was not employed and compared in their work. Third, they manually projected the data either onto one ICA direction or over one eigenvector direction, while we employed an intrinsic selection based on class information [34].

Finally, some experiments have been performed by Yang et al. [42] who found that the whitening step is responsible for increasing face recognition accuracy not the pure ICA.

IX. Aknowledgement

This work was partially supported by the European Union Research Training Network "MUHCI" on Multi-modal Human-Computer Interaction and the European Network of Excellence "SIMILAR" on Multimodal Interfaces of the IST Programme of the European Union.

References

[1] J. M. Susskind, G. Littlewort, M. S. Bartlett, J. Movellan, and A. K. Anderson, "Human and computer recognition of facial expressions of emotion," *Neuropsychologia*, vol. 45, no.1, pp. 152–162, 2007.

[2] M. Pantic and L .J. M. Rothkrantz, "Automatic analysis of facial expressions: The state-of-the-art," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1424–1445, Dec., 2000.

[3] M. Pantic and L .J. M. Rothkrantz, "Facial action recognition for facial expression analysis from static face images," *IEEE Trans. Systems, Man, and Cybernetics, Part B*, vol. 34, no. 3, pp. 1449–1461, June, 2004.

[4] B. Fasel and J. Luettin, "Automatic Facial Expression Analysis: A survey," *Pattern Recognition*, vol. 1, no. 30, pp. 259–275, 2003.

[5] Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu, "Comparison between geometry-based and Gabor-wavelets-based facial expression recognition using multi-layer perceptron," *in Proc. Third IEEE Int. Conf. Automatic Face and Gesture Recognition*, April 14-16 1998, Nara Japan, pp. 454-459, 1998.

[6] L. Wiskott, J. -M. Fellous, N. Kruger, and C. von der Malsburg, "Face recognition by elastic bunch graph matching," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 775–779, July 1997.

[7] Y.-Li Tian, T. Kanade, and J. Cohn, "Evaluation of Gabor-wavelet-based facial action unit recognition in image sequences of increasing complexity," in *Proc. Fifth IEEE Int. Conf. Automatic Face and Gesture Recognition*, May, pp. 229–234, 2002.

[8] Y.-Li Tian, T. Kanade, and J. Cohn, "Recognizing action units for facial expression analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 23, no.2, pp. 97–115, Feb. 2001.

[9] G. Littlewort, M. Bartlett, I. Fasel, J. Susskind, and J. Movellan, "Dynamics of facial expression extracted automatically from video," *Image and Vision Computing*, vol. 24, no. 6, pp. 615–625, 2006.

[10] G. Donato, M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski, "Classifying facial actions," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 21, no. 10, pp. 974–989, October 1999.

[11] J. Kim, J. Choi, J. Yi, and M. Turk, "Effective representation using ICA for face recognition robust to local distortion and partial occlusion," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 12, pp. 1977-1981, December 2005.

[12] B. A. Draper, K. Baek, M. S. Bartlett and J. R. Beveridge, "Recognizing faces with PCA and ICA," *Computer Vision and Image Understanding*, vol. 91: Special issue on Face Recognition, pp. 115–137, 2003.

[13] B. Moghaddam, "Principal manifolds and Bayesian subspaces for visual recognition," in *Int. Conf. Computer Vision (ICCV'99)*, pp. 1131–1136, 1999.

[14] G. Guo and C. R. Dyer, "Learning from examples in the small sample, case: Face expression recognition," *IEEE Trans. Systems, Man, and Cybernetics, Part B*, vol. 35, no. 3, pp. 477–488, 2005.

[15] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 1129–1159, 1995.

[16] T.-W. Lee, M. Girolami, and T. J. Sejnowski, "Independent component analysis using an extended Infomax algorithm for mixed sub-Gaussian and super-Gaussian sources," *Neural Computation*, vol. 11, no. 2, pp. 417–441, 1999.

[17] J. F. Cardoso and A. Souloumiac, "Blind beamforming for non Gaussian signals," *IEE Proceedings-F*, vol. 140, no. 6, pp. 362-370, 1993.

[18] A. Hyvarinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Trans. Neural Networks*, vol. 10, no. 3, pp. 626–634, 1999.

[19] J. V. Stone and J. Porrill, "Undercomplete independent component analysis for signal separation and dimension reduction," Technical Report, 1998.

[20] F. R. Bach and M. J. Jordan, "Kernel independent component analysis," *Machine Learning Research*, vol. 3, pp. 1–48, 2002.

[21] J. H. Friedman, "Exploratory projection pursuit," *Journal American Statistical Association*, vol. 82, no. 397, pp. 249–266, 1987.

[22] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, N. Y: J. Wiley, 2001.

[23] M. McKeown, S. Makeig, G. Brown, T. Jung, S. Kindermann, and T. Sejnowski, "Spatially independent activity patterns in functional magnetic resonance imaging during the stroop color-naming task," in *Proc. Nat. Acad. Sci.*, vol. 95, pp. 803–810, 1998.

[24] M. S. Bartlett, J. R. Movellan, and T. K. Sejnowski, "Face recognition by independent component analysis," *IEEE Trans. Neural Networks*, vol. 13, no. 6, pp. 1450–1464, 2002.

[25] M. Asunción Vicente, P. O. Hoyer, and A. Hyvärinen, "Equivalence of some common linear feature extraction techniques for appearance-based object recognition tasks," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, no. 5, pp. 896-900, May 2007.

[26] T. Kanade, J. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *Proc. Fourth IEEE Int. Conf. Face and Gesture Recognition*, pp. 46-53, March, 2000.

[27] M. Pantic and L. J. M. Rothkrantz, "Expert system for automatic analysis of facial expressions," *Image and Vision Computing*, no. 18, pp. 881–905, March, 2000.

[28] M. J. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with Gabor wavelets," in *Proc. Third IEEE Int. Conf. Automatic Face and Gesture Recognition*, pp. 200–205, 1998.

[29] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, 1995.

[30] J. C. Platt, "Fast training of support vector machines using sequential minimal optimization," *Advances in Kernel Methods - Support Vector Learning*, vol. 12, pp. 185–208, 1999.

[31] J. C. Platt, N. Cristianini, and J. S.-Taylor, "Large margin DAGs for mutliclass classification," *Advances in Neural Information Processing Systems*, vol. 12, pp. 547–553, 2000.

[32] I. Guyon, J. Makhoul, R. Schwartz, and V. Vapnik, "What size test set gives good error rate estimates?," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 1, pp. 52–64, January 1998.

[33] T. C. W. Landgrebe and R. P. W. Duin "Efficient Multiclass ROC Approximation by Decomposition via Confusion Matrix Perturbation Analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 30, no. 5, pp. 810–822, May 2008.

[34] N. Kwak, C. - H. Choi, and N. Ahuja "Face recognition using feature extraction based on independent component analysis," in *Proc. 2002 IEEE Int. Conf. Image Processing*, pp. 337–340, 2002.

[35] Y. Petrov and Z. Li, "Local correlations, information redundancy, and the sufficient pixel depth in natural images," *Journal Optical Society of America A*, vol. 20, no. 1, pp. 56–66, 2003.

[36] D. L. Donoho, "Sparse component of images and optimal atomic decomposition," Tech. Rep. Dept, Statistics, Stanford University, 1998.

[37] D. D. Lee and H. S. Seung, "Learning the parts of the objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[38] S. Z. Li, X. W. Hou, and H. J. Zhang, "Learning spatially localized, parts-based representation," *Int. Conf. Computer Vision and Pattern Recognition*, pp. 207–212, 2001.

[39] I. Buciu, C. Kotropoulos, and I. Pitas, "ICA and Gabor representation for facial expression recognition," in *Proc. 2003 IEEE Int. Conf. Image Processing*, pp. 855–858, 2003.

[40] N. Brady and D. J. Field, "Local contrast in natural images: normaliation and coding efficiency," *Perception*, vol. 29, no. 9, pp. 1041–1055, 2000.

[41] M. Wainwright, O. Schwartz, and E. Simoncelli, "Natural image statistics and divisive normalization: Modeling nonlinearity and adaptation in cortical neurons," in R. Rao, B. Olshausen, and M. Lewicki (Eds), *Statistical Theories of the Brain: Perception and Neural Function*, MIT Press, 2002.

[42] J. Yang, D. Zhang, and J.-y Yang, "Is ICA Significantly Better than PCA for Face Recognition?," in *Proc. of the 10th IEEE International Conference on Computer Vision*, vol. 1, pp. 198–203, 2005.