Face detection by using independent component decomposition

I. Buciu C. Kotropoulos I. Pitas *

Department of Informatics, Aristotle University of Thessaloniki
GR-540 06, Thessaloniki, Box 451, Greece, {costas,pitas}@zeus.csd.auth.gr

Abstract. In this paper we explore the independent component decomposition for face detection. The minimization of the Kullback - Leibler divergence and the maximization of the entropy are two methods employed to decompose an original image into its independent components. We built nearest neighbor classifiers based on their resulting independent components and compare their ability to detect faces to that of support vector machines.

1 Introduction

There are many applications in which human face detection plays a very important role. For example, it can be used in content-based image database indexing/searching, surveillance systems, and human-centered computer interfaces. It also constitutes the first step in a fully automatic face recognition system. A comprehensive survey on face detection methods is given in [1]. A face detection technique based on independent component decomposition is developed in this paper. The principal components matrix of the original face and non-face patterns is assumed to represent a mixture of independent image sources which are retrieved by using independent component analysis (ICA) through an unmixing matrix. We can reconstruct the original images by combining linearly these sources. The matrix which contains the coefficients of those combinations is further use as the first input of the two nearest neighbor classifiers employed in the paper. The second input is a combination of the test image with principal components matrix and the unmixing matrix. The classification is then performed according to the nearest neighbor rule. Testing this approach against support vector machines (SVMs), we found the latter is outperformed by the proposed method in the face detection task.

2 Spatial independent component analysis

The goal of is to decompose a set of observations into a basis whose components are statistically independent or, at least, are as independent as possible. ICA

* This work was supported by the European Union Research Training Network "Multimodal Human-Computer Interaction (HPRN-CT-2000-00111).

originally applied to blind source separation [2]. Two ICA representations of facial patterns have been proposed in [3] for face recognition. The discriminating ability of ICA alone or when combined with other discriminant criteria, such as Bayesian framework or Fisher's linear discriminant, was analyzed in [4].

In our analysis we follow the model proposed in [3]. Consider a matrix $\mathbf{X}$ whose rows contain vectors formed by scanning lexicographically face and non-face patterns (i.e., image regions). We assume that $\mathbf{X}$ contains a mixture of the original independent sources $\mathbf{U}$. The matrix is decomposed into a family of $\mathbf{Y}$ independent sources passing it through an unmixing matrix $\mathbf{D}$ in the attempt to recover $\mathbf{U}$. Each source (row of $\mathbf{Y}$) is an image whose pixel values are independent of those in every other image. Accordingly, these images are said to be spatially independent. We refer to this model as the *spatial* ICA. Having a number of n face and non-face images, the number of independent components will be n as well. In order to have a control on the number of independent components, we choose m linear combinations of face and non-face patterns, namely the principal component vectors of the image set. Let $\mathbf{P}_m^T$ denote the matrix that is formed by the m principal components in its rows. The objective of ICA applied onto $\mathbf{P}_m^T$ is to find the matrix $\mathbf{Y}$ whose rows are the statistically independent sources by appropriately determining the unmixing matrix $\mathbf{D}$. The relationship between the three aforementioned matrices is given by [3]:

$$\mathbf{Y} = \mathbf{D}\mathbf{P}_m^T. \tag{1}$$

Frequently, a *whitening* process applied to $\mathbf{P}_m^T$ is necessary to decorrelate and normalize the data. If the row means are substracted from $\mathbf{P}_m^T$ and the resulting matrix is passed through a zero-phase whitening filter which is twice the inverse square root, the whitening transformation is written as $\mathbf{W} = 2(\mathbf{P}_m^T\mathbf{P}_m)^{-\frac{1}{2}}$. Therefore, the zero - mean input matrix can be computed as the product of the unmixing matrix and the whitening matrix $\mathbf{D}_w = \mathbf{D}\mathbf{W}$. Eq. (1) is rewritten as follows:

$$\mathbf{Y} = \mathbf{D}_w\mathbf{P}_m^T \Longrightarrow \mathbf{P}_m^T = \mathbf{D}_w^{-1}\mathbf{Y}. \tag{2}$$

The reconstructed image by ICA is:

$$\mathbf{X}_{recICA} = (\mathbf{X}\mathbf{P}_m\mathbf{D}_w^{-1})\mathbf{Y} = \mathbf{C}_{train}\mathbf{Y}. \tag{3}$$

The matrix $\mathbf{C}_{train}$ contains the coefficients of the linear combination of spatial independent sources $\mathbf{Y}$. Each row of $\mathbf{Y}$ comprises the independent component representation of the face images. Once we have finished training and obtained $\mathbf{Y}$, a test image can be presented as:

$$\mathbf{c}_{test} = \mathbf{D}_w^{-1}\mathbf{P}_m\mathbf{x}_{test}. \tag{4}$$

2.1 Entropy maximization

Given $\mathbf{P}_m^T$, the component in (1) which is responsible for obtaining the independent sources is the unmixing matrix $\mathbf{D}$ that must be updated in order to

obtain sources that are as independent as possible. Different approaches exist for this purpose. One way is the so called maximum entropy method which has been developed in [5]. The matrix $\mathbf{Y}$ is transformed into a matrix $\mathbf{Z}$ by passing it through a component-wise nonlinearity denoted by $\mathbf{G}[\cdot]$. As ICA is applied on the columns of $\mathbf{P}_m^T$, a realization $\mathbf{p}_j$ is a combination of the original sources $\mathbf{u}_j$ via a mixing matrix $\mathbf{A}$, $\mathbf{p}_j = \mathbf{A}\mathbf{u}_j$. Therefore, the sources can be restored through the unmixing matrix $\mathbf{D}$ as $\mathbf{y}_j = \mathbf{D}\mathbf{p}_j \approx \mathbf{u}_j$. For simplicity we omit the index j from now on. Passing the sources $\mathbf{y}$ through $\mathbf{G}$ yields:

$$\mathbf{z} = \mathbf{G}(\mathbf{y}) = \mathbf{G}(\mathbf{D}\mathbf{p}) = \mathbf{G}(\mathbf{D}\mathbf{A}\mathbf{u}). \tag{5}$$

Therefore:

$$\mathbf{u} = \mathbf{A}^{-1}\mathbf{D}^{-1}\mathbf{G}^{-1}(\mathbf{z}) = \mathbf{\Psi}(\mathbf{z}). \tag{6}$$

The entropy is given by:

$$h(\mathbf{z}) = -E[\log(f_{\mathbf{Z}}(z))] = -E\left[\log\left(\frac{f_{\mathbf{U}}(\mathbf{u})}{|\det(\mathbf{J}(\mathbf{u}))|}\right)\right], \tag{7}$$

where $f_{\mathbf{Z}}(\mathbf{z})$ and $f_{\mathbf{U}}(\mathbf{u})$ are the probability density functions of $\mathbf{Z}$ and the sources $\mathbf{U}$, and $\mathbf{J}$ is the Jacobian matrix $\mathbf{J} = \partial\mathbf{z}/\partial\mathbf{y}$. Using the chain rule, the determinant of $\mathbf{J}$ can be evaluated as:

$$|\det(\mathbf{J}(\mathbf{u}))| = \left|\det\left(\frac{\partial\mathbf{z}}{\partial\mathbf{y}}\right)\right| = |\det(\mathbf{D}\mathbf{A})|\prod_{i=1}^{m}\frac{\partial z_i}{\partial y_i}. \tag{8}$$

Maximizing the entropy $h(\mathbf{z})$ requires to maximize the expectation of the denominator term $\log|\det(\mathbf{J}(\mathbf{u}))|$ with respect to the matrix $\mathbf{D}$:

$$\frac{\partial}{\partial\mathbf{D}}(\log|\det(\mathbf{J}(\mathbf{u}))|) = [\mathbf{D}^{-1}]^T + \sum_{i=1}^{m}\frac{\partial}{\partial\mathbf{D}}\log\left(\frac{\partial z_i}{\partial y_i}\right). \tag{9}$$

If $z_i = g(y_i) = 1/(1+e^{-y_i})$ is a component-wise nonlinearity applied to all elements of matrix $\mathbf{Y}$, and taking into account that:

$$\frac{\partial z_i}{\partial s_i} = z_i(1-z_i), \tag{10}$$

and $\mathbf{y} = \mathbf{G}^{-1}(\mathbf{z})$, (9) becomes:

$$\frac{\partial}{\partial\mathbf{D}}(\log|\det(\mathbf{J}(\mathbf{s}))|) = [\mathbf{D}^{-1}]^T + (\mathbf{1}-2\mathbf{z})\mathbf{p}^T. \tag{11}$$

Using the gradient ascent algorithm, the change of the unmixing matrix $\mathbf{D}$ is [5]:

$$\Delta\mathbf{D} = \eta(\mathbf{D}^{-T} + (\mathbf{1}-2\mathbf{z})\mathbf{p}^T). \tag{12}$$

It is more convenient to use the natural gradient instead of the actual one to avoid inverting $\mathbf{D}$ at each step, therefore, the formula for unmixing matrix change becomes:

$$\mathbf{D}_{k+1} = \eta[\mathbf{I} + (\mathbf{1}-2\mathbf{z})\mathbf{y}^T]\mathbf{D}_k. \tag{13}$$

2.2 Minimization of the Kullback-Leibler divergence

Another way to obtain independent sources is equivalent with minimizing the Kullback-Leibler divergence between the probability density function $f_{\mathbf{S}}(\mathbf{s};\mathbf{D})$ parameterized by $\mathbf{D}$ and the corresponding factorial distribution defined by [6]:

$$\widehat{f_{\mathbf{Y}}}(\mathbf{y};\mathbf{D}) = \prod_{i=1}^{m} \widehat{f_{\mathbf{Y}}}(\mathbf{y}_i;\mathbf{D}). \tag{14}$$

The Kullback-Leibler divergence is given by:

$$\mathcal{D}_{f\|\widehat{f}}(\mathbf{D}) = -h(\mathbf{y}) + \sum_{i=1}^{m} \widehat{h}(\mathbf{y}_i), \tag{15}$$

where $h(\mathbf{y})$ is the entropy of the random vector $\mathbf{y}$ at the output of the unmixer and $\widehat{h}(\mathbf{y}_i)$ is the marginal entropy of the ith element of $\mathbf{y}$. The minimization can be implemented using the method of gradient descent. Following [6], the unmiximg matrix will be updated at each iteration k as follows:

$$\mathbf{D}_{k+1} = \mathbf{D}_k + \eta[\mathbf{I} - \theta(\mathbf{y}_k)\mathbf{y}_k^T]\mathbf{D}_k^{-T}, \tag{16}$$

where $\mathbf{I}$ is the identity matrix and the analytical form of the *activation function* $\theta(\mathbf{y})$ is also given by [6].

3 ICA performance evaluation

The ability of ICA for face detection was evaluated using face patterns derived from the AT&T face database. A description of the data is given in [7]. A number of 294 non-face patterns was collected and added to 306 face patterns, achieving a total data base of 600 patterns. 80 of them were used to form the training set. Each row of the training matrix contains a 238 - dimensional vector. This matrix was updated according to (13) and (16) for the first and second method respectively, for 1000 iterations. The learning rate η was set to 10^{-6}. The evaluation of the ICA performance was assessed by means of two classifiers. The first one is based on the nearest neighbor rule and measures the angle between a test vector and a training one. Let us denote the class of face feature vectors by $\mathcal{L}_1$ and those of the non-face feature vectors by $\mathcal{L}_{-1}$. Let $\mathbf{c}_{+1}$ be a row vector of $\mathbf{C}_{train}$ matrix that corresponds to the nearest face pattern. Let us denote the nearest non-face neighbor of $\mathbf{c}_{test}$ by $\mathbf{c}_{-1}$. Then we compute the quantities:

$$d_f = \frac{\mathbf{c}_{test}^T\mathbf{c}_{+1}}{\|\mathbf{c}_{test}\|\|\mathbf{c}_{+1}\|} \quad \text{and} \quad d_{nf} = \frac{\mathbf{c}_{test}^T\mathbf{c}_{-1}}{\|\mathbf{c}_{test}\|\|\mathbf{c}_{-1}\|}, \tag{17}$$

where d_f and d_{nf} are the cosines of the angle between a test feature vector and the nearest training one. We assign $\mathbf{c}_{test}$ to $\mathcal{L}_1$ if $d_f > d_{nf}$, otherwise $\mathbf{c}_{test} \in \mathcal{L}_{-1}$. Notice that the labels for the training set are preserved, therefore we know the labels corresponding to $\mathbf{C}_{train}$.

The second classifier is the a minimum Euclidean distance classifier. The Euclidean distance from $\mathbf{c}_{test}$ to $\mathbf{c}_k$, where $k \in \{\pm 1\}$ is expressed as

$$\begin{aligned} \|\mathbf{c}_{test} - \mathbf{c}_k\|^2 &= -2[\mathbf{c}_k^T \mathbf{c}_{test} - \frac{1}{2}\mathbf{c}_k^T \mathbf{c}_k] + \mathbf{c}_{test}^T \mathbf{c}_{test} \\ &= -2h_k(\mathbf{c}_{test}) + \mathbf{c}_{test}^T \mathbf{c}_{test}, \end{aligned} \tag{18}$$

where $h_k(\mathbf{c}_{test})$ is a linear discriminant function of $\mathbf{c}_{test}$. A test pattern is classified by this classifier (also known as "maximum correlation classifier") by computing two linear discriminant function $h_{+1}(\mathbf{c}_{test})$ and $h_{-1}(\mathbf{c}_{test})$ and assigning $\mathbf{c}_{test}$ to the class corresponding to the maximum discriminant function.

We have investigated the performance of the two previously mentioned classifiers (17) and (18) by varying the number of principal components extracted from the training set. The results are depicted in Figure 1. A minimum error of 5.2% was achieved using 20 principal components in the case of the second classifier. However, the performance of this classifier seems to be almost insensitive to the number of the principal components used. On the contrary, for the nearest neighbor rule, the classification error decreases as the number of principal components involved increases. A minimum 3.9% classification error is achieved by keeping 70 linear combinations of 80 training vectors. For comparison, support vector machines (SVMs) with different kernels [8] were applied to discriminate between the face and the non-face patterns. The error rates for different SVMs are included the Table 1, in the same experiment for comparison purposes.

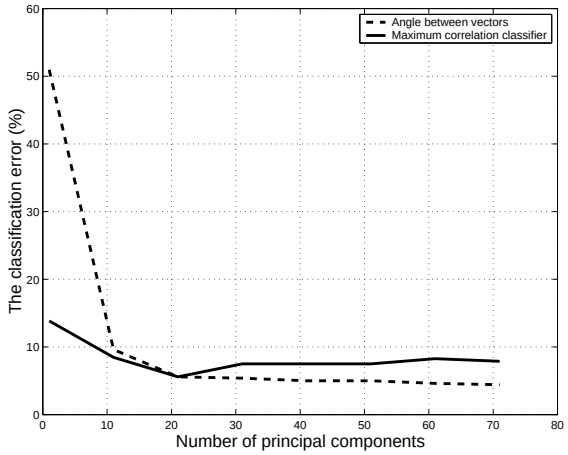

Fig. 1. Classification error (false acceptance rate plus false rejection rate) versus the number of principal components for both classifiers.

Table 1. Number of errors (%) for several classifiers.

Face detection methods	Errors (%)
ICA-based classifier 1	3.9
ICA-based classifier 2	5.2
linear SVM	6.1
polynomial SVM with degree equals 2	6.3
polynomial SVM with degree equals 3	11.1
radial basis function SVM	5.5
exponential radial basis function SVM	6.1

4 Conclusions

We have exploited the ability of ICA to provide useful features in order to conduct a face detection task. The combination of ICA with nearest neighbor classifiers seems to provide a reliable face detector that can outperform SVMs.

References

1. M.-H. Yang, N. Ahuja, and D. Kriegman, "A survey on face detection methods," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, no. 1, pp. 34–58, January 2002.
2. J.F. Cardoso, "Blind signal separation: statistical principles," *Proceedings of the IEEE, Special Issue on Blind Identification and Estimation*,vol. 90, no. 8, pp. 2009–2026, October, 1998.
3. M.S. Bartlett, H.M. Lades, and T.J. Sejnowski, "Independent component representations for face recognition," in *Proc. SPIE Conf. on Human Vision and Electronic Imaging III*, vol. 3299 pp. 528–539, 1998.
4. C. Liu and H. Wechsler, "Comparative Assessment of Independent Component Analysis (ICA) for Face Recognition," in *Proc. Second Int. Conf. on Audio- and Video-based Biometric Person Authentication*, 1999.
5. A.J. Bell and T.J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 6, pp. 1129–1159, 1995.
6. S. Haykin, *Neural Networks. A Comprehensive Foundation*. New Jersey: Prentice-Hall, Inc. 1999.
7. I. Buciu, C. Kotropoulos and I. Pitas, "Combining support vector machines for accurate face detection," in *Proc. 2001 IEEE Int. Conf. on Image Processing*, pp 1054–1057, 2001.
8. V.N. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer Verlag, 1995.