Normalized Radial Basis Function Networks and Bilinear Discriminant Analysis for Face Recognition

Muriel Visani, Christophe Garcia, Jean-Michel Jolion

Normalized Radial Basis Function Networks and Bilinear Discriminant Analysis for Face Recognition

Muriel Visani, Christophe Garcia
France Telecom Division R&D TECH / IRIS
4, rue du Clos Courtel
35512 Cesson-Sevigne, France
{muriel.visani,christophe.garcia}@rd.francetelecom.com

Jean-Michel Jolion
Laboratoire LIRIS, INSA Lyon
20, Avenue Albert Einstein
Villeurbanne, 69621 cedex, France
Jean-Michel.Jolion@insa-lyon.fr

Abstract

In this paper, we present a novel approach for face recognition using a new subspace method called Bilinear Discriminant Analysis (BDA), and Normalized Radial Basis Function Networks (NRBFNs). In a first step, BDA extracts the features that enhance separation between classes by using a generalized bilinear projection-based Fisher criterion, computed from image matrices directly. In a second step, the features are fed into a NRBFN that learns class conditional probabilities. This results in an efficient and computationally simple open-world identification process. Experimental results assess the performance and robustness of the proposed algorithm compared to other subspace methods combined withs NRBFNs, in the presence of variations in head poses, facial expressions, and partial occlusions.

1. Introduction

During the last decade, automatic recognition of human faces has grown into a key technology, especially in the fields of multimedia indexing and security. In this context, the views of the face to recognize can differ drastically from the training set, especially in head poses, facial expressions and partial occlusions, which makes face recognition a difficult task.

Like any other pattern recognition task, face recognition can be basically defined as a two-step process: feature extraction and classification. In most surveillance applications, the features classification consists in 1) checking whether the corresponding face is registered in a database of known faces or not; and possibly 2) assigning an identity to that face. Since the seminal work of Sirovich and Kirby [13], statistical projection-based methods have been widely used for facial representation. In the Eigenfaces [14] and Fisherfaces [1] methods, a costly and potentially unstable statistical analysis is applied to the high-dimensional image vector space. To overcome these drawbacks, Yang *et al.* [18] proposed the Two Dimensional PCA (2DPCA) method, performing PCA using directly the face image matrices. It has been shown that 2DPCA is more efficient [18] and more robust [15] than the eigenfaces. Recently, we proposed the Two-Dimensional-Oriented Linear Discriminant Analysis (2DoLDA) approach [16], consisting in applying LDA on image matrices and outperforming the 2DPCA and Fisherfaces methods. The first contribution of this paper is a new feature extractor named Bilinear Discriminant Analysis (BDA), generalizing and outperforming 2DoLDA. This method is based on the optimization of a generalized Fisher criterion, relying on bilinear projections that are computed from image matrices directly.

The second contribution of this paper is the use of Normalized Radial Basis Function Networks (NRBFs) [5, 10] for the features classification. Indeed, as they generally provide better results in high dimensional problems [10, 12], and offer better generalization than the traditional RBFNs [2, 10, 6, 17], NRBFNs are particularly convenient for face recognition. Moreover, as they provide class conditional probabilities, NRBFNs can be easily used to check whether a person is registered or not.

The remainder of the paper is organized as follows. In section 2, we describe in details the principle and the algorithm of the proposed BDA method. In section 3, we present the classifier based on Normalized Radial Basis Functions Neural Networks. Then, in section 4, we present a series of two experiments, demonstrating the effectiveness and the robustness of the proposed algorithm and comparing its performances with respect to other subspace methods combined with NRBFNs. Finally, conclusions and closing remarks are drawn in section 5.

2 Bilinear Discriminant Analysis

The model is constructed from a training set containing N face image matrices X_l, of size $h \times w$. The set of images corresponding to one person is called a *class*. Each class contains multiple views. The training set contains k classes

$(\Omega_j)_{j=\{1...k\}}$ of registered faces. Let us consider two projection matrices $Q \in \mathbb{R}^{h\times g}$ and $P \in \mathbb{R}^{w\times g}$. Let us define the signature $X_l^{Q,P}$ of the sample X_l by its bilinear projection onto (Q, P):

$$X_l^{Q,P} = Q^T X_l P \tag{1}$$

We are searching for the optimal pair of matrices (Q^*, P^*) maximizing the separation between signatures from different classes while minimizing the separation between signatures from the same class:

$$(Q^*, P^*) = \underset{(Q,P)\in\mathbb{R}^{h\times g}\times\mathbb{R}^{w\times g}}{\texttt{Argmax}} \frac{|S_b^{Q,P}|}{|S_w^{Q,P}|} \tag{2}$$

$$= \texttt{Argmax} \frac{|\sum_{j=1}^{k} n_j(\overline{X_j^{Q,P}} - \overline{X^{Q,P}})^T(\overline{X_j^{Q,P}} - \overline{X^{Q,P}})|}{|\sum_{j=1}^{k}\sum_{X_l\in\Omega_j}(X_l^{Q,P} - \overline{X_j^{Q,P}})^T(X_l^{Q,P} - \overline{X_j^{Q,P}})|} \tag{3}$$

where $S_w^{Q,P}$ and $S_b^{Q,P}$ are respectively the within-class and between-class covariance matrices of the set $(X_l^{Q,P})_{l\in\{1,...,N\}}$, $\overline{X_j^{Q,P}}$ is the mean of all the projected samples belonging to class j, and $\overline{X^{Q,P}}$ is the mean of all the projected samples of the training set.

The objective function given in equation (3) is biquadratic and has no analytical solution. We therefore propose an iterative procedure that we call *Bilinear Discriminant Analysis*. Let us expand the expression (3):

$$(Q^*, P^*) = \texttt{Argmax} \frac{|\sum_{j=1}^{k} n_j P^T(\overline{X_j} - \overline{X})^T Q Q^T(\overline{X_j} - \overline{X})P|}{|\sum_{j=1}^{k}\sum_{X_l\in\Omega_j} P^T(X_l - \overline{X_j})^T Q Q^T(X_l - \overline{X_j})P|} \tag{4}$$

Thus, for any fixed $Q \in \mathbb{R}^{h\times g}$, the objective function (3) can be rewritten:

$$P^* = \underset{P\in\mathbb{R}^{w\times g}}{\texttt{Argmax}} \frac{|P^T\left[\sum_{j=1}^{k} n_j(\overline{X_j^Q} - \overline{X^Q})^T(\overline{X_j^Q} - \overline{X^Q})\right]P|}{|P^T\left[\sum_{j=1}^{k}\sum_{X_l\in\Omega_j}(X_l^Q - \overline{X_j^Q})^T(X_l^Q - \overline{X_j^Q})\right]P|} \tag{5}$$

$$= \underset{P\in\mathbb{R}^{w\times g}}{\texttt{Argmax}} \frac{|P^T S_b^Q P|}{|P^T S_w^Q P|} \tag{6}$$

with S_w^Q and S_b^Q being respectively the *generalized within-class covariance matrix* and the *generalized between-class covariance matrix* of the set $(X_l^Q)_{l\in\{1...N\}}$, where $X_l^Q = Q^T \cdot X_l$. Therefore the columns of the matrix P^* are the g eigenvectors of ${S_w^Q}^{-1}S_b^Q$ with largest eigenvalues. If $Q = I_h$, the identity matrix of size $h \times h$, P^* is the projection matrix of Row-Oriented Linear Discriminant Analysis (RoLDA), a version of 2DoLDA [16].

Let us denote $A = P^T(\overline{X_j} - \overline{X})^T Q$, matrix of size $g \times g$. Given that, for every square matrix A, $|A^T A| = |AA^T|$, the objective function (3) can be rewritten:

$$(Q^*, P^*) = \texttt{Argmax} \frac{|\sum_{j=1}^{k} n_j Q^T(\overline{X_j} - \overline{X})PP^T(\overline{X_j} - \overline{X})^T Q|}{|\sum_{j=1}^{k}\sum_{X_l\in\Omega_j} Q^T(X_l - \overline{X_j})PP^T(X_l - \overline{X_j})^T Q|} \tag{7}$$

For any fixed $P \in \mathbb{R}^{w\times g}$, using equation (7) the objective function (3) can be rewritten $Q^* = \underset{Q\in\mathbb{R}^{h\times g}}{\texttt{Argmax}} \frac{|Q^T\Sigma_b^P Q|}{|Q^T\Sigma_w^P Q|}$, Σ_w^P and Σ_b^P being the generalized within-class and between-class covariance matrices of the set $((X_l^P)^T)_{l\in\{1...N\}}$, where $X_l^P = X_l \cdot P$. Therefore, the columns of Q^* are the g eigenvectors of $(\Sigma_w^P)^{-1}\Sigma_b^P$ with largest eigenvalues. If $P = I_w$, the matrix Q^* is the projection matrix of Column-Oriented Linear Discriminant Analysis (CoLDA) [16].

We can note that BDA leads to a significant reduction in the dimensionality of the signatures compared to 2DPCA and 2DoLDA: the size of a signature using BDA is g^2, versus $h \cdot g$ for RoLDA and 2D PCA, and $w \cdot g$ for CoLDA.

The algorithm of the BDA approach is as follows: let us initialize $P_0 = I_w$ and $\alpha_0 = 0$. The number of feature vectors g is fixed (see section 4).

While $\alpha_t < \tau$

1. For $l \in \{1, \ldots, N\}$, compute $X_l^{P_t} = X_l \cdot P_t$.
2. Compute $\Sigma_w^{P_t}$, $\Sigma_b^{P_t}$ and $(\Sigma_w^{P_t})^{-1} \cdot \Sigma_b^{P_t}$;
3. Compute Q_t, whose columns are the first g eigenvectors of $(\Sigma_w^{P_t})^{-1} \cdot \Sigma_b^{P_t}$;
4. For $l \in \{1, \ldots, N\}$, compute $X_l^{Q_t} = (Q_t)^T \cdot X_l$.
5. Compute $S_w^{Q_t}$, $S_b^{Q_t}$, and $(S_w^{Q_t})^{-1} \cdot S_b^{Q_t}$;
6. Compute P_t, whose columns are the first g eigenvectors of $(S_w^{Q_t})^{-1} \cdot S_b^{Q_t}$;
7. Compute $\alpha_t = \sqrt{(\|P_t - P_{t-1}\|_2^2 + \|Q_t - Q_{t-1}\|_2^2}$.

The parameter τ can be determined empirically. However, experimental results have shown that after one iteration the recognition results are satisfying. Therefore, in the following, we will use the preceding algorithm with only one iteration, freeing us from determining τ.

3. Normalized RBF Network

Radial Basis Function Networks can be used for classification as a two-layer neural network implementing a mapping function $\mathbb{R}^n \rightarrow \mathbb{R}^k$, where n is the size of the input signal and k is the number of classes. The two layers are: the RBF layer, and the output layer. For any input signal X_l, the Euclidean distance to the center of each of the r RBF units is computed, and then an activation function R is applied. Although there are many possible activation functions, the Gaussian function is preferred for high-dimensional data [12]: for each input observation X_l, the output of the i^{th} RBF is given by:

$$R_{i\,l} = e^{-\frac{1}{2\sigma_i^2}(X_l - C_i)^T(X_l - C_i)} \tag{8}$$

where C_i and σ_i are respectively the center and the width of the i^{th} RBF. The k output units are linear: if W_0 denotes the bias of the system, the response of the j^{th} output is:

$$Y_{j\,l} = W_0 + \sum_{i=1}^{r} W_{j\,i} \cdot R_{i\,l} \tag{9}$$

The system constructed from (9) can be rewritten:

$$Y = W \cdot R \tag{10}$$

where the elements of the matrix $W \in \mathbb{R}^{g \times r+1}$ are the $W_{j\,i}$ and, $\forall j \in \{1 \dots k\}, W_{j\,0} = W_0$. The matrix $R \in \mathbb{R}^{r+1 \times N}$ contains the elements $R_{i\,l}$, and $\forall l \in \{1 \dots N\}, R_{0\,l} = 1$. The elements of $Y \in \mathbb{R}^{k \times N}$ are the $Y_{j\,l}$.

The specificity of the Normalized RBF Network, which architecture is illustrated in Figure 1, is that the output of each RBF unit is normalized by the total activity of the RBF layer. The outputs of the NRBFs are normalized by the total activity of the RBF layer, and therefore if we consider that $R_{i\,l} = \mathbb{P}[i/X_l]$ and using Bayes' theorem, the NRBF outputs are the posterior probabilities of X_l belonging to the i^{th} NRBF hypersphere:

$$\mathbb{P}[i/X_l] = \frac{R_{i\,l}}{\sum_{i=0}^{r} R_{i\,l}} \tag{11}$$

The j^{th} output of the overall network can be viewed as the posterior probability of X_l belonging to class Ω_j:

$$Y_{j\,l} = \mathbb{P}[\Omega_j/X_l] = \sum_{i=0}^{r} W_{j\,i} \cdot \mathbb{P}[i/X_l] \tag{12}$$

and the weights can be expressed as: $W_{j\,i} = \mathbb{P}[\Omega_j/i]$.

Initialization of the Normalized RBFs. The number, positions and widths of the RBFs have a great influence on the performance, and depend on the geometrical properties of the training features and the type of the activation function [3]. There are basically three classes of initialization strategies for RBFNs: 1) RBFNs with a fixed number of RBFs whose centers are selected randomly from the training set [4]; 2) RBFNs with a fixed number of RBFs whose centers are selected using unsupervised clustering techniques [9] and 3) RBFNs using supervised procedures for selecting a fixed number of RBFs centers [11, 6]. The first class of methods is poorly performing in high-dimensional spaces, and a major problem with unsupervised techniques is that in some cases they can converge to a poor local optimum [8]. We therefore evaluated two supervised initialization methods.

We first tested the initialization paradigm proposed in [6]: initially, each class is associated to one RBF, whose hypersphere contains all the samples from the associated class and further defines a cluster. Then we check the following two criteria: 1) *embody criterion*: if cluster j is embodied in cluster k, then cluster k is splitted into two clusters; 2) *misclassification criterion*: if cluster j contains many data from cluster k, then cluster j must be splitted into two clusters. This splitting procedure is repeated until no cluster meet one of the above criteria.

We also tested a simple supervised initialization technique, consisting in assigning to each class a fixed number of RBFs. If the number of RBFs per class is set to one, the center C_i of the RBF corresponding to the j^{th} class Ω_j is the centroid of the set of the samples corresponding to class j, and its width σ_i is initialized to $\max_{X_l \in \Omega_j} \|X_l - C_i\|_2$. If more than one RBF per class are necessary, the additional RBFs are added randomly so that they are included into the active area of the initial RBF.

After using whatever initialization paradigm, the widths of the RBFs are adjusted, so as to provide a good compromise between specialization and generalization: the overlapping between the RBFs must be sufficient to enable good generalization, while a necessary separation between different classes must be kept to avoid ambiguity during classification. Let us denote $d_i^W = \max_{X_l \in \omega_i} \|X_l - C_i\|_2$, where ω_i is the set of samples belonging to the hypersphere delimited by the i^{th} RBF, and $d_i^B = \min_{C_p \in \omega_i'} \|C_p - C_i\|_2$, where ω_i' is the set of the RBF centers, except C_i. The proposed adjustment, inspired from [6], is:

$$\sigma_i^W = \frac{d_i^W}{\sqrt{|\log(\beta)|}}, \qquad \sigma_i^B = \mu \cdot d_i^B \tag{13}$$

$$\sigma_i = \max(\sigma_i^W, \sigma_i^B) \tag{14}$$

where the parameter μ can be estimated as follows:

$$\mu \approx \frac{\sum_{j=1}^{k} \sigma_j^W}{\sum_{j=1}^{k} d_j^B} \tag{15}$$

The parameter $\beta \in [0.5, 1[$ is determined by the relative positions of the classes: the more largely scattered are the data, the smaller should be β.

Hybrid Learning Algorithm. In hybrid learning, the adjustment of the centers and widths of the RBFs is a nonlinear process while the weight learning is a linear one. It combines the gradient paradigm and the Linear Least Squares (LLS) optimization method. If we consider, for each training sample X_l, the pair of obtained and desired outputs $(Y_{\cdot\,l}, T_{\cdot\,l})$ (the matrix T contains only 'O's and '1's), we can jointly optimize the set of parameters of the RBFs $\{C_i, \sigma_i\}_{i=\{1 \dots r\}}$ and the weights W by minimizing the following cost function:

$$E = \sum_{l=1}^{N} E^l = \frac{1}{2} \sum_{l=1}^{N} (T_{\cdot\,l} - Y_{\cdot\,l})^T (T_{\cdot\,l} - Y_{\cdot\,l}) \tag{16}$$

At each epoch t, the centers and widths are adjusted for each sample X_l:

$$C_i = C_i - \xi_c \Delta C_i^l \quad (17)$$

$$\sigma_i = \sigma_i - \xi_\sigma \Delta \sigma_i^l \quad (18)$$

where ΔC_i^l and $\Delta \sigma_i^l$, whose associated learning rates are respectively ξ_c and ξ_σ, can be computed as follows:

$\forall\, m \in \{1 \dots n\}$,

$$\Delta C_i^l(m) = \frac{\partial E^l}{\partial C_{i\,m}} = -\frac{(X_{l\,m} - C_{i\,m})}{\sigma_i^2 \cdot (\sum_{i=1}^r R_{i\,l})^2} \cdot R_{i\,l} \cdot (\sum_{i=1}^r R_{i\,l} - R_{i\,l}) \cdot \sum_{j=1}^k W_{j\,i}(T_{j\,l} - Y_{j\,l}) \quad (19)$$

and

$$\Delta \sigma_i = \frac{\partial E^l}{\partial \sigma_i} = -\frac{\sum_{m=1}^n (X_{l\,m} - C_{i\,m})^2}{\sigma_i^3 \cdot (\sum_{i=1}^r R_{i\,l})^2} \cdot R_{i\,l} \cdot (\sum_{i=1}^r R_{i\,l} - R_{i\,l}) \cdot \sum_{j=1}^k W_{j\,i}(T_{j\,l} - Y_{j\,l}) \quad (20)$$

Once the parameters $\{C_i, \sigma_i\}$ have been settled, the NRBFN outputs $\mathbb{P}[i/X_l]$ can be computed from (11). Then, the weight matrix W can be determined by applying the LLS method based on the pairs $(Y_{\cdot\,l}, T_{\cdot\,l})_{l=\{1\dots N\}}$:

$$W = TR^T(RR^T)^{-1} \quad (21)$$

Checking whether a Face is Registered in the Database. For most surveillance system, an alert should be given only when a registered face is detected. Let us denote j_l^1 the most probable class. We propose the following criterion: X_l corresponds to a registered face (and further to class j_l^1) if and only if: $\mathbb{P}[j_l^1/X_l] > \tau_1$ (22) where the threshold τ_1 depends on the number of registered classes.

4. Experimental Results

In this section, we perform two series of experiments on a subset of the Asian Face Image Database PF01 [7] containing 60 people under neutral illumination conditions. The training set contains four views per person (see Figure 2(a)). Three classifiers are built from that training set: BDA+NRBFN, PCA+NRBFN and PCA+LDA+NRBFN. For BDA+NRBFN, each signature matrix $X_l^{Q,P}$ is concatenated to obtain a vector, which is considered as an input signal. The parameters of the three models are set to their optimal values. The BDA-based subspace is built from $g = 15$ feature vectors; the PCA subspace is built from $g = 150$ Eigenfaces and the PCA+LDA subspace is built from 150 Eigenfaces, and $g = 59$ Fisherfaces. According to the experimental results obtained when testing the different initialization schemes, NRBFNs are initialized by setting one NRBF per class with $\beta = 0.7$. The learning rates of the associated NRBFNs are $\xi_c = \frac{\|C\|_2}{n}$ and $\xi_\sigma = \frac{\|\sigma\|_2}{n}$; they are updated each 1000 epochs. We set $\tau_1 = 0.25$ and the false rejections are considered as misclassifications.

The first experiment is designed to evaluate these classifiers when the persons in the test set are registered in the training set, but with strong dissimilarities between the two sets. The first test set contains strong facial expression changes (see Figure 2(b)). The second test set contains one top right-head pose view per person (see Figure 2(c)). The third test set (see Figure 2(d)) contains one digitally modified image per person, where a black strip of width 25 pixels, simulating a scarf, has been added. The original image does not belong to the training set.

The experimental results show that the use of the NRBF Network does not improve the recognition results, compared to the Nearest Neighbour Classifier (NNC), in the PCA+LDA subspace. Moreover, the performances of PCA+LDA+NRBFN are highly dependent on the number of RBFs, and on their initialization; that is why they are not presented in this paper. Table 1 provides a comparison of the performances of PCA, PCA+LDA, 2DPCA, the two versions of 2DoLDA (namely CoLDA and RoLDA) and BDA, using NNC. From this table we can see that, as highlighted in [15], 2DPCA outperforms PCA, especially when dealing with partial occlusions. Table 1 also highlights the fact that BDA outperforms the other projection-based methods, when using NNC.

The Mean Square Errors (MSE) between the obtained and desired outputs Y and T, using BDA+NRBFN and PCA+NRBFN, are given in Figure 3(a–b). From Figure 3(a) we can see that the learning errors are initially comparable for the two methods; both decrease when the number of epochs increase, but the MSE for PCA+NRBFN decreases much more slowly than the MSE for BDA+NRBFN. Figure 3(b) shows that the MSE for PCA+NRBFN slightly increases with the number of epochs over the two first test sets, while the MSE for BDA+NRBFN significantly decreases. Therefore, BDA+NRBFN seems to have a better generalization power than PCA+NRBFN. The compared recognition rates on the three test sets, given in Figure 4(a–b), verifies that assumption. It should be noted that, for both methods and after a sufficient number of iterations, the false rejection rate is null among the three test sets, ie all the faces are correctly classified as registered in the training data.

When evaluating BDA+NRBFN on the "pose" test set, one misclassification and two false rejections occur; however, after 4800 epochs the recognition rate reaches 100%. Let us denote "$\mathbb{P}$[target]" the probability of belonging to the target class, and "$\mathbb{P}$[wrong]" the probability of belonging to the wrong class with highest associated probability. Figure 5(a) shows that after 4800 iterations, $\mathbb{P}$[target]>$\mathbb{P}$[wrong] and $\mathbb{P}$[target]>τ_1, therefore the sample is recognized. Figure 5(b-c) illustrates the two false rejections: in both cases,

after respectively 2450 and 1900 epochs, P[target] reaches $\tau_1 = 0.25$ while P[wrong] remains inferior to 0.25 and then the corresponding sample is recognized.

The first experiment results show 1) the superiority of BDA+NRBFN over PCA+NRBFN and PCA+LDA+NNC (and, consequently, over PCA+LDA+NRBFN) in the presence of variations in facial expression, head pose, and facial occlusion and 2) that, after a sufficient number of epochs, all the registered faces are correctly classified as registered.

The second experiment aims at testing whether the proposed classifier correctly classifies the unregistered people as unknown. The training set and the parameters of the model are the same as in the first experiment. The test set (see Figure 2(e)) contains one image of each of 15 unregistered persons, under the same illumination conditions as in the training set and near-frontal head poses. Figure 4(c) shows that BDA+NRBFN generates less false alarms than PCA+NRBFN, as it correctly classifies 93.3% of the faces as unregistered, versus 86.7% for PCA+NRBFN.

5. Conclusion

In this paper, we have presented a novel approach for face recognition, combining a new feature extractor named Bilinear Discriminant Analysis and Normalized Radial Basis Function Networks, that to the best of our knowledge have never been applied to face recognition. This method, by providing class conditional probabilities, can be used to check whether a person is registered in the training data and, possibly, assign an identity to that person. The experimental results assess the effectiveness and robustness of the proposed approach and show its superiority over the usual projection-based methods, combined with Normalized Radial Basis Function Networks and the Nearest Neighbour rule.

References

[1] P.N. Belhumeur, J.P. Hespanha, D.J. Kriegman, "Eigenfaces vs Fisherfaces : Recognition Using Class Specific Linear Projection". In *IEEE Trans. on Pattern Analysis and Machine Intelligence*. 19(7), pp. 711–720, 1997.

[2] C.M. Bishop, *Neural Networks for Pattern Recognition*, New York, NY: Oxford University Press, 1995.

[3] G. Bors, M. Gabbouj, "Minimal Topology for a Radial Basis Functions Neural Networks for Pattern Classification". In *Digital Processing*, 4, pp. 173–188, 1994.

[4] D.S. Broomhead, D. Lowe, "Multivariable Functional Interpolation and Adaptive Networks". In *Complex Systems*, 2, pp. 321–355, 1988.

[5] G. Bugmann, "Classification using Networks of Normalized Radial Basis Functions". In *Proc. of the Int. Conf. on Advances in Pattern Recognition ICAPR'98*, S. Singh, editor, Springer, London, pp. 435–444, 1998.

[6] M.J. Er, S. Wu, J. Lu, H.L. Toh, "Face Recognition With Radial Basis Function (RBF) Neural Networks". In *IEEE Trans. on Neural Networks*, 13(3), pp.697–709, 2002.

[7] B.W. Hwang, M.C. Roh, S.W. Lee, "Performance Evaluation of Face Recognition Algorithms on Asian Face Database". In *Proc. of the IEEE Int. Conference on Automatic Face and Gesture Recognition*, pp. 278–283, Seoul, Korea, 2004.

[8] M. Meila, D. Heckerman, "An Experimental Comparison of Several Clustering and Initialization Methods" In *Proc. of the Conference on Uncertainty in Artificial Intelligence UAI'98*, pp. 386-395, 1998.

[9] J. Moody, J. Darken, "Fast Learning in Networks of Locally-Tuned Processing Unit". In *Neural Computing*, vol. 1, pp. 281–294, 1989.

[10] O. Nelles, *Nonlinear System Identification: from Classical Approaches to Neural Networks and Fuzzy Models*, New York, NY: Springer, 2001.

[11] T. Poggio, F. Girosi, "Regularization Algorithms for Learning that are Equivalent to Multilayer Networks". In *Science*, 247, pp. 978–982, 1990.

[12] D. Potts, C. Sammut, "Online Nonlinear System Identification in High Dimensional Environments. In *Proc. of the Australasian Conf. on Robotics and Automation*, 2004.

[13] L. Sirovitch, M.Kirby, "A Low-Dimensional Procedure for the Characterization of Human Faces". In *Journal of Optical Society of America A*, 4(3), pp. 519–524, 1987.

[14] M.A. Turk, A.D. Pentland, "Eigenfaces for Recognition". *Journal of Cognitive Neuroscience*. **3**(1), pp. 71–86, 1991.

[15] M. Visani, C. Garcia, C. Laurent, "Comparing Robustness of Two-Dimensional PCA and Eigenfaces for Face Recognition". In *Proc. of the Int. Conf. on Image Analysis and Recognition ICIAR'04*. Springer LNCS 3211, A. Campilho, M. Kamel (eds), pp. 717–724, 2004.

[16] M. Visani, C. Garcia, J.M. Jolion, "Two-Dimensional-Oriented Linear Discriminant Analysis for Face Recognition". In *Proc. of the Int. Conf. on Computer Vision and Graphics ICCVG'04*. Computational Imaging and Vision Series, pp. 1008–1017, 2004.

[17] Y. Wang, Y. Jia, C. Hu, M. Turk, "Face Recognition Based on Kernel Radial Basis Function Networks". In Proc. of the *Asian Conference on Computer Vision*, Korea, 2004.

[18] J. Yang, D. Zhang, A.F. Frangi, "Two-Dimensional PCA: A New Approach to Appearance-Based Face Representation and Recognition". In *IEEE Trans. on Pattern Analysis and Machine Intelligence*. 26(1), pp. 131–137, 2004.

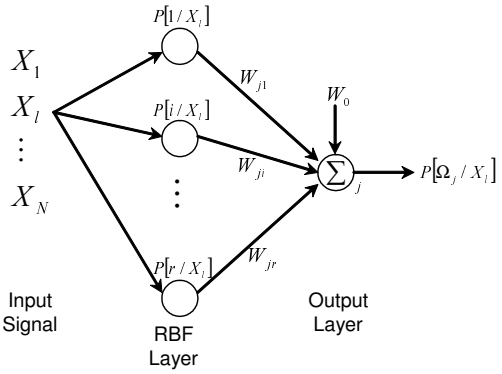

Figure 1: *Architecture of a NRBF Network.*

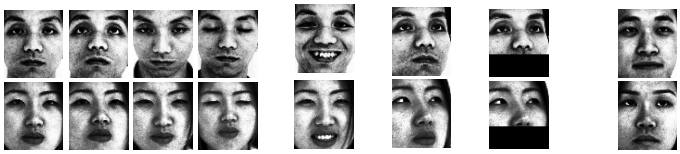

Figure 2: *(a): training set; (b) "expression" test set; (c) "pose" test set; (d) "occlusion" test set; (e): set of unregistered people.*

	PCA	PCA+ LDA	2DPCA	CoLDA	RoLDA	BDA
'pose'	96.7%	98.3%	96.7%	96.7%	98.3%	100%
'expression'	86.7%	95%	93.3%	95%	96.7%	96.7%
'occlusion'	13.3%	50%	55%	70%	53.3%	70%

Table 1: *Compared Recognition Rates of PCA, PCA+LDA, 2DPCA, CoLDA, RoLDA and BDA using the Nearest Neighbour Rule.*

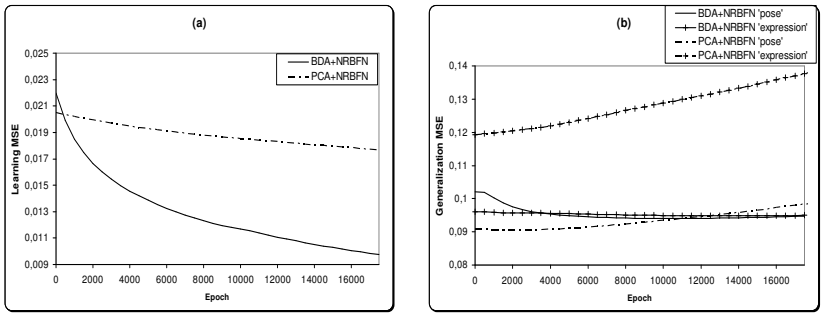

Figure 3*: *Compared Mean Square Errors of BDA+NRBFN (solid line) and PCA+NRBFN (dashed line) computed (a): from the training set; b): from the first two test sets, when varying the number of epochs.*

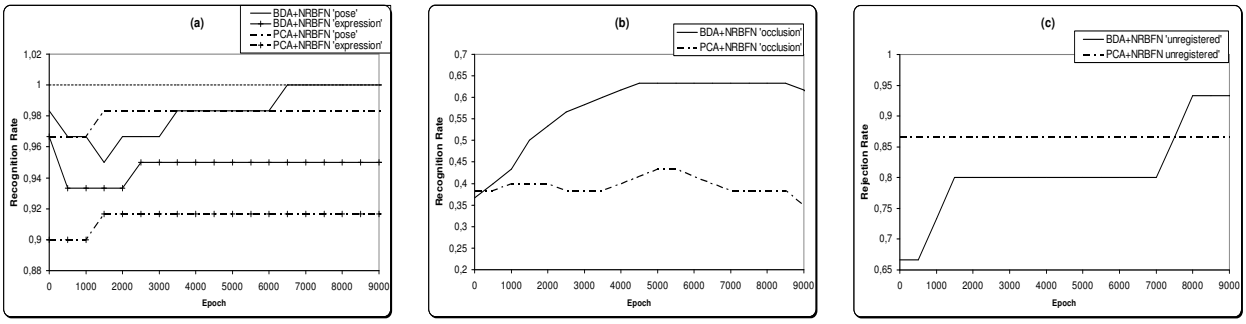

Figure 4*: *Evolution of the recognition rates of BDA+NRBFN and PCA+NRBFN computed: (a) from the first two test sets and (b) from the "occlusion" set; (c): rejection rates computed from the set of unregistered people.*

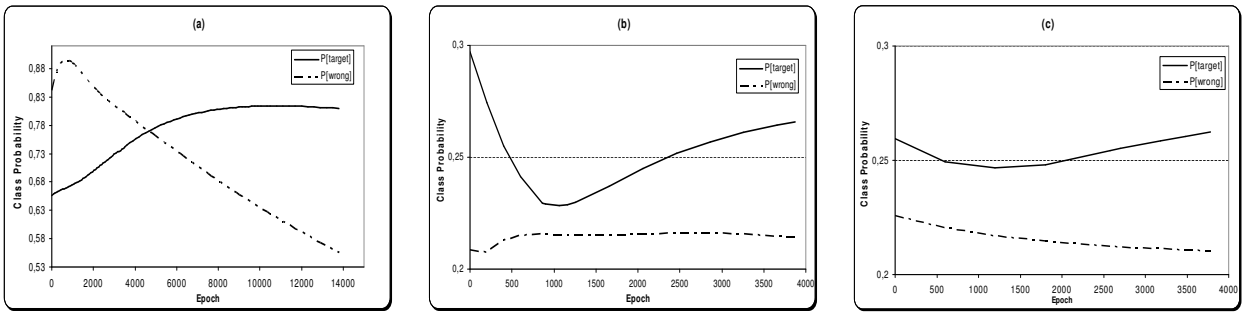

Figure 5*: *Class conditional probabilities of the target class ("P[target]") and of the most probable wrong class ("P[wrong]"), for (a) the misclassification; (b-c) the two false rejections, when varying the number of epochs.*

*Note that the vertical scales vary from plot to plot in this Figure.