Facial Feature Extraction by Kernel Independent Component Analysis

T. Martiriggiano, M. Leo, P.Spagnolo, T. D'Orazio
Istituto di Studi sui Sistemi Intelligenti per l'Automazione - C.N.R.
Via Amendola 122/D-I, 70126 Bari, ITALY
{martiggiano,leo, spagnolo, dorazio }@ba.issia.cnr.it

Abstract

In this paper, we introduce a new feature representation method for face recognition. The proposed method, referred as Kernel ICA, combines the strengths of the Kernel and Independent Component Analysis (ICA) approaches. For performing Kernel ICA, we employ an algorithm developed by F. R. Bach and M. I. Jordan. This algorithm has proven successful for separating randomly mixed auditory signals, but it has never been applied on bidimensional signals such as images. We compare the performance of Kernel ICA with classical algorithms such as PCA and ICA within the context of appearance-based face recognition problem using the FERET and ORL databases. Experimental results show that both Kernel ICA and ICA representations are superior to representations based on PCA for recognizing faces across days and changes in expressions.

1. Introduction

Face recognition has become one of most important biometrics technologies during the past 20 years. It has a wide range of applications such as identity authentication, access control, and surveillance.

Human face image appearance has potentially very large intra-subject variations due to 3D head pose, illumination, facial expression, occlusion due to other objects or accessories (e.g., sunglasses, scarf, ect.), facial hair, and aging. On the other hand, the inter-subject variations are small due to the similarity of individual appearances. This makes face recognition a great challenge. Two issues are central: 1) what features to use to represent a face and 2) how to classify a new face image based on the chosen representation. This work focuses on the issue of feature selection. The main objective is to find techniques that can introduce low-dimensional feature representation of face objects with enhanced discriminatory power. Among various solutions to the problem (see [1] for a survey), the most successful are the appearance-based approaches, which generally operate directly on images or appearances of face objects.

Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) are two powerful tools largely used for data reduction and feature extraction in the appearance-based approaches. Two face recognition methods, Eigenfaces [2] and Fisherfaces [3], built on the two techniques respectively, have been proved to be very successful.

Independent Component Analysis (ICA) is a generalization of PCA which separates the high-order moments of the input in addition to the second-order moments utilized by PCA. Bartlett *et al.* [4] provided two architectures based on ICA, statistically independent basis images and a factorial code representation, for the face recognition task. Both ICA representations were superior to representations based on PCA for recognizing faces across days and changes in expression. A classifier [5] that combined the two ICA representation gave the best performance.

Although successful in many cases, linear methods fail to deliver good performance when face patterns are subject to large variations due to 3D head pose, illumination, facial expression, and aging. The limited success of these methods should be attributed to their linear nature. As a result, it is reasonable to assume that a better solution to this inherent nonlinear problem could be achieved using non linear methods, such as the so-called kernel machine techniques [6][7].

A kernel Principal Component Analysis, recently proposed as a nonlinear extension of a PCA [8][9], computes the principal component in a high-dimensional *feature space F*, which is nonlinearly related to the input space. Kim *et al.* [10] adopted a kernel PCA as a mechanism for extracting facial information. Through the use of a polynomial kernel, higher order correlations can be utilized between input pixels in the analysis of facial images. Using SVMs as the recognizer, experimental results with the ORL database [11], where the images vary in expression and pose, showed the effectiveness of the proposed method.

Yang [12] investigated the use of Kernel PCA and Kernel Fisher Linear Discriminant for learning low dimensional representations for face recognition, which he called Kernel Eigenface and Kernel Fisherface methods. He compared the performance of kernel methods with classical algorithms such as Eigenface, Fisherface, and ICA within the context of appearance-based face recognition problem using two data sets where images vary in pose, scale, lighting and expressions. Experimental results showed that kernel methods provided better representations and achieved lower error rates for face recognition.

In [13], Lu *et al.* presented a kernel machine based Discriminant analysis method, which combines kernel-based methodologies with Discriminant analysis techniques. The new algorithm has been tested, in terms of error rate performance, on the multi-view UMIST Face Database [14]. Results indicated that the proposed methodology outperform other commonly used approaches, such as the Kernel PCA.

Recently, Liu *et al.* [15] proposed a nonlinear ICA to model face appearance, which combines the nonlinear kernel trick with ICA. First, the kernel trick was employed to project the input image data into a high-dimensional implicit feature space *F* with a nonlinear polynomial mapping, and then the InfoMax algorithm[18] was performed in *F* to produce nonlinear independent components of input data. They proved the effectiveness of the approach on a test set of the FERET database containing people with the same expression and acquired in the same session.

In this paper we investigate a Kernel ICA approach proposed by F. R. Bach and M. I. Jordan [16] for the face recognition problem. The considered Kernel ICA approach is not based on a single nonlinear function (as the one proposed in [15]) but on an entire function space of candidate nonlinearities. The advantages of this approach over single nonlinear function, widely shown to estimate Independent Component and separate randomly mixed auditory signals, gave us the motivation for our choice.

We compare the performance of Kernel ICA with classical algorithms such as PCA and ICA within the context of appearance-based face recognition problem using two different databases: FERET and ORL. In order to test the effectiveness of the considered algorithms, different experiments of recognitions were carried out on three test set of the FERET database: 1) same session and different expressions of each individual, 2) different day and same expressions; 3) different day and different expressions. The ORL database contains images taken at different times, varying the light condition, the facial expressions (with some side movements) and also the facial details (glasses/no glasses). Results showed that both Kernel ICA and ICA representations were superior to representations based on PCA especially when there were great variations across days and changes in expressions.

2. Independent Components Analysis

Independent Component Analysis (ICA) is a computational technique for revealing hidden factors that underlie sets of measurements or signals.

ICA assumes a statistical model whereby the observed multivariate data, typically given as a large database of samples, are assumed to be linear or nonlinear mixtures of some unknown latent variables. The mixing coefficients are also unknown. The latent variables are nongaussian and mutually independent, and they are called the independent components of the observed data. By ICA, these independent components, also called sources or factors, can be found.

In other words an observed data vector $y = (y_1,\dots,y_m)$ is modelled by ICA as

$$y = Ax \qquad (1)$$

where x is a latent vector with independent components, and where A is the m×m matrix of mixing parameters.

Given N independently, identically distributed observations of y, Independent Component Algorithms estimate the mixing matrix A and thereby they recover the latent vector x corresponding to any particular y by solving equation 1.

ICA techniques are usually performed by introducing proper contrast functions and relative iterative procedure able to optimise them. A considerable portion of open literature is dedicated to define contrast functions associated with the estimation of the mixing matrix A by the Maximum Likelihood principle (ML) [17] $p(x; A, q) = |\det(A)|^{-1}\, q(A^{-1}x)$ or by minimizing the mutual information between the components of the estimated latent variables $\hat{x} = \hat{A}^{-1} y$ [18][19]. Alternative contrast functions derived as expansion based approximations of the mutual information have been also proposed [20,21].

The class of algorithms for Independent Component Analysis (ICA), recently proposed by F. R. Bach and M. I. Jordan, uses instead contrast functions based on canonical correlations in a reproducing kernel Hilbert space.

Given a reproducing-kernel Hilbert space (RKHS) F [22], we define the F-correlation as the maximal correlation between the random variables $f_1(x_1)$ and $f_2(x_2)$, where f_1 and f_2 range over F:

$$\rho_F = \max_{f_1,f_2 \in F} \operatorname{corr}(f_1(x_1), f_2(x_2)) = \max_{f_1,f_2 \in F} \frac{\operatorname{cov}(f_1(x_1), f_2(x_2))}{(\operatorname{var} f_1(x_1))^{1/2} (\operatorname{var} f_2(x_2))^{1/2}} \qquad (2)$$

If the variables x_1 and x_2 are independent, then the F-correlation is equal to zero. Moreover, if the set F is large enough, the converse is also true. In [16], Bach and Jordan show that the converse is also true for the reproducing kernel Hilbert spaces based on Gaussian kernels. In particular, they define their first contrast function as $I_{\rho_F} = -\frac{1}{2}\log(1-\rho_F)$.

Their converse result implies that I_{ρ_F} is a valid contrast function; a function that is always nonnegative and equal to zero if and only if the variables x_1 and x_2 are independent.

The authors show that ρ_F can be interpreted in term of linear projections and they derived a computationally efficient algorithm starting from the Standard Canonical Correlation Analysis algorithm [23].

3. Image data

The face images employed for this research are two subset of the FERET [24] and ORL [11] face databases.

The FERET dataset contain images of 38 individuals. There are four frontal views of each individual: A neutral expression and a change of expression from one session, and a neutral expression and change of expression from a second session that occurred three weeks after the first. Examples of the four views are shown in fig. 1.

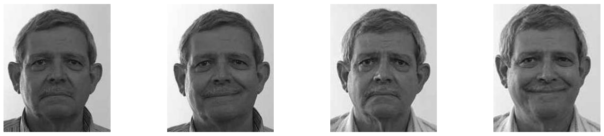

Fig. 1. Example from the FERET database of the four frontal image viewing conditions: neutral expression and change of expression from session 1; neutral expression and change of expression from session 2. Reprinted with permission from Jonathan Phillips.

The algorithms are trained on a single frontal view of each individual. The training set is comprised of 50% neutral expression images and 50% change of expression images. The algorithms are tested for recognition under three different conditions: same session, different expression (Test Set 1); different day, same expression (Test Set 2); and different day, different expression (Test Set 3).

Coordinates for eye and mouth locations are provided with the FERET database. These coordinates are used to center the face images, and then crop and scale to 240×200 pixels. Scaling is based on the area of triangle defined by the eyes and mouth. At last, we apply the Histogram Equalization to improve the contrast.

The ORL database contains ten different images of each of 40 distinct subjects. For some subjects, the images were taken at different times, varying the lighting, facial expressions (open / closed eyes, smiling / not smiling) and facial details (glasses / no glasses). All the images were taken against a dark homogeneous background with the subjects in an upright, frontal position (with tolerance for some side movement). Fig. 2 shows some examples of the ORL images.

All the 400 images from the ORL database are used to evaluate the face recognition performance of our Kernel ICA method. Five images are randomly chosen from the ten images available for each subject for training, while the remaining five images (unseen during training) are used for testing. In particular, fig. 2 shows in the top two rows the examples of training images used in our experiments, and in the bottom two rows the examples of test images.

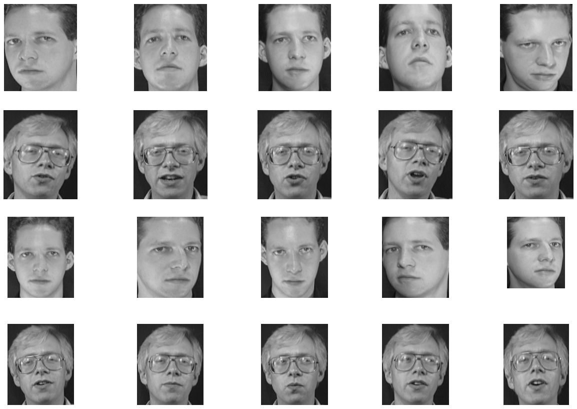

Fig. 2. Example ORL images. In particular, the above figure shows in the top two rows the examples of training images used in our experiments, and in the bottom rows the examples of test images.

4. A factorial face code

We use Kernel ICA to find a representation in which the coefficients used to code images are statistically independent, i.e., a factorial face code. Barlow and Atick discussed advantages of factorial codes for encoding complex objects that are characterized by high-order combinations of features [25], [26].

To archive this goal, we organize the data matrix X so that rows represent different pixels and columns represent different images. Performing Kernel ICA on X, we find a matrix a W such that the columns of $U = WX$ are the ICA representation of training images.

The representation code for test images is obtained by

$$WX_{test} = U_{test}, \tag{3}$$

where X_{test} is the matrix of test images.

Let N denote the number of training images, and let m denote the number of pixels. In our experiments and in most image recognition applications, especially in biometric ones, the number of training examples is limited ($N<<m$). Unfortunately, Kernel ICA operates well only if $m<<N$. Therefore, in order to reduce the dimensionality of the input, Kernel ICA is performed on

the first M ($M<<m$) PCA coefficients of the face images. The representation for the training images is therefore contained in the columns of $U = WR_M$, where $R_M = E^T X$ and E is the matrix containing the first M PCA axes in its columns.

The Kernel ICA weight matrix W is $M \times M$. The representation for test images is obtained in the columns of U_{test} as follows:

$$U_{test} = WR_{test}, \tag{4}$$

where $R_{test} = E^T X_{test}$.

The basis images for this representation consist of the columns of $A = W^{-1}$.

5. Experimental results

Face recognition performance is evaluated by the nearest neighbor algorithm (to the mean), which is defined as follows:

$$\delta\left(b_{test}, M_k^0\right) = \min_j \delta\left(b_{test}, M_j^0\right) \rightarrow b_{test} \in \omega_k, \tag{5}$$

where M_k^0 is the mean of the training samples for class ω_k. Therefore, the image feature vector b_{test} is classified to the class of the closest mean M_k^0 based on the similarity measure δ. Similarity measures used in our experiments are the Euclidean distance measure δ_{euc} and the cosine similarity measure $\delta_{\cos}$, which are defined as follows:

$$\delta_{euc}\left(b_{test}, b_{train}\right) = \sqrt{\sum_i \left(b_{test_i} - b_{train_i}\right)^2} \tag{6}$$

$$\delta_{\cos}\left(b_{test}, b_{train}\right) = \frac{-b_{test}{}^T \cdot b_{train}}{\left\|b_{test}\right\| \left\|b_{train}\right\|}, \tag{7}$$

where $\|.\|$ denotes the norm operator.

1) Experiments using the FERET dataset: The first set of experiments is carried out using the FERET dataset. Figures 3 and 4 report the face recognition performances with the Kernel ICA, ICA factorial code representations (for performing ICA, we employ the FastICA algorithm developed by A. Hyvärinen [27]) and PCA representations (the eigenface representation used by Pentland *et al*. [2]) using 36 PCA coefficients. In figure 3 and 4 the performances have been evaluated with the δ_{Euc} and the $\delta_{\cos}$ similarity measures, respectively. The first 36 PCs account for over 98,7% of the variance in the images.

There is a trend for the Kernel ICA and ICA representation to give superior face recognition performance to the PCA representation. The difference in performance is statistically significant for Test Set 2 and Test Set 3, when the test and training images differ not only in expression but also in lighting, scale and the date on which they were taken. Therefore, the high-order relationships among pixels, estimated by Kernel ICA and ICA, improve notably the performance when the face recognition is more difficult.

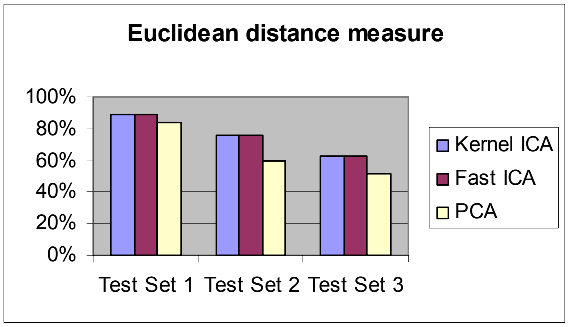

Fig. 3. Recognition performance of the Kernel ICA, ICA factorial code representations and PCA representations using 36 coefficients corresponding to the δ_{Euc} similarity measure.

2) Experiments using the ORL dataset: The second set of experiments is carried out using the ORL dataset. Figures 5 and 6 give the face recognition performances with the Kernel ICA, ICA factorial code representations and PCA representation using 30 (41% of the variance), 60 (58% of the variance), 120 (80,7% of the variance), 180 (96,2% of the variance) PC coefficients. Also in these experiments the two δ_{Euc} and $\delta_{\cos}$ similarity measures have been used to evaluate the performances of figures 5 and 6 respectively.

These experimental results lead to the same finings obtained using the FERET: both Kernel ICA and ICA representations are superior to representations based on PCA and there isn't a great difference in the performances of the Kernel ICA and ICA representations.

The lack of a substantial difference between the performances of the Kernel ICA and ICA algorithms, as found in their mono-dimensional applications, is probably due to the PCA preprocessing which is necessary in order to reduce the dimensionality of the data. In our opinion, the new orthogonal representation of the data provided by PCA precludes the kernel methods to improve their ability of represent the knowledge. In other words the evaluation of ICA produces the same results if it is applied directly after PCA or after a further transformation of PCA in a non-linear space (kernel method).

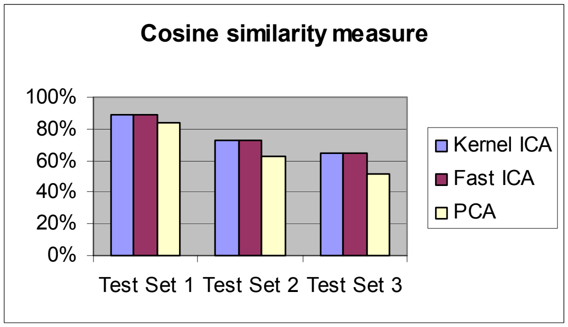

Fig. 4. Recognition performance of the Kernel ICA, ICA factorial code representations and PCA representations using 36 coefficients corresponding to the $\delta_{\cos}$ similarity measure.

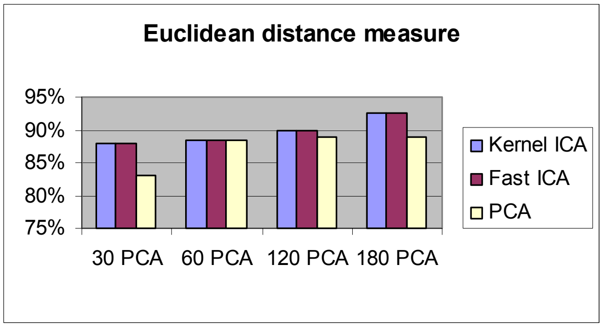

Fig. 5. Recognition performance of the Kernel ICA, ICA factorial code representations and PCA representations using 30, 60, 120, 180 coefficients corresponding to the δ_{Euc} similarity measure.

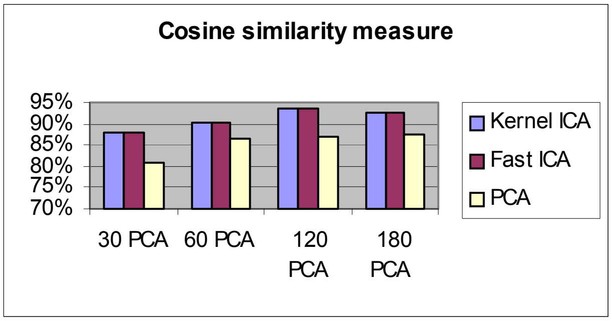

Fig. 6. Recognition performance of the Kernel ICA, ICA factorial code representations and PCA representations using 30, 60, 120, 180 coefficients corresponding to the $\delta_{\cos}$ similarity measure.

5. Conclusion and future work

A new method for face recognition has been introduced in this paper. The proposed method combines the strengths of the Kernel and Independent Component Analysis (ICA) approaches. Experiments results indicate that the performance of Kernel ICA is superior to that obtained by PCA, but are quite the same of those obtained by ICA for recognizing faces across days and changes in expression.

Future work will be addressed to the experimentation of different preprocessing techniques for feature reduction in order to assess if the Kernel ICA representation, without PCA, will increase the performances in more difficult situations.

6. References

[1] W. Zhao, R. Chellappa, A. Rosenfeld, P.J. Phillips, "Face Recognition: A literature survey", *Technical Report CART-TR-948*. University of Maryland, Aug. 2002.

[2] M.A. Turk, A.P. Pentland, "Eigenfaces for recognition", *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[3] P.N. Belhumeur, J.P. Hespanha, D.J. Kriegman, "Eigenfaces vs. Fisherfaces: recognition using class specific linear projection" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.

[4] M.S. Bartlett, H.M. Lades, T.J. Sejnowski, "Independent component representations for face recognition", In *Proc. of SPIE Conf. Human Vision and Electronic Imaging III*, vol. 3299, pp. 528-539, 1998.

[5] M.S. Bartlett, H.M. Lades, T.J. Sejnowski, "Face Recognition by Independent Component Analysis" *IEEE Transactions on Neural Networks*, vol. 13, NO. 6, November 2002.

[6] A. Ruiz, P.E. López de Teruel, "Nonlinear kernel-based statistical pattern analysis" *IEEE Transactions on Neural Networks*, vol. 12, no. 1, pp. 16–32, January 2001.

[7] K.R. Müller, S. Mika, G. Rätsch, K. Tsuda, B. Schölkopf, "An introduction to kernel-based learning algorithms", *IEEE Transactions on Neural Networks*, vol. 12, no. 2, pp. 181–201, March 2001.

[8] B. Schölkopf, A. Smola, K. Müller, "Non-linear component analysis as a kernel eigenvalues problem", *Neural Comput*. vol. 10, pp. 1299-1319, 1998.

[9] B. Schölkopf, A. Smola, K. Müller, "Kernel principal component analysis", *Advances in Kernel Methods - Support Vector Learning*, 327-352. (Eds.) B. Schölkopf, C.J.C. Burges and A.J. Smola, MIT Press, Cambridge, MA (1999).

[10] K.I. Kim, K. Jung, H.J. Kim, "Face Recognition Using Kernel Principal Component Analysis", *IEEE Signal Processing Letters*, vol. 9, no. 2, February 2002.

[11] Web site of ORL face database: http://www.uk.research.att.com/facedatabase.html. AT&T Laboratories Cambridge.

[12] M.H. Yang, "Kernel eigenfaces vs. Kernel fisherfaces: Face Recognition using kernel methods", In *Proc. 5th Int. Conf.*

Automat. Face Gesture Recognition, Washington, DC, May 2002, pp. 215-220.

[13] J. Lu, K.N. Plataniotis, A.N. Venetsanopoulos, “A kernel machine based approach for multi-view face recognition”, In *Proc. of the IEEE International Conference on Image Processing at Rochester*, New York, USA, September 22-25 2002.

[14] D. Graham, N. Allinson, Web site of umist multi-view face database: http://images.ee.umist.ac.uk/danny/database.html. Image Engineering and Neural Computing Lab, UMIST, UK, 1998.

[15] Q. Liu, J. Cheng, H. Lu, S. Ma, “Modeling face appearance with nonlinear independent component analysis”, *In Proc. Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, 17-19 May 2004.

[16] F.R. Bach, M. I. Jordan, “Kernel Independent Component Analysis”, *J. Machine Learning Res.*, vol. 3, pp. 1-48, 2002.

[17] J.F. Cardoso, “Blind Signal Separation: Statistical Principles”, *Proc. of the IEEE*, vol. 9, no. 10, pp. 2009-2026, October 1998.

[18] A.J. Bell, T.J. Sejnowski,, “An information-maximization approach to blind separation and blind deconvolution”, *Neural Computation*, 7:1129-1159, 1995.

[19] J.F. Cardoso, “High order contrasts for ICA”, *Neural Computation*, 11(1):157-192, 2001.

[20] A. Hyvärinen, E. Oja, “Independent Component Analysis: Algorithms and Applications”, *Neural Networks*, 13(4-5):411-430, 2000.

[21] S. Amari, A. Cichocki, H.H. Yang, “A new learning algorithm for blind signal separation” In *adv. In NIPS*, 8, 1996.

[22] S. Saitoh, “Theory of Reproducing Kernels and its Applications Harlow”, *Longman Scientific & Technical*, UK, 1988.

[23] T.W. Anderson, “An Introduction to Multivariate Statistical Analysis”, *A Wiley publication in mathematical statistics*. New York: John Wiley & Sons, 1958.

[24] P.J. Phillips, H. Moon, S.A. Rizvi, P.J. Rauss, “The FERET evaluation methodology for face-recognition algorithms”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, pp. 1090–1104, 2000.

[25] H.B. Barlow, “Unsupervised learning”, *NeuralComput*, vol. 1, pp. 295-311, 1989.
[26] J.J. Atick, “Could information theory provide an ecological theory of sensory processing?” *Network*, vol. 3, pp. 213-251, 1992.

[27] A. Hyvärinen, “Fast and Robust Fixed-Point Algorithms for Independent Component Analysis”, *IEEE Transactions on Neural Networks*, 10(3):626-634, 1999.