Available online at www.sciencedirect.com

SCIENCE DIRECT®

Pattern Recognition Letters 25 (2004) 1377–1388

Pattern Recognition Letters

www.elsevier.com/locate/patrec

Feature selection in the independent component subspace for face recognition

H.K. Ekenel *, B. Sankur

Department of Electrical and Electronic Engineering, Bogazici University, Bebek 34342, Istanbul, Turkey

Received 2 October 2003; received in revised form 19 March 2004
Available online 11 June 2004

Abstract

This paper addresses the feature selection problem for face recognition in the independent component subspace. While there exists, at least, energy principle to guide the selection of the principle components, the independent components (ICs) are devoid of any energy ranking, and must therefore selected based on their discriminatory power. In addition the independent component features can be selected starting from a much larger pool, or from a combination pool of ICA and PCA features. Four feature selection schemes have been comparatively assessed, and feature subsets are tested on a face database constructed from CMU PIE and FERET databases. The discriminatory features from larger pools are observed to be concentrated around fiduciary spatial details of the nose, the eyes and the facial contour. Overall, face recognition benefits from the feature selection of ICA or PCA components and from the combination of ICA and PCA feature pools.

1. Introduction

Face recognition has become one of the most active research areas of pattern recognition since the early 1990s. The interest on face recognition is mainly fueled by the identification requirements for access control and for surveillance tasks against terrorism. This interest is still increasing, since face recognition is also seen as an important part of next-generation smart environments (Pentland and Choudhury, 2000).

Among the plethora of face recognition methods, the paradigm based on face appearance data, template-based algorithms and their concomitant subspace versions, such as PCA and LDA methods are the most popular (Turk and Pentland, 1991; Belhumeur et al., 1997). Recently a blind source separation technique, called independent component analysis (ICA) has been adopted for face recognition as an alternative subspace method. Face recognition algorithms, however, encounter several difficulties due to changes of the face appearances caused by such factors as occlusion, illumination, expression, pose, make-ups and aging. In fact the subsequent intra-individual variability of face images can be larger than the inter-individual variability (Gong et al., 2000).

* Corresponding author. Tel.: +90-212-358-15-00; fax: +90-212-287-24-65.
E-mail address: ekenelha@boun.edu.tr (H.K. Ekenel).

0167-8655/$ - see front matter © 2004 Published by Elsevier B.V.
doi:10.1016/j.patrec.2004.05.013

One expects to surmount these difficulties, or at least to mitigate their effect by judicious choice of features that are insensitive to the variations in the facial appearance. The purpose of this study is to explore subsets of ICA and PCA features for face recognition that increase recognition performance, and that are purportedly robust against expression variations and differences in illumination. The feature selection criterion used in PCA-based face recognition is the eigenvalue variance. On the other hand the independent components (ICs) are devoid of any energy/importance ranking, therefore relatively high-dimensional feature vectors are used in face recognition using ICA, i.e., 200 dimensions (Baek et al., 2001; Bartlett et al., 1998; Draper et al., 2003) with respect to PCA. To handle this problem, proportion of variance (PoV) has so far been the only feature selection technique used in the literature on ICA face recognition (Bartlett et al., 1998; Deniz et al., 2001; Yuen and Lai, 2002; Havran et al., 2002). Besides using the feature selection methods based on individual properties of the features, like proportion of variance and best individual feature, we also utilize more sophisticated feature selection schemes that take into account the combinatorial properties of the features such as sequential forward selection and sequential floating forward selection. We assess comparatively these four different feature selection schemes on the basis of the improvement they bring in the recognition performance. Furthermore, since the first face recognition architecture of ICA provides local features, we tried to discover which one of these local features has more discriminatory power, and thus contributes most to the person identification.

The paper is organized as follows. In Section 2 we review briefly PCA and ICA techniques. The feature selection methods are presented in Section 3. Experimental results are discussed in Section 4 and conclusions are drawn in Section 5.

2. Subspace analysis methods

Face data, as obtained from the raster scanning of the face images, constitutes a very high-dimensional space. However the intrinsic dimensionality of the face space is known to be much smaller (Gong et al., 2000), this despite the variations in expression, pose and lighting. In fact the faces are believed to be clustered on some low-dimensional manifolds. The subspace techniques aim to reduce the inherent excessive dimensionality of scanned data to make the face recognition algorithms viable and to capture or approximate the underlying face manifolds. The most widely used subspace analysis tools are the principal component analysis (PCA), independent component analysis (ICA), linear discriminant analysis (LDA) and their nonlinear varieties via kernel tools (Turk and Pentland, 1991; Belhumeur et al., 1997; Kim et al., 2002; Yang, 2002; Moghaddam, 2002).

In the sequel, lowercase and uppercase letters denote scalar values, bold uppercase letters denote matrices, and bold lowercase letters denote row vectors.

2.1. Principal component analysis (PCA)

PCA aims to determine a new orthogonal basis vector set that best reconstructs the face images, in other words with the smallest mean-square error for any given subspace dimensionality. These orthogonal basis vectors, also called eigenfaces, are the eigenvectors of the covariance matrix of the face images. The most parsimonious eigenvector set, say of dimension M, for the face reconstruction problem is chosen as the subset of the M most energetic eigenvectors, that is the eigenvectors corresponding to the first M rank ordered eigenvalues.

Consider the $K \times D$-dimensional face data matrix $\boldsymbol{X}$, where each D-dimensional row corresponds to the lexicographically ordered pixels of one of the faces, and where there are K face images. The PCA method tries to approximate this face space using an M-dimensional feature vector, that is using M eigenfaces, where typically $M \ll \min(D, K)$. These M eigenvectors span a face subspace, such that $\|\boldsymbol{X}\|^2 - \|\boldsymbol{XV}\|^2$ is minimum, where $\boldsymbol{V}$ is the $D \times M$-dimensional matrix that contains orthogonal basis vectors of the face space in its columns. Once the projection bases $\boldsymbol{V}$ are formed, when a test image $\boldsymbol{x}_t$ arrives, it is projected onto the face subspace to yield the feature vector,

$\boldsymbol{r}_t = \boldsymbol{x}_t \boldsymbol{V}$. The classifier decides for the identity of the individual, according to a similarity score between $\boldsymbol{r}_t$ and the feature vectors of the individuals in the database $\{\boldsymbol{r}_1, \boldsymbol{r}_2 \ldots, \boldsymbol{r}_K\}$.

2.2. Independent component analysis (ICA)

Independent component analysis is an unsupervised learning method based on high order statistics. Briefly, ICA is the separation of independent sources from their observed linear mixtures (Hyvarinen and Oja, 2000). The system model of ICA is given as

$$\boldsymbol{X} = \boldsymbol{AS} \tag{1}$$

where $\boldsymbol{A}$ denotes the mixing matrix, $\boldsymbol{S}$ denotes the source matrix containing statistically independent source vectors in its rows and $\boldsymbol{X}$ denotes the data matrix. In the ICA method, the only information we possess is the observations, and neither the mixing matrix nor the distribution of the sources is known. Under the assumptions that the sources are statistically independent and non-Gaussian (at most one of them can have Gaussian distribution), we find the unmixing matrix $\boldsymbol{W}$ by maximizing some measure of independence. In other words, a separation matrix, $\boldsymbol{W}$, is estimated, which, under ideal conditions, is the inverse of the mixing matrix $\boldsymbol{A}$.

$$\boldsymbol{Y} = \boldsymbol{WX} \quad \text{and} \quad \boldsymbol{W} = \boldsymbol{A}^{-1} \quad \text{and} \quad \boldsymbol{Y} \approx \boldsymbol{S} \tag{2}$$

In the context of face recognition, the use of ICA features was first proposed in (Bartlett et al., 1998) and several other studies followed in (Liu and Wechsler, 1999; Baek et al., 2001; Kwak et al., 2002; Deniz et al., 2001; Yuen and Lai, 2002; Ding et al., 2001; Havran et al., 2002; Draper et al., 2003). In (Bartlett et al., 1998) two different approaches are presented for face recognition. In the first approach (called ICA1 architecture), the face images are assumed to be a linear mixture of an unknown set of statistically independent source images. The source images obtained in this architecture are spatially local and sparse in nature. In the second approach (called ICA2 architecture), the representation (weighting) coefficients are assumed to be statistically independent. In this second architecture, while mixing coefficient vectors are independent, source images tend to have global face appearance, as in the case of PCA. In contrast to the PCA method, where feature subset selection is based on energy criterion, the selection of an ICA basis subset is not immediately obvious since the energies of the independent components cannot be determined. One typically first trims the $\min(D, K)$-dimensional space to M using a PCA stage, and then proceeds to extract the ICA components from this M-dimensional space.

The architecture of ICA1 and that of PCA have analogous structures, as they are both based on the statistical properties of the basis images. In ICA1, one tries to find statistically independent basis images, whereas in PCA one extracts uncorrelated basis images. ICA2 is based on the statistical properties of the representation coefficients, thus it has a different structure with respect to ICA1 and PCA. In this work, both for a fair comparison with PCA-based method and to explore the local face regions that are appropriate for identification, the first architecture of the ICA is used.

In summary, the face recognition algorithm using ICA1 architecture is as follows:

(i) Prior to ICA, PCA is performed on a training set and the M eigenvectors, associated with the largest eigenvalues are selected to form the matrix $\boldsymbol{V}$.

(ii) The training face images are projected onto the PCA-based face space and the $K \times M$ matrix of their representation coefficients, $\boldsymbol{R} = [\boldsymbol{r}_1 | \boldsymbol{r}_2, \ldots, | \boldsymbol{r}_K]^{\mathbf{T}}$, are obtained:

$$\boldsymbol{R} = \boldsymbol{X} * \boldsymbol{V} \tag{3}$$

(iii) An ICA analysis is performed on $\boldsymbol{V}^{\mathrm{T}}$, where face eigenvectors form the rows of this matrix; the unmixing matrix, $\boldsymbol{W}$, reveals the "ICA faces"

$$\boldsymbol{S} = \boldsymbol{W} * \boldsymbol{V}^{\mathrm{T}} \tag{4}$$

(iv) By using PCA representation coefficients, $\boldsymbol{R}$, and independent basis images, $\boldsymbol{S}$, ICA representation coefficients of the faces in the training set, $\{\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_K\}$, are calculated as follows:

$$A = R * W^{-1} \tag{5}$$

since $V^{\mathrm{T}} = W^{-1} * S$, $X = R * V^{\mathrm{T}}$, and $X = R * W^{-1} * S$. When a test image x_t arrives, it is projected onto the face subspace to extract its PCA representation $r_t = x_t V$, which is then multiplied with the inverse of the separation matrix to yield the ICA feature vector $a_t = r_t * W^{-1}$. This vector is compared vis-à-vis the feature vectors $\{a_1, a_2, \ldots, a_K\}$ of the individuals in the face database.

3. Feature selection techniques

3.1. Rationale for feature selection for PCA/ICA faces

The main goal of this study is the automatic selection of the best feature subset for classification purposes given a high-dimensional ICA feature vector. Recall that the feature selection criterion used in PCA-based face recognition is the eigenvalue variance. While this criterion yields the most compact set for the reconstruction, it does not necessarily follow that it is optimal from face recognition point of view. The ICA methods do not have such an opportunity of feature selection, in that the number of ICA features is pre-determined in the PCA stage of data processing. It is conjectured that some feature selection scheme focused on "recognition" rather then on "reconstruction" could augment face classification performance. With this goal in mind, we have not initially reduced the subspace dimensionality via PCA to the target feature size, M. Instead we have reduced the initial dimensionality $\min(D, K)$ down to an intermediate size M', $M < M' < \min(D, K)$. The judicious choice of M', on the one hand, should make the implementation of the ICA algorithm viable and avoid overlearning effects (Bartlett et al., 1998; Liu and Wechsler, 1999). On the other hand M' should allow sufficient freedom or richness of choice for the feature selection algorithms to be effective. For example, if the target feature vector will be 30-dimensional, we do not effect a $\min(D, K)$-to-30 PCA algorithm, but use, for example, a $\min(D, K)$-to-200 PCA reduction. Following this, the feature selection algorithm proceeds to find the most discriminating $M = 30$ ICA features from the intermediate set of $M' = 200$.

We have comparatively assessed the recognition performances resulting from different feature selection algorithms, namely, proportion of variance (PoV), best individual feature (BIF), sequential forward selection (SFS) and sequential floating forward selection (SFFS) techniques (Feri et al., 1994). PoV has so far been the only feature selection technique used in the literature on ICA face recognition (Bartlett et al., 1998; Deniz et al., 2001; Yuen and Lai, 2002; Havran et al., 2002). PoV would be the best approach if the features' contributions to the face recognition performance were independent from each other. BIF is also based on the same assumption as PoV, the difference being that, while in PoV the variances of features are used, in BIF the classification performances of the individual features are taken into consideration. SFS technique is a special case of the well known "plus l – take away r" methods, and SFFS is an enhanced version of SFS. These last two techniques search for the best performing feature set while taking into consideration the correlation between features.

3.2. Proportion of variance (PoV)

In this technique, for all feature points the ratio of between-class variance to within-class variance is calculated and the feature points that have high ratio are selected. In this context "class" denotes the ensemble of images belonging to one individual. Thus the mth feature of the qth face image ($q = 1, \ldots, Q$) of the kth ($k = 1, \ldots, K$) individual will be expressed as $f_m(k, q)$. The class mean will be obtained as

$$\bar{f}_m(k) = \frac{1}{Q} \sum_{q=1}^{Q} f_m(k, q) \tag{6}$$

while the average of this feature over the database will be denoted as

$$\bar{\bar{f}}_m = \frac{1}{KQ} \sum_{k=1}^{K} \sum_{q=1}^{Q} f_m(k, q) \tag{7}$$

The ratio of between-class variance to within-class variance for the mth feature point in the feature vector is calculated as

$$r_m = \frac{\sigma_{\text{between},m}}{\sigma_{\text{within},m}} = \frac{\sum_{k=1}^{K}\left(\bar{f}_m(k) - \bar{\bar{f}}_m\right)^2}{\sum_{k=1}^{K}\sum_{q=1}^{Q}\left(f_m(k,q) - \bar{f}_m(k)\right)^2} \tag{8}$$

This measure reveals out whether a change in facial appearance corresponds to a change in the mth feature point or not. Thus we rank the features according to their r_m ratio where $f_{r(m)}$ is the mth ranking feature, and select the highest ranking M features, $\{f_{r(1)}, f_{r(2)}, \ldots, f_{r(M-1)}, f_{r(M)}\}$.

3.3. Best individual feature (BIF)

In this technique, classification performance of each feature point is calculated separately, that is, on individual basis, and the features giving rise to highest correct recognition rate are selected. Let P_m be the classification performance of the mth feature point, that is, P_m is the probability of correct detection using solely feature f_m. Then the feature points are ordered according to their individual classification performances P_m and the first M ones, having highest P_m values, are selected.

3.4. Sequential forward selection (SFS)

In the SFS method, features are selected successively by adding the locally best feature point, the feature point that provides the highest incremental discriminatory information, to the existing feature subset. The SFS technique starts as the BIF by identifying the first feature that has the highest discrimination power. It proceeds, however, by adding sequentially to the selected subset, those features that contribute most to the classification performance on top of the already selected ones. Thus, from a single initial BIF feature, the SFS subset grows to a pair, to a triple…till an M-fold subset is found.

3.5. Sequential floating forward selection (SFFS)

SFFS is similar to the SFS in adding features to the subset; however, in addition, it goes through cleansing periods, in that features are removed systematically so long as the performance improves after pruning. Note that the performance of the pruned subset is compared with that of a previous one with the same population. This stratagem helps to avoid the nesting effect, which results when one is stuck with a suboptimal subset.

4. Results and discussion

4.1. Experimental setup

We have used a face database constructed with images chosen from CMU PIE and FERET databases (Phillips et al., 2000; Sim et al., 2002). In our experiment set, there are 214 individuals each having four different frontal face images, making a total of 856 images. 584 out of the 856 images are chosen randomly from the FERET database, fafb image set. These FERET-based images mainly differ in expression. The remaining 272 images derive from the CMU PIE database, which contain variations in illumination. All the face images are aligned with respect to the manually detected eye coordinates, scaled to 60×50 pixels resolution and histogram equalized. For each individual in the set, two of their images that contain normal facial expression and have frontal illumination are used for training, and the remaining two images that contain alternative facial expressions and illumination from left and right sides are used for testing purposes (Fig. 1).

Fig. 1. Sample face images: four images of each of two individuals with differing illumination effects, expression or accessories.

Feature vectors are extracted via independent component analysis and principal component analysis schemes. The FastICA algorithm (Hyvarinen and Oja, 2000) is used to perform independent component analysis. In either case we extracted 200-dimensional feature vectors, that is, the original 3000-dimensional image feature vectors consisting of raw pixel values were reduced to, respectively, 200 PCA features conserving 96.46% of the energy and 200 ICA features. Any further feature selection was carried on these intermediate sets of 200.

We used the nearest-neighborhood classifier in our feature selection study. We opted for the nearest-neighbor method, as it is a powerful non-parametric classifier without any costly training stage since otherwise one needs to re-design the classifier repetitively for each selected feature subsets. Finally we evaluated comparatively three different distance metrics, namely, the $L1$ norm, the $L2$ norm, and the normalized correlation coefficient, defined as follows:

$$L1 : d = \sum_{j=1}^{M} |f_{\text{training},j} - f_{\text{test},j}| \tag{9}$$

$$L2 : d = \left(\sum_{j=1}^{M} |f_{\text{training},j} - f_{\text{test},j}|^2 \right)^{1/2} \tag{10}$$

$$CC : d = \frac{\boldsymbol{f}_{\text{training}} \cdot \boldsymbol{f}_{\text{test}}}{\|\boldsymbol{f}_{\text{training}}\| * \|\boldsymbol{f}_{\text{test}}\|} \tag{11}$$

In Fig. 2, sample ICA and PCA basis images are shown. In the first row, samples of eigenfaces are seen. The basis images in the second row are instances of ICA faces obtained by applying source separation to 30 eigenfaces, whereas in the third row we display basis images that are obtained by separating 200 eigenfaces. It can be observed that, as the number of independent sources increases, the ICA images become more spatially localized and sparse. On the other hand, when a small number of sources are used, the ICA basis images resemble the eigenfaces.

4.2. Experimental results

In the sequel we present the classification results as well as the cumulative matching scores (CMS) of face images under various feature selection schemes and distance metrics.

In Table 1, the correct classification performances of ICA and PCA are given when, respectively, 30 and 200-dimensional feature vectors are used. As it has also been reported in the literature (Moghaddam, 2002), there are no significant performance differences between ICA and PCA. As

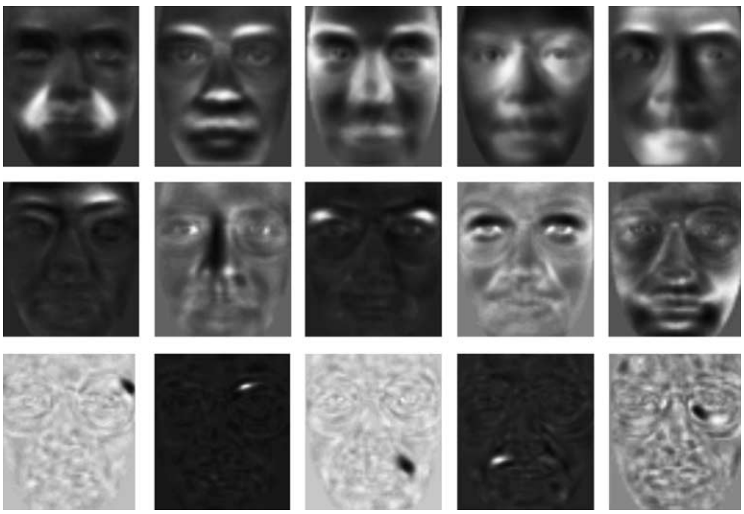

Fig. 2. Sample basis images: first row—eigenfaces; second row—typical instances of 30 ICA faces obtained from 30 eigenfaces; third row—typical instances of 30 ICA faces obtained from 200 eigenfaces.

Table 1
Correct classification performance of ICA and PCA

	$L1$	$L2$	CC
ICA-30	75.47	74.30	73.60
PCA-30	74.77	74.30	73.60
ICA-200	80.61	79.67	78.50
PCA-200	81.07	79.67	78.50

one should expect, the correct recognition rate gradually increases with the increase in the dimension of the feature vector. For both feature sets, the $L1$ norm gives the best results; hence the $L1$ distance metric is taken into consideration while performing feature selection. Recall that in this experiment the most energetic 30 (200) PCA components were selected as feature sets, and that the ICA features were simply obtained by further processing these PCA components via the Fast-ICA algorithm. The rationale for this choice is explicated in Fig. 3 on the basis of the relative energy accounted for by the chosen number of eigenfaces.

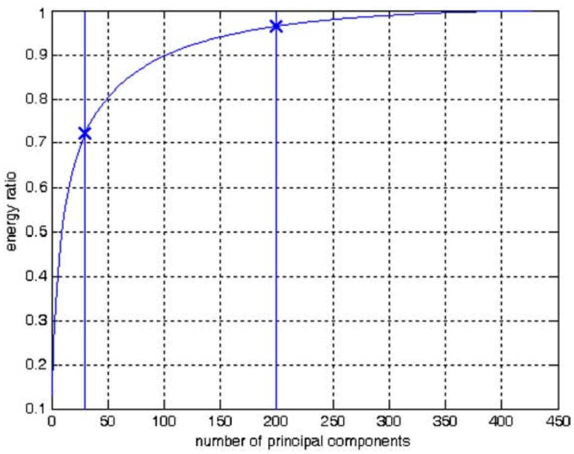

Fig. 3. Energy percentage accounted by the eigenfaces.

To prove the conjecture that a judiciously selected subset from the individual feature sets or from the merged set of both ICA and PCA faces can yield better performance, we carried the following experiments. In the first experiment, we selected subsets of 30 features out of the initial set of 200, individually for the PCA and ICA cases using various selection algorithms. In the second experiment, we merged the two sets to constitute a superset of 400 dimensions (200 PCA + 200 ICA) and then proceeded to select the subset of 30 features from this augmented feature set. Note that before merging of the two feature sets, they were rendered commensurate by variance normalization of the ICA and PCA feature vectors. In Table 2, the performance values of the selected feature subsets are shown. Sequential selection methods perform better than the PoV and BIF methods. The reason can be attributed to the fact that, in classification, the class-conditional features are not independent. Therefore combinatorial feature selection proves to be more effective than selecting them based only upon their individual performances. In other words, selecting the feature points that have high individual classification performances does not imply the best subset for classification. For example, in Fig. 4, the ratio of between-class variance to within-class variance is shown. Although ICA coefficients have higher ratio as compared to those of PCA coefficients, this does not necessarily imply higher classification performance when considered jointly as can be observed from Table 2. These results show that the correct combination of the feature points in the selected subset is more important than any selection based on individual properties.

Table 2
Performance results of the feature selection techniques, when the initial feature set is reduced from 200 to 30. From 400 to 30 in the combined case

Feature selection method	PCA-30/200	ICA-30/200	Combined 30/400
PoV	74.30	71.73	72.90
BIF	74.07	75.93	71.73
SFS	77.34	80.84	79.67
SFFS	78.74	83.64	86.68

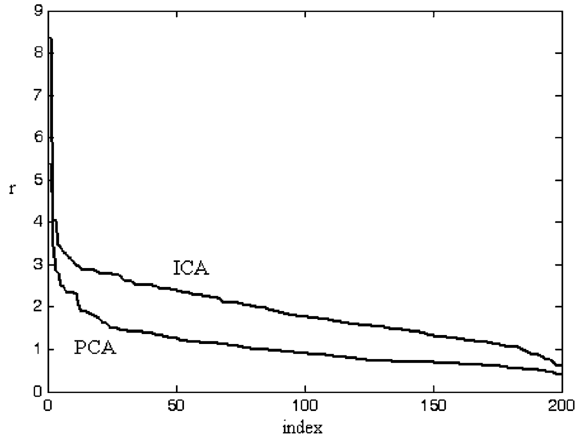

Fig. 4. The ratio of between-class to within-class variances.

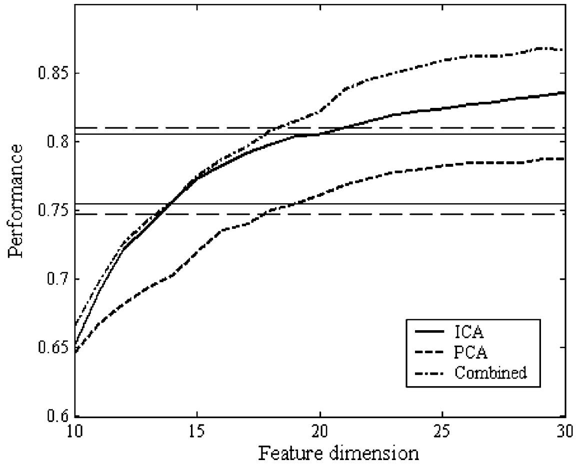

Fig. 5. Cumulative matching score curves of SFFS.

One can observe from Table 2 that, in the sequential selection techniques, ICA achieves higher correct recognition rates as compared to PCA. While trying to bring down the dimensionality of the PCA feature set proves detrimental, the converse is true for the ICA set. As the ICA feature set is reduced from the original 200 components down to dimension 30, the performance improves by 3%, that is from the score of 80.61 for ICA-200 to the score of 83.64 for ICA-30 using the SFFS method. This increase in classification performance is a consequence of the removal of the ICA basis images localized on those face regions experiencing major changes in the appearance and having less discriminatory information.

It is interesting to observe that the subset selected from merged 400 dimensional ICA and PCA feature vectors perform even better than separate ICA or PCA feature vectors, despite the fact the ICA features were obtained from PCA features via a linear method. Finally we can observe that the SFFS feature selection method proves to be uniformly superior in all experiments. In fact the ranking in increasing performance of the feature selection methods in this experiment is as follows: PoV < BIF < SFS < SFFS. In Fig. 5, cumulative matching score curves of SFFS method are given. The horizontal lines in the figure represent the performance values of the 30 and 200-dimensional PCA and ICA feature vectors obtained without any selection. As can be seen from the figure, in ICA, nearly a 20-dimensional selected feature vector can outperform the initial 200-dimensional feature vector, or similarly, a 14-dimensional selection of features is on a par with the 30-dimensional feature performance.

In Figs. 6–8, we show the indices of the selected 30 features from the original set of 200 PCA and ICA features and from the combined set using SFFS method. In Fig. 6, the features are ordered according to the energy index, that is the size of

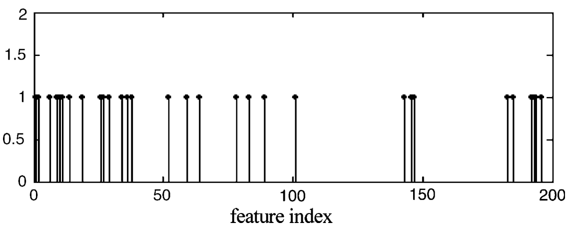

Fig. 6. Selected PCA feature points by SFFS.

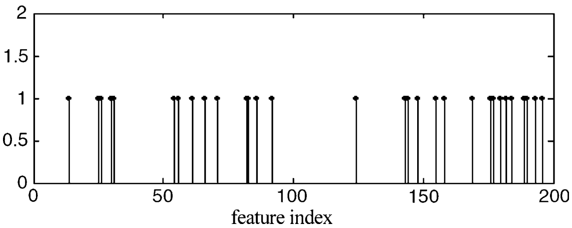

Fig. 7. Selected ICA feature points by SFFS.

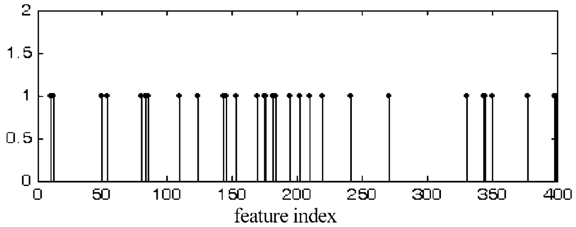

Fig. 8. Selected feature points from the combined set by SFFS: note that the first 200 indices refer to the ICA features, and the second set of features, with indices from 201 to 400 refer to the PCA features.

their eigenvalue. In Fig. 7, the ordering is what results from the ICA algorithm, after the PCA reduction stage. Finally note that in Fig. 8, the ICA and PCA features are juxtaposed in a single graph, such that the first 200 indices correspond to the ICA features and the last 200 indices correspond to the PCA features. Contrary to the common practice of selecting the most energetic PCA features, as can be observed in Fig. 6, only two thirds of the PCA feature are from the first 100 points. This shows that some of the less energetic PCA coefficients, which are normally discarded for image representation and compression, may still contain valuable discriminatory information. The selected ICA features are spread more evenly over the index range. Finally, it is interesting to observe that the number of selected feature indices from PCA and ICA feature pools are approximately equal.

In Fig. 9, the selected ICA features are presented. One notices that these features are characterized by sparse regions localized around the eyes, on the face contours and at the nose. One can interpret these regions in the selected ICA components as being less sensitive to the variations in the facial appearance and at the same time to contain more discriminatory information. These findings confirm the results of a recent study on optimal Gabor kernel location selection (Gokberk et al., 2003). Another observation from Fig. 9 is that, these basis images resemble the receptor fields

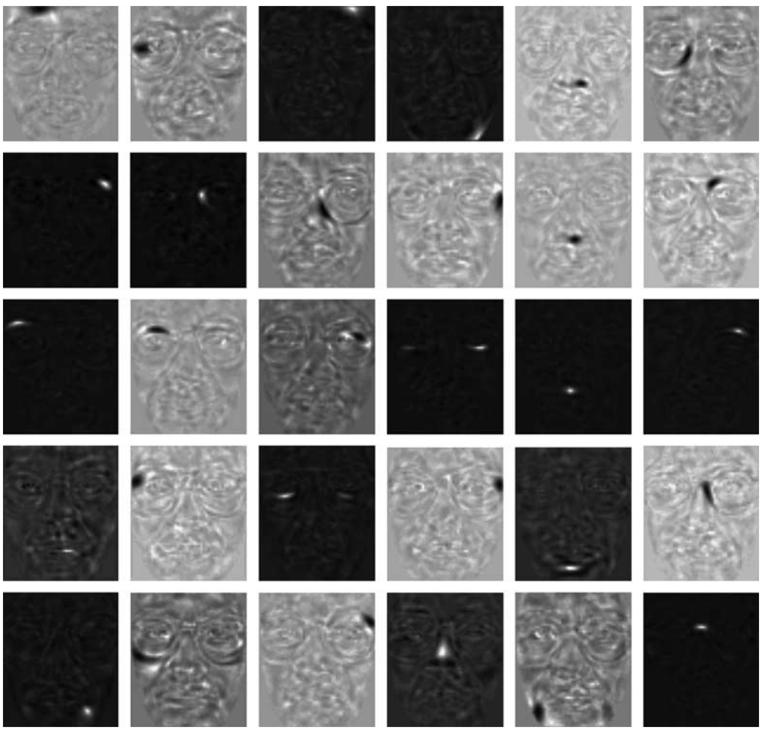

Fig. 9. Selected ICA faces with SFFS method from a training set of normal face appearances.

obtained in local feature analysis (LFA) (Penev and Atick, 1996).

4.3. Discussion of the results

The rationale to resort to a feature selection scheme was to improve the face recognition performance and to mitigate the effects of intra-variation. We investigated the source of the attained improvement (compare for example, ICA-30 performances in Tables 1 and 2). Some of the face image, which were erroneously classified without feature selection in ICA, but were then correctly labeled after feature selection are shown in Fig. 10, in pairs side by side the test image and erroneously matched training image. As can be seen from these face pairs, their appearances are very similar and the individuals can be discriminated by only a careful consideration of the local face regions. Appropriate selection of the ICA basis images provides the utilization of this valuable local discriminatory information.

Fig. 10. Samples of misclassified face images. First image in the pair is the test image; second image in each pair is the matched training image.

A desirable aspect of the proposed scheme should be the robustness from one face database to another. In other words, one would like to encounter the same or very similar subset of fea-

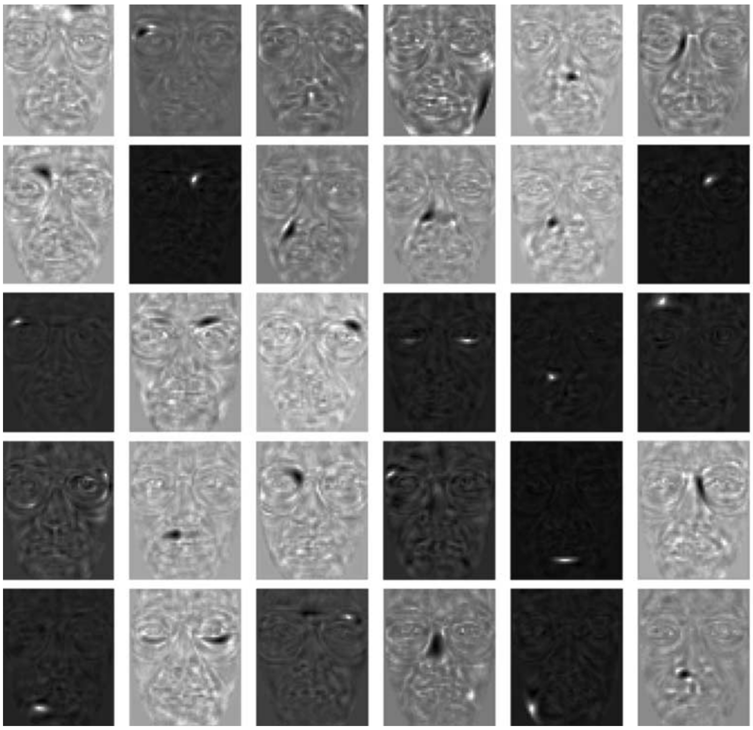

Fig. 11. Selected ICA faces with SFFS method from a training set of faces with expression and illumination variations.

tures, as one switches from one face database to another one. To validate the usefulness of the features we have selected, we perform ICA on a different training set. We swap the images in the test set with those in the training set. Thus, in this new arrangement, the face images in the training set contain differences in expression and variations in illumination, and the face images in the test set contain normal facial expression and have frontal illumination. The independent components that correspond to the selected ICA feature points with SFFS method are presented in Fig. 11. As can be observed from Fig. 11, again the ICA basis images that localize on the eye region, outlines of face and nose are selected.

5. Conclusions

In this study we have explored feature selection techniques on ICA and PCA bases for face recognition. Feature selection techniques are warranted especially for ICA features since these are devoid of any importance ranking based on energy content as the PCA components. The study was carried out on a face database that contains both facial expression and illumination variations. Four different feature selection techniques were used comparatively and the sequential floating forward selection method was observed to be uniformly superior in all cases, in that the maximum correct classification rate was obtained with its feature subset.

The major conclusion from this study was that the feature selection applied on ICA features definitely improves the recognition performance by 8.17%. Furthermore, if the features are selected from the augmented pool of both ICA and PCA features, the performance improvement becomes 11.21%. In other words, it pays to select ICA/PCA features subset from a larger set of them, rather than deciding a priori for the dimensionality of the final feature subset. Indeed, instead of selecting features on the basis of the first M most energetic PCA components or their ICA versions, it was more beneficial to search for the subset resulting in the best possible classification performance from a larger pool. It is interesting to note that even the PCA features benefited from this approach, when 30 of them were selected from the set of 200 via the SFFS algorithm instead simply selecting the first most energetic ones, as is commonly done in the literature.

When the ICA features are selected from a larger initial set, we observe that the resulting features are more localized, as the sharp dark or bright spots in Figs. 9 and 11 indicate. In fact, not surprisingly the majority of these local accents are around the eyes and nose, as well as close to the facial contours.

References

Baek, K., Draper, B.A., Beveridge, J.R., She, K., 2001. PCA vs. ICA: A comparison on the FERET data set. In: Internat. Conf. on Computer Vision, Pattern Recognition and Image Processing, North Carolina.

Bartlett, M.S., Lades, H.M., Sejnowski, T.J., 1998. Independent component representations for face recognition. In: Conf. on Human Vision and Electronic Imaging III, San Jose, California.

Belhumeur, P.N., Hespanha, J.P., Kriegman, D.J., 1997. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. IEEE Trans. Pattern Anal. Machine Intell. 19 (7), 711–720.

Deniz, O., Castrillon, M., Hernandez, M., 2001. Face recognition using independent component analysis and support vector machines. In: Third Internat. Conf. on Audio and Video Based Biometric Person Authentication, Sweden.

Ding, P., Kang, X., Zhang, L., 2001. Personal recognition using ICA. In: Eighth Internat. Conf. on Neural Information Processing, China.

Draper, B.A., Baek, K., Bartlett, M.S., Beveridge, J.R., 2003. Recognizing faces with PCA and ICA. Computer Vision and Image Understanding 91 (1–2), 115–137.

Feri, F.J., Pudil, P., Hatef, M., Kittler, J., 1994. Comparative study of techniques for large-scale feature selection. In: Pattern Recognition in Practice IV. Elsevier, pp. 403–413.

Gokberk, B., Irfanoglu, M.O., Akarun, L., Alpaydın, E., 2003. Optimal Gabor kernel location selection for face recognition. In: Internat. Conf. on Image Processing, Spain.

Gong, S., McKenna, S.J., Psarrou, A., 2000. Dynamic vision: From images to face recognition. Imperial College Press.

Havran, C., Hupet, L., Czyz, J., Lee, J., Vandendorpe, L., Verleysen, M., 2002. Independent component analysis for face authentication. In: KES'2002 Proc., Italy, pp. 1207–1211.

Hyvarinen, A., Oja, E., 2000. Independent component analysis: Algorithms and applications. Neural Networks 13 (4–5), 411–430.

Kim, K.I., Jung, K., Kim, H.J., 2002. Face recognition using kernel principal component analysis. IEEE Signal Process. Lett. 9 (2), 40–42.

Kwak, N., Choi, C., Ahuja, N., 2002. Face recognition using feature extraction based on independent component analysis. In: Internat. Conf. on Image Processing, Rochester.

Liu, C., Wechsler, H., 1999. Comparative assessment of independent component analysis (ICA) for face recognition. In: Second Internat. Conf. on Audio and Video Based Biometric Person Authentication, Washington, DC.

Moghaddam, B., 2002. Principal manifolds and probabilistic subspaces for visual recognition. IEEE Trans. Pattern Anal. Machine Intell. 24 (6), 780–788.

Penev, P., Atick, J., 1996. Local feature analysis: A general statistical theory for object representation. Network: Computation in Neural Systems 7 (3), 477–500.

Pentland, A., Choudhury, T., 2000. Face recognition for smart environments. IEEE Computer 33 (2), 50–55.

Phillips, P.J., Moon, H., Rizvi, S.A., Rauss, P.J., 2000. The FERET evaluation methodology for face-recognition algorithms. IEEE Trans. Pattern Anal. Machine Intell. 22 (10), 1090–1104.

Sim, T., Baker, S., Bsat, M., 2002. The CMU pose, illumination, and expression (PIE) database. In: Proc. IEEE Internat. Conf. on Automatic Face and Gesture Recognition, Washington, DC.

Turk, M., Pentland, A., 1991. Eigenfaces for recognition. J. Cognitive Sci. 3 (1), 71–86.

Yang, M.H., 2002. Kernel eigenfaces vs. kernel fisherfaces: Face recognition using kernel methods. In: Fifth IEEE Internat. Conf. on Automatic Face and Gesture Recognition, Washington, DC.

Yuen, P.C., Lai, J.H., 2002. Face representation using independent component analysis. Pattern Recognition 35 (6), 1247–1257.