NMF vs ICA for Face Recognition

Menaka Rajapakse and Lonce Wyse
Institute for Infocomm Research
21 Heng Mui Keng Terrace, Singapore 119613
menaka@i2r.a-star.edu.sg

Abstract

This paper deals with the application of spatially localized, non-overlapping features for face recognition. The analysis is carried out by using the features generated from two closely related techniques known as Independent Component Analysis (ICA) and Non-negative Matrix Factorization (NMF). A set of statisically independent basis vectors with sparse features is derived from ICA. Likewise, NMF is used to yield sparse representation of localized features to represent distributed parts over a human face. Similarities between reconstructed faces of test images and a set of synthesised face representations from the basis vectors derived from an image database using the two techniques are measured. The strengths and weaknesses of each method in the context of face recogntion are discussed.

1. Introduction

Multivariate data can be represented by subspace projections such as principal component analysis (PCA) [13], independent component analysis (ICA) [11], or non negative matrix factorization (NMF) [7]. These methods can be categorized as global or local feature based methods depending on the amount of influence imposed by the image pixels on the output features. The local feature based methods where only a few pixels in the input image contribute to the forming of the output image have better stablility over global feature based methods where each pixel in the image is contributed to the formalization of output features. Some of the advantages of using local feature based methods over global methods are their less responsiveness to occlusions, scale and lighting variations, and rotations. In this paper, we are interested in analysing only the local feature based approaches. Two such local feature based methods are ICA and NMF.

The Independent Component Analysis(ICA) provides a linear representation that minimizes the statistical dependencies among its components, based on higher order statistics of the data. These dependencies among higer order features could be removed by isolating independent components. The ability of the ICA to handle higher-order statistics in addition to the second order statistics is useful in achieving an effective separatation of feature space for given data. The higher order features are capable of capturing invariant features of natural images. The distribution of face images is unlikely to be Gaussian due to the reason that a face can populate to more than one region in the image space due to its multifariousness. Therefore, the distribution of a face in a low dimensional feature space after a projection such as PCA is also unlikely to be Gausssian [12]. These assumptions support the use of ICA for facial feature extraction under non reduced and reduced face representations. In representing the non-gaussian sources such as face images, the ICA has potential advantages over global PCA. Such advantages are, provision of a better probabilistic model of the data, which gives a better identification of the data clusters in the n-dimensional space, a unique unmixing of data and the ability to handle higher order data [3]. Further, previous research indicate that the features extracted from ICA are similar to those observed in the primary visual cortex [11] and have resemblence to the features extracted by Gabor wavelets.

A recently emerged approach known as Non-negative Matrix Factorization (NMF) is also suggestive of some aspects of activation patterns in response to images in the mammalian visual cortex. In NMF, as the name implies, the non-negativity adds constraints to the matrix factorization, allowing only additions in the synthesis; there are no cancellations or interference of patterns via subtraction or negative feature vector values. This leads more naturally to the notion of parts-based representation of images [7] [8]. With the underlying non-negative constraints, NMF is able to learn localized parts based representations. Sparse coding with NMF seems befitting especially for face recognition applications as the features of face images are naturally represented as a small collection of features, namely eyes, nose and mouth, which are distributed over the face. Because the outputs of NMF are localized features, we can use these parts based features collectively to represent a face.

2 Independent Component Analysis

Finding a suitable transformation to best represent the data is essential for many or all pattern recognition tasks. Linear transforms such as PCA and factor analysis based on second order statistics are commonly used due to their simplicity in manipulation. In second order methods, a representation with minimum reconstruction error of the data is found using the information contains in the covariance matrix of the data. It is assumed that all the information of Gaussian variables (zero mean) is contained in the covariance matrix. However, most data sets such as face images are characterized by non Gaussian and higher order features where higher order methods are needed for providing meaningful representations for such data.

There exist linear and non-linear forms of the ICA that can represent data in the sense of higher order statistics. For most applications, the linear transform is sufficitent. The ICA of the random vector $\mathbf{x} = (x_1, ..., x_m)^T$ finds a linear transform $\mathbf{s} = \mathbf{Wx}$ so that the components s_i are as independent as possible in the sense of highest order statistics. This is achieved by maximizing some function $F(s_1, ..., s_m)$ known as contrast function that measures independence among the components s_i.

The ICA estimates the following generative model for the random vector $\mathbf{x}$

$$\mathbf{x} = \mathbf{As} \tag{1}$$

The matrix $\mathbf{A}$ is a constant mxm mixing matrix. Using the ICA algorithm an estimation of the independent components can be achieved. The columns of A represent features, and s_i is the coefficient of the i^{th} basis feature in the observed data vector $\mathbf{x}$.

Several algorithms have been proposed for the estimation of ICA model. Here, we describe the ICA algorithm introduced by Bell & Sejnowski [5] that is based on the principle of optimal information transfer through sigmoidal neurons. Given an n dimensional random data vector $\mathbf{x}$ and a n x n invertible matrix W, we can write

$$U = W\mathrm{x}. \tag{2}$$

The variables in n dimensional random variable $\mathbf{U}$ are the linear combinations of input data that can be interpreted as activations of an n input neurons in a network. Let the random variable $\mathbf{Y}$ represent the output of n neurons of the network. The aim of this algorithm is to maximize the mutual information between the input data vector $\mathbf{x}$ and the neural network output $\mathbf{Y}$. By maximizing the mutual information, the neural network achieves independence at the output.Mutual Information can be written as

$$I(\mathbf{x}, \mathbf{y}) = \mathbf{H}(\mathbf{y}) - \mathbf{H}(\mathbf{y}|\mathbf{x}) \tag{3}$$

where $H(\mathbf{y})$ denotes the entropy of the output of the network and $H(\mathbf{y}|\mathbf{x})$ is the entropy that is transfered to the output, which does not come from the input.

Prior to learning, sphering is performed on the data in order to uncorrelate the data. The sphering is carried out by subtracting the mean of $\mathbf{x}$ from $\mathbf{x}$ and then passing the zero mean data to a whitening matrix W_z, where $W_z = 2*(Cov(\mathbf{x}))^{-\mathbf{1/2}}$. The ICA transform matrix, learned with the sphered data can be given as $W_{ica} = WW_z$ where W is the learned weight matrix calculated by ICA. The inverse matrix W^{-1} of the weight matrix W is known as the source *mixing* matrix.

2.1 ICA Basis

Given a matrix F where each row is an image of the set $\{\mathbf{f}_1, \mathbf{f}_2, ...\mathbf{f}_n\}$ of n images, and where S is a set of unknown *original* sources. According to Eq. (1), we can write the matrix F as $F = AS$ where A is a mixing matrix which mixes the original sources S to produce the faces in the matrix F.

The function of ICA is to find the independent sources in S. Due to the computational complexity involved this is usually carried out in two stages. In the first stage, the data set is reduced using the PCA. The implementation is same as the one carried out by [3] to evaluate performance of ICA for a set of images with different view points and lighting variations using statistically independent basis images.

The data matrix F whose row vectors are the images in the training data set, we can write $F = [\mathbf{f}_1\mathbf{f}_2....\mathbf{f}_n]^T$. By performing PCA on F, the reduced representation of the training data set can be given as a $P = [\mathbf{p}_1\mathbf{p}_2....\mathbf{p}_m]$, where $\mathbf{p}_i$ are the m principal component vectors corresponding to the highest eigen values where $m < n$.

The next step is to perform ICA on the reduced representation P. As the matrix P has more rows (size of each image) than columns (number of principal components) ICA is performed on the transpose of the matrix P ($P_t = P^T$) resulting in a number of basis vectors equal to number of principal components retained for computations (m).

The *recovered* set of sources $\{\mathbf{u}_1, \mathbf{u}_2, ...\mathbf{u}_m\}$ in the rows of the matrix U of the reduced data matrix P_t can be written as $U = WP_t$ where W is the unmixing matrix which separates the mixes from the matrix P such that $A = W^{-1}$.

The m rows of the matrix U contain the statistically independent basis vectors derived for the reduced training image representation.

The projection of a data matrix F on to reduced feature space of can be expressed as $\mathbf{P} \rightarrow f(\mathbf{F})$.
and the recovered sources $U = W_{ica}P_t$ where $P_t = W_{ica}^{-1}U$. Hence the reconstructed image set $F' = FPP_t$ and by substituting for P_t,

$$F' = FPW_{ica}{}^{-1}U$$

Let

$$H = FPW_{ica}^{-1} \tag{4}$$

then

$$F' = HU \tag{5}$$

where each row of H contains coefficients which can be used to linearly combine the the basis images in U. By doing so, we can reconstruct the images in the rows of the matrix F.

For a given training set F_{train} and test set F_{test}, we can write H_{train} and H_{test} as $H_{train} = F_{train}PW_{ica}{}^{-1}$ and $H_{test} = F_{test}PW_{ica}{}^{-1}$, respectively. A distance metric applied between the elements of H_{test}, $\mathbf{h}_j{}^{test}$ and H_{train}, $\mathbf{h}_j{}^{train}$ is used to measure the similarity of a given test image to an image in the database.

3 Non-Negative Matrix Factorization

Given a data matrix $F = \{F_{ij}\}_{nxm}$, non-negative matrix factorization refers to the decomposition of the matrix F into two matrices W and H of size nxr and rxm, respectively, such that

$$F = WH \tag{6}$$

where the elements in W and H are all positive values. From this decomposition, a reduced representation is achieved by choosing r such that $r < n$ and $r < m$.

In NMF, no negative entries are allowed in matrix factors W and H whereby nonnegativity constraint is imposed in factorizing the data matrix F limiting data manipulation only to additions (no subtractions are allowed). This leads to the idea of reconstructing an object by adding its representative parts collectively. Each column in the matrix W is called a basis image, and a column in the matrix H is called an encoding. An image in F can be reconstructed by linearly combining basis images with the coefficients in an encoding. The encodings influence the activation of pixels in the original matrix via basis images.

Given a data matrix F, Lee and Seung [7] developed a technique for factorizing the F to yield matrices W and H as given in Eq. (6). Each element in the matrix F can be written as $F_{ij} = \sum_{\rho=1}^{r} W_{i\rho}H_{\rho j}$ where r represents the number of basis images and the number of coefficients in an encoding. The following iterative learning rules are used to find the linear decomposition [7]:

$$H_{\rho j} \leftarrow H_{\rho j} \sum_{i=1}^{n} \left(\frac{W_{i\rho}F_{ij}}{\sum_{k=1}^{r} W_{ik}H_{kj}}\right) \tag{7}$$

$$W_{i\rho} \leftarrow W_{i\rho} \sum_{j=1}^{m} \left(\frac{F_{ij}H_{\rho j}}{\sum_{k=1}^{r} W_{ik}H_{kj}}\right) \tag{8}$$

$$W_{i\rho} \leftarrow \frac{W_{i\rho}}{\sum_{k=1}^{n} W_{k\rho}} \tag{9}$$

The above *unsupervised* multiplicative learning rules are used iteratively to update W and H. The initial values of W and H are fixed randomly. At each iteration, a new value for W or H is evaluated. Each update consists of a multiplication and sums of positive factors. With these iterative updates, the quality of the approximation of the Eq. (6) improves monotonically with a guaranteed convergence to a locally optimal matrix factorization [8].

3.1 NMF Basis

The data matrix F, is constructed such that the training face images occupy the columns of the F matrix. Let the training face set be $\Gamma^{train} = \{\mathbf{f}_1{}^{train}, \mathbf{f}_2{}^{train}, \cdots, \mathbf{f}_m{}^{train}\}$, $F = [\mathbf{f}_1\mathbf{f}_2\cdots\mathbf{f}_m]$ and μ represents the mean of all training images. Learning is done using Eq.s (7)-(9) to decompose the matrix F into 2 matrices, H and W. Let the basis images be represented as $W = [\mathbf{w}_1\mathbf{w}_2\cdots\mathbf{w}_r]$ and encodings as $H = [\mathbf{h}_1\mathbf{h}_2\cdots\mathbf{h}_m]$, where each face $\mathbf{f}_i$ in F can be approximately reconstructed by linearly combining the basis images, and the corresponding encoding coefficients $\mathbf{h}_i{}^{\mathrm{T}} = [h_{1i}h_{2i}\cdots h_{ri}]$ as shown in Figure 1. Hence, a face can be modeled in terms of a linear superposition of basis functions together with encodings as follows:

$$\mathbf{f}_i = \sum_{j=1}^{r} \mathbf{w}_j\mathbf{h}_i \tag{10}$$

For each face $\mathbf{f}_i$ in the training set and test set, we calculate the corresponding encoding coefficients. The basis images in W are generated from the set of training faces, Γ^{train}. The encodings, $\mathbf{h}_i$ of each training face $\mathbf{f}_i$ is given by

$$\mathbf{h}_i = W^{\dagger}\mathbf{f}_i$$

where $W^{\dagger}$ is the pseudoinverse of the matrix W. Once trained, the face image set, $\{\mathbf{f}_1, \mathbf{f}_2, \cdots, \mathbf{f}_m\}$ is represented by a set of encodings $\{\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_m\}$ with reduced dimension, r. A distance metric is used to calculate the similarity between $\mathbf{h}_i{}^{train}$ and $\mathbf{h}_j{}^{test}$; encodings of a training image and a test image.

4 Experiment

4.1 Database

The experiments were carried out on M2VTS face database from University of Surrey, which consists of 1180 images, with 4 images per person taken at four different times (one month apart). Though similar lighting conditions and backgrounds have been used during image acquisition, significant changes in hair styles, facial hair, presence and absence of glasses introduce variability into the images. These images

are of frontal and near frontal views with somewhat dissimilar facial expressions. The original image size is 726 x 576 pixels and the database contains images of Caucasians and Asian males and females. The images are normalized for scale, rotation, translation and illumination invariance.

4.2 Preprocessing and Normalization

The preprocessing of images facilitates minimizing the variances among faces of the same individual while maximizing the variance between different individuals. These variances occur in scale, rotation, translation and lighting conditions existing among captured images at different time instances.

The normalized output images have consistent grey levels across all images in the database. The eye positions are fixed at preset coordinates. For the experiments, the final image size is reduced to 64x64 from the original size of 150x200 by re-sampling the images.

The geometric normalization used in our approach is based on the manually located eye positions. In order to achieve faces invariant to rotation, translation and scale, a transformation matrix is computed by joining the located eye positions on a horizontal segment having a length of 52 pixels separating the two eyes in the original dimensions [10]. The redistribution of intensity values of the image is carried out using histogram equalization thereby producing an image with equally distributed intensity values.

4.3 Data Preparation

The database is divided into 2 subsets, a training set and a test set. The training data set consists of an equal number of male and female images. The total number of images used for training is 834. Each individual contributes 3 face images for the training set and a single image to the test set.

4.4 Recognition Performance

Here we compare the recognition performance of ICA and NMF in face recognition. As a baseline measure for recognition, we further compare the results with PCA. The cosine of the angle between the two data vectors, one from the training set and the other from the testing set is taken to calculate the similarity measure, s.

$$s(\mathbf{h}_j^{test}, \mathbf{h}_i^{train}) = \frac{\mathbf{h}_j^{test} \cdot \mathbf{h}_i^{train}}{|\mathbf{h}_j^{test}||\mathbf{h}_i^{train}|} \qquad (11)$$

For our experiment, 834 images in the database were projected on to a reduced image space using PCA. Figure (1), illustrates the amount of variance information captured by eigen components as a cumulative percentage. From this plot we see that the first 200 principal components alone are able to capture 97% of the variance distributed among the 834 images in the database. A subset of these eigenfaces are shown in the 2nd row of the Figure (5).

The ICA was then performed on these eigenfaces to extract a set of statistically independent non-global basis. A subset of these extracted basis images are shown in the 3rd row of the Figure (5). This was performed for several sets of principal components, $50, 100, 150, 200, 300$ separately. For each set of principal components, the sources were recovered and a mixing matrix was calculated. With the similarity measure given in the Eq. (11), we then measured the recognition accuracy for each subset at an incremental order of 25 components up to the maximum number of components extracted. The results are shown in the Figure (2). The Figure(3) illustrates the search rates of the first correct hits evaluated for each feature subset at an incremental order of 10 components up to the maximum number of components extracted. From this graph, we see that, for our probe set the search rate for the correct hit is remarkably low even with the increasing number of independent components.

Next we compared the recognition results from PCA and ICA using the Euclidean and the cosine measure given in Eq. (11). In both cases, the results achieved for PCA were superior to the results achieved for ICA. These results are in accordance with the test results acquired by Baek, K. et al.[2] by using ICA for a similar probe set in FERET database and are shown in Figure (4). Likewise for NMF, we experimented with different ranks so as to measure the recognition strength with varying ranks. Under the same initial conditions, we generated basis images and encodings for the ranks, $r = 25, 49, 81, 121, 144$. Figure (4) also illustrates the recognition performance of NMF at each rank given above. From this, we see that NMF outperforms both PCA and ICA significantly.

4.5 Analysis

According to the results achieved, we see that the PCA is able to capture overall variances in face images and able to minimize the reconstruction error when reconstructing images from the reduced data set. With ICA, the goal is to minimize the statistical dependence between the basis vectors. Though these basis vectors are statistically informative they seem to lack the ability to capture significant variances when the data set exhibits larger variances as in the case of our experiment where the probe set was collected over a long period of time. Moreover, ICs are spatially localized and they do not exhibit any direct correspondence to the facial parts as in the case of NMF.

Unlike ICA, the extracted components using NMF

preserve spatial relationships corresponding to the facial landmark features. Further, the basis images extracted from NMF seem to retain some global structures of facial features (see last row of Figure (5)). According to [1], in the brain, the information is represented using an intermediate structure between local features and whole objects. The basis vectors extracted from NMF seem to be in compliance with this by contributing well for the task of face recognition.

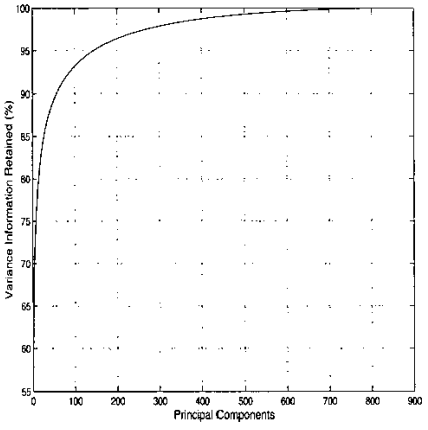

Figure 1. Percentage of variances accounted by eigen components of the face images.

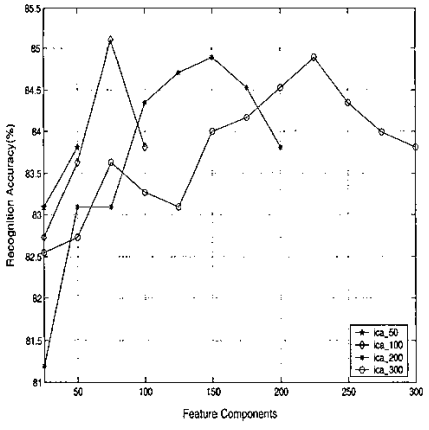

Figure 2. Recognition Rate performance of ICA with subsets of eigen basis images.

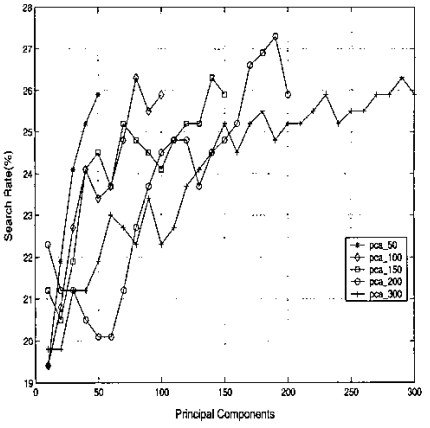

Figure 3. Search Rate performance of ICA with subsets of eigen basis images.

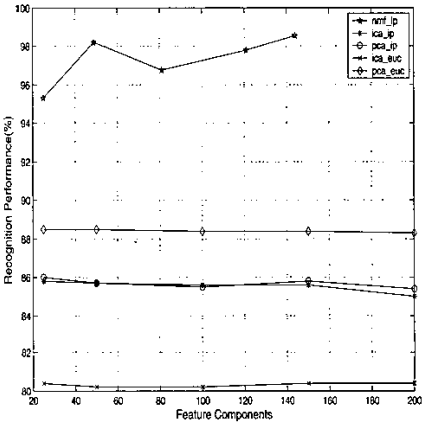

Figure 4. Recognition performance comparison of ICA, PCA and NMF with varying components.

5 Conclusions

We investigated the performance of the ICA and NMF on the MXVTS face database. We have applied parts-based NMF and ICA to learn face images from the M2VTS database. Experimental results show that NMF is quite robust and yields better results compared to the described ICA feature approach and PCA for face recognition, especially when the probe set consists of images taken over a period of time. Having an intermediate representation (parts-based) between local and global structures may be the reason behind for achieving better results with NMF compared to PCA and ICA. A reliable face recognition engine must be able to handle the variabilities introduce by images when the images are taken over a time span as it is what is expected from a real-time face recognition system.

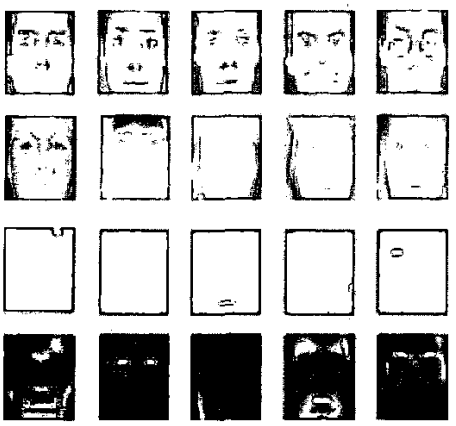

Figure 5. Sample images in the database are given in the £rst row followed by the £rst 5 eigen faces. A subset of basis images derived from the ICA and NMF are given in the 3rd and 4th rows, respectively.

References

[1] Babazadeh, L., "Development of Visual Shape Primitives",*PhD thesis,* University of Southern California, 1999.

[2] Baek, K., Draper, B. A., et al., "PCA vs.ICA:A Comparison on the FERET Data Set", ,,*International Conference on Computer Vision, Pattern Recognition and Image Processing in conjunction with the 6th JCIS,*Durham, North Carolina, March 8-14, 2002.June 2001.

[3] Bartlett, M., Lades, H.,Sejnowski, T., "Independent component representations for face recognition", ',*Proc.of the SPIE, conference on Human vision and Electronic Imaging III,* vol.3299,1998, pp.528-539.

[4] Bartlett, M., Sejnowski, T., "Independent components of face images: A representation for face recognition", *Proceedings of the 4th Annual Joint Symposium on Neural Computation,* May 1997.

[5] Bell, A., Sejnowski, T., "An information maximization approach to blind separation and blind deconvolution", *Neural Computation 7,* 1995, pp.1129-1159.

[6] Bell, A., Sejnowski, T., "The independent component of natural scenes are edge filters", *Vision Research,* Vol. 37(23)1997, pp.3327-3338.

[7] Daniel, D. Lee, H. Sebastian Seung. "Learning the parts of objects by non-negative matrix factorization.", *Nature,*Vol 401,pp 788-791, 1999.

[8] Daniel, D. Lee, H. Sebastian Seung. "Algorithms for Non-negative Matrix Factorization.", *Proc. Neural Information Processing Systems,* 2000.

[9] Gonzalez, R, Woods, R., " Digital Image Processing", *Addison-Wesely,* 1993.

[10] Huang, W, et al., "A robust approach to face and eye detection from images with cluttered background",*Proc. IEEE 14th International Conference on Pattern Recognition,* vol. 1,1998, pp.110-113.

[11] Hyvarinen, A., Oja, E., "Independent Component Analysis: Algorithms and Applications."m *Neural Networks,* 13(4-5):411-430, 2000.

[12] Shaogang Gong, Stephen J McKenna and Alexandra Psarrou., " Dynamic Vision, From Images to Face Recognition", *Imperial College Press,* 2000.

[13] Turk, M., Pentland, A., "Eigenfaces for recognition ",*J. Cog. neurosci.,* vol. 3(1),1991, pp.71-86.