Ref: Valentin, D., Abdi, H., O'Toole, A.J. (in press, 1996). Principal component and neural networks analyses of face images: Explorations into the nature of information available for classifying faces by sex. In C. Dowling, F.S. Roberts,P. Theuns, *Progress in mathematical psychology.* Hillsdale: Erlbaum.

Principal component and neural network analyses of face images: Explorations into the nature of information available for classifying faces by sex

Dominique Valentin*, Hervé Abdi*† and Alice J. O'Toole*

* *School of Human Development: The University of Texas at Dallas, MS: GR.4.1., Richardson, TX75083–0688, U.S.A.*

† *Université de Bourgogne à Dijon, Boulevard Gabriel, 21000 Dijon, France*

ABSTRACT

In this paper we present an overview of the major findings of the principal component analysis (PCA) approach to facial analysis. In a neural network or connectionist framework this approach is known as the linear autoassociator approach. Faces are represented as a weighted sum of macrofeatures (eigenvectors or eigenfaces) extracted from a cross-product matrix of face images. Using sex categorization as an illustration, we analyzed the robustness of this type of facial representation. We show that eigenvectors representing general categorical information can be estimated using a very small set of faces and that the information they convey is generalizable to new faces.

1. Introduction

One of the major problems in modeling face processing is to find a way of representing faces that allows for the wide range of tasks typical of human performance. Traditionally, computational models of face recognition represent faces in terms of geometric descriptors that include distances, angles, and areas between elementary features such as eyes, nose, or chin (Harmon & Hunt, 1977; Harmon, Khan, Lash & Ramig, 1981; Kaya & Kobayashi, 1972; Sakai, Nagao & Kidode, 1971) or in terms of template parameters (Yuille, 1991), or isodensity lines (Nakamura, Mathur & Minami, 1991). Although this approach[1] constitutes an economical way of representing faces that is relatively insensitive to variations in scale, tilt, or rotation of the faces, it is not without problems.

The major difficulty with representing faces as a set of features is that it assumes some *a priori* knowledge about what are the features and/or what are the relationship between features essential for the task at hand. For example, a recent study by Burton, Bruce, and Dench (1993) illustrates the difficulty of finding a set of features useful to discriminate

[1] for a review, cf. Samal & Iyengar, 1992.

accurately between male and female faces. In a series of five experiments, they examined the usefulness of different kinds of feature measures for predicting the sex of a set of faces. The measures they used ranged from simple raw distances between facial landmarks (*e.g.*, pupils) to more complex measures including ratios of distances or angles taken from full faces and/or profile views of the faces. They showed that no simple set of features can predict the sex of faces accurately. By combining all the measurements used in their five experiments, however, they obtained 94% correct classification by sex (for the learning set). They concluded that "explicit measurement" of facial features is probably not the best "basis for automated face recognition systems".

Recent connectionist models (*i.e.*, artificial neural network) of face processing have provided a different way of representing faces using "features" automatically derived from the statistical structure of a set of learned faces. In the literature, this approach is generally referred to as the *principal component analysis* (PCA) *approach* to face processing. In what follows we shall use this terminology. In brief, PCA models represent the faces, explicitly or *via* a neural network architecture, as a weighted sum of eigenvectors extracted from the cross-product matrix of a set of face images. This type of representation has the advantages of eliminating the difficult problem of feature selection and extraction, and of capturing information useful for modeling a wide range of tasks such as face recognition and categorization.

The general purpose of the present work is to analyze the robustness of the kind of face representation proposed by the PCA approach. The human face is a complex visual pattern that contains general categorical information as well as idiosyncratic, identity specific information. By categorical information, we mean that some aspects of a face are not specific to that particular face but are shared by subsets of faces (*e.g.*, female faces share some visual characteristics such as smoothness of the skin, prominence of the cheeks, or roundness of the face). These aspects can be used to assign both unfamiliar and familiar faces to general semantic categories such as sex or race. Previous work (O'Toole, Abdi, Deffenbacher & Valentin, 1993, see also Valentin, Abdi & O'Toole, 1994, for a review) suggests that different ranges of eigenvectors represent different kinds of information in faces. Eigenvectors with large eigenvalues seem to capture information that is common to subsets of faces (*i.e.*, categorical information). Eigenvectors with small eigenvalues tend to capture information specific to individual faces (*i.e.*, identity specific information).

Our objective was to examine the extent to which the categorical information conveyed by the eigenvectors extracted from a particular set of face images is generalizable to other face images. Specifically, we wanted to show that 1) while all ranges of eigenvectors are necessary to categorize *learned* faces optimally along general categories such as sex or race, only the information contained by the eigenvectors with relatively large eigenvalues (*e.g.*, information relative to the general shape of the faces) can be generalized to categorize *new* faces. 2) The stability of the information conveyed by the eigenvectors with relatively large eigenvalues (*i.e.*, categorical information) is such that only a small number of faces is necessary to estimate these eigenvectors.

The present paper is organized as follows. The PCA approach is presented briefly first (for a more detailed presentation see, for example, Valentin, Abdi, O'Toole & Cottrell, 1994; Abdi & Valentin, 1994) followed by a discussion of the interpretation of eigenvectors as "macrofeatures". Next, the usefulness of this approach for analyzing the perceptual information in faces is discussed along with its relationship to some earlier work on the role of different spatial frequencies in face processing. Finally, two series of simulations concerning the statistical properties of eigenvectors derived from a set of face images are described. The first series examines the "generalizability" of the information conveyed by the eigenvectors extracted from small sets of faces; the second series examines their "stability". The results of these simulations are discussed in relation to temporal properties of the visual system and are put in perspective with some neuropsychological data.

2. Principal component and linear neural network approach

This approach is based on earlier work by Anderson, Silverstein, Ritz, and Jones (1977) and Kohonen (1977) on autoassociative memories. Autoassociative memories are a special case of associative memories in which the input patterns are associated with themselves. The goal of autoassociative memories is to find the values or weights for the connections between input units so that when a portion of an input is presented as a memory key, the memory retrieves the complete pattern, filling in the missing components. Kohonen (1977) used faces as stimuli to illustrate some properties of autoassociative memories. Specifically, he showed that an autoassociative memory can act as a content addressable memory for faces. Both Kohonen (1977) and Anderson *et al.* (1977) pointed out that using an autoassociative memory to store a set of patterns is equivalent to computing the eigen-decomposition of the cross-product matrix created from the set of features describing these patterns, or, in other words, to performing the principal component analysis of the set of patterns. The model is presented first, followed by a discussion of the interpretation of eigenvectors as "macrofeatures".

2.1 Model description

To construct the input patterns, each face is digitized and coded as a $I \times 1$ vector $\mathbf{x}_k$ concatenated from the rows of the face image (with I representing the number of pixels, and k indexing the faces). For example, if the k-th face image is a 225×151 pixel image, it is represented by the 33975 elements vector $\mathbf{x}_k$. For convenience, the vectors are assumed to be normalized so that $\mathbf{x}_k^T \mathbf{x}_k = 1$. The set of K faces composing the learning set is represented by an $I \times K$ matrix $\mathbf{X}$ in which the k-th column is equal to $\mathbf{x}_k$.

The faces are stored in an autoassociative memory composed of I neuron-like units as follows: each unit is connected to all the other units and the intensity of the connections is represented by an $I \times I$ matrix $\mathbf{W}$. When standard Hebbian learning is used, $\mathbf{W}$ is given by the sum of the outer product matrices of each face vector

$$\mathbf{W} = \sum_{k=1}^{K} \mathbf{x}_k \mathbf{x}_k^T = \mathbf{X}\mathbf{X}^T \; . \tag{1}$$

The reconstruction of the k-th face is obtained by premultiplication by the matrix $\mathbf{W}$:

$$\hat{\mathbf{x}}_k = \mathbf{W}\mathbf{x}_k \tag{2}$$

where $\hat{\mathbf{x}}_k$ represents the estimation of the k-th face by the memory. The quality of this estimation can be measured by computing the cosine of the angle between $\hat{\mathbf{x}}_k$ and $\mathbf{x}_k$:

$$\cos(\hat{\mathbf{x}}_k, \mathbf{x}_k) = \frac{\hat{\mathbf{x}}_k^T \mathbf{x}_k}{||\hat{\mathbf{x}}_k||\ ||\mathbf{x}_k||} \ . \tag{3}$$

A cosine of 1 indicates a perfect reconstruction of the stimulus.

The performance of the autoassociator can be improved by using a Widrow-Hoff error correction learning rule. The Widrow-Hoff learning rule corrects the difference between the response of the system and the expected response by iteratively changing the weights in matrix $\mathbf{W}$ as follows:

$$\mathbf{W}_{[t+1]} = \mathbf{W}_{[t]} + \eta(\mathbf{X} - \mathbf{W}_{[t]}\mathbf{X})\mathbf{X}^T \tag{4}$$

where η is a small positive constant (typically smaller than one).

2.2 Eigenvectors as "macrofeatures"

As pointed out by Anderson *et al.* (1977) and Kohonen (1977), since the weight matrix $\mathbf{W}$ is a cross-product matrix, it is semi-positive definite (*i.e.*, all its eigenvalues are positive or zeros, and its eigenvectors are real). As a consequence, $\mathbf{W}$ can be expressed in a convenient way as a weighted sum of its eigenvectors:

$$\mathbf{W} = \sum_{\ell=1}^{L} \lambda_\ell \mathbf{u}_\ell \mathbf{u}_\ell^T = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T \qquad \text{with} \quad \mathbf{U}^T\mathbf{U} = \mathbf{I} \tag{5}$$

where $\mathbf{u}_\ell$ is the ℓ-th eigenvector of $\mathbf{W}$, λ_ℓ the ℓ-th eigenvalue, $\mathbf{I}$ stands for the identity matrix, $\mathbf{\Lambda}$ represents the $L \times L$ diagonal matrix of eigenvalues, $\mathbf{U}$ is the $I \times L$ matrix of the eigenvectors, and L is the rank of the matrix $\mathbf{W}$. The eigenvectors in $\mathbf{U}$ are generally ordered according to their eigenvalues. In what follows, the eigenvector with the largest eigenvalue is referred to as the first eigenvector, the eigenvector with the second largest eigenvalue is referred to as the second eigenvector and so on.

Likewise, the original matrix of faces $\mathbf{X}$ can be expressed in terms of eigenvectors and eigenvalues using the singular value decomposition technique (cf. Horn & Johnson, 1985), specifically:

$$\mathbf{X} = \mathbf{U}\mathbf{\Delta}\mathbf{V}^T \ . \tag{6}$$

Where $\mathbf{U}$ represents the matrix of eigenvectors of $\mathbf{X}\mathbf{X}^T$, $\mathbf{V}$ represents the matrix of eigenvectors of $\mathbf{X}^T\mathbf{X}$, and $\mathbf{\Delta}$ is the diagonal matrix of singular values which are equal to the squareroot of the eigenvalues of $\mathbf{X}\mathbf{X}^T$ and $\mathbf{X}^T\mathbf{X}$ (they are the same).

When visually displayed, the eigenvectors of the weight matrix span the entire image and appear face-like. They can be thought of as a set of "global features" or "macrofeatures" from which the faces are built (Abdi, 1988; O'Toole & Abdi, 1989; Sirovich & Kirby, 1987). The estimation of a face by the system can, thus, be represented as a weighted sum of eigenvectors (from Eq. 2 and Eq. 5)

$$\hat{\mathbf{x}}_k = \sum_{\ell=1}^{L} \lambda_\ell \mathbf{u}_\ell \mathbf{u}_\ell^T \mathbf{x}_k = \sum_{\ell=1}^{L} \lambda_\ell \gamma_\ell \mathbf{u}_\ell \quad \text{with } \gamma_\ell = \mathbf{u}_\ell^T \mathbf{x}_k \tag{7}$$

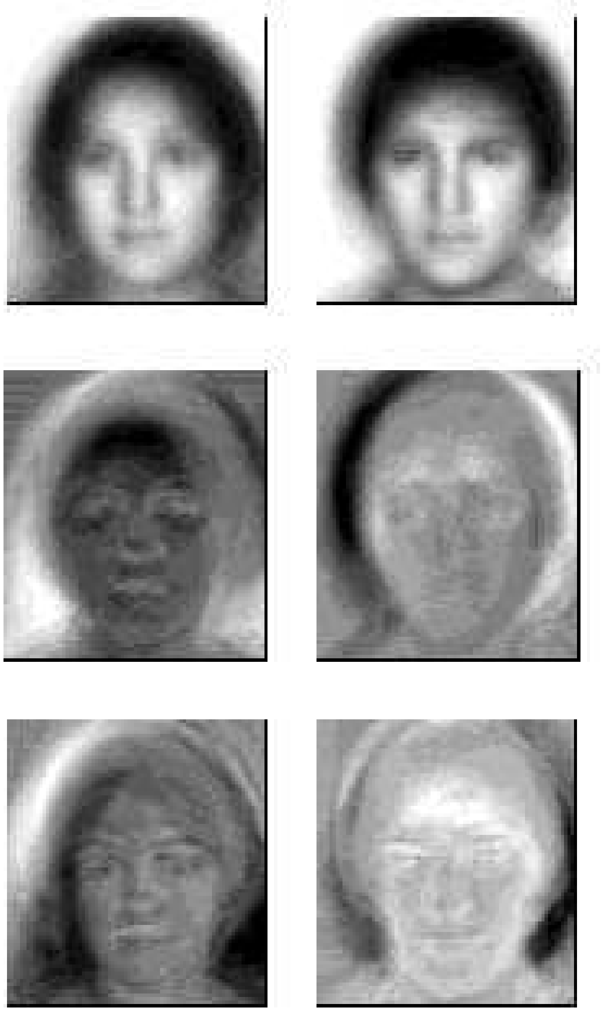

FIG. 1. — The first 3 eigenvectors of an autoassociative memory trained with 80 Caucasian female faces appear in the *left panels*, the first 3 eigenvectors of an autoassociative memory trained with 80 Caucasian male faces in the *right panels*. Global differences in shape and form can be observed between the two sets of eigenvectors.

where the weights γ_ℓ are the projections of the faces onto each of the eigenvectors. These weights (also described as coefficient profiles in O'Toole *et al.*, 1991) can be interpreted as an indication of the extent to which a given eigenvector (or "macrofeature") characterizes a particular face.

A first advantage of using eigenvectors to represent faces is that they are determined *a posteriori*. A second advantage is that they reflect the statistical structure of the set of faces from which they are extracted. As an illustration, Figure 1 displays the first 3 eigenvectors extracted from a set of 80 female faces (left panel) and the first 3 eigenvectors extracted from a set of 80 male faces (right panel). Clearly, these two sets of eigenvectors differ in global shape and form.

FIG. 2. — The *top panels* show three original faces and the *bottom panels* the "responses" produced by an autoassociative memory trained with 160 Caucasian faces when these faces are presented as input after complete Widrow-Hoff learning. The faces are 1) a Caucasian face that has been learned by the memory, 2) a Caucasian face that has not been learned by the memory, and 3) a Japanese face that has not been learned by the memory. The learned face is reconstructed perfectly by the memory ($r = 1$). The Japanese face is more distorted by the memory than the new Caucasian face ($r = .85$ versus $r = .91$, respectively).

From a signal processing point of view, the system acts somewhat like a Wiener filter (Abdi, 1994a). In a neural network terminology this type of systems is known as a linear content addressable memory. When new faces are presented as memory keys, they are filtered through the features extracted from the set of learned faces. Hence, new faces that resemble the learned faces are less distorted by the memory than new faces different from the learned faces. In other words, the more different a face is from the learned faces, the poorer the quality of reconstruction of this face will be. This is illustrated by Figure 2.

An additional *bonus* of the eigenvector approach to linear neural network models is that the Widrow-Hoff learning acts only on the eigen-structure of the weight matrix **W** (Abdi, 1994a). Specifically:

$$\mathbf{W}_{[t]} = \mathbf{U}\{\mathbf{I} - (\mathbf{I} - \eta\mathbf{\Lambda})^t\}\mathbf{U}^T \qquad (\mathbf{8})$$

which shows that the Widrow Hoff learning rule converges only if:

$$0 < \eta < 2\lambda_{\max}^{-1} \qquad (\mathbf{9})$$

where $\lambda_{\max}$ is the largest eigenvalue of **W**.

2.3 Analysis of perceptual information in faces

An interesting aspect of the PCA approach is that it provides a tool for analyzing the perceptual information in faces. Recently, O'Toole *et al.* (1993) examined the kind of information that is provided by different ranges of eigenvectors, and the usefulness of this information with respect to specific tasks. They showed that different tasks make different demands in terms of the information that needs to be processed, and that this information is not contained in the same ranges of eigenvectors. More specifically, they showed that the eigenvectors with larger eigenvalues convey information relative to the basic shape and structure of the faces. This kind of information is most useful in categorizing faces along general semantic dimensions (*e.g.*, sex, race). In contrast, the eigenvectors with smaller eigenvalues tend to capture information that is specific to single or small subsets of learned faces. These eigenvectors are most useful for distinguishing a particular face from any other face.

It is worth noting that the O'Toole *et al.* results can be related to some earlier work by Sergent (1986a, 1986b), in which facial information was analyzed in terms of spatial frequencies. Sergent reported that different kinds of facial information are conveyed by different physical characteristics of the face. Specifically, faces contain both featural and configural properties that coexist. These properties are not conveyed by the same spatial frequency ranges and are not equally useful depending on the nature of the processes involved in performing a particular task. A low frequency representation provides information concerning facial configuration (*i.e.*, general shape, outer contour, and hairline of a face) but does not provide a detailed representation of the individual features. As a consequence, categorization tasks may be accurately achieved by processing only the low frequency range of a face image. In contrast, because recognition and identification tasks require finer information than a simple categorization task, performance for these tasks can be improved by the addition of a high frequency range. High frequencies carry information concerning the inner features of a face (*e.g.*, specific shape of the eyes, nose, and mouth). Figure 3 shows the similarity between the principal components and the spatial frequencies analysis of a visual pattern. While it is clear from this figure that some similarity exists between these two approaches (eigenvectors with large eigenvalues contain mostly low frequency information and eigenvectors with small eigenvalues contain essentially high frequency information), more formal analysis is necessary to clarify the exact nature of this relationship.

Despite the resemblance between these two types of analysis, an important difference exists however. Principal components analysis is performed on a set of faces and so, in addition to the analysis that tends to relate to spatial frequencies, it provides some information about the statistical structure of the sample of faces from which the eigenvectors are extracted. Spatial analysis, operating on the basis of individual faces, does not provide that type of information. Figure 4 provides an illustration of the way PCA and spatial frequencies analysis differ.

In summary, the principal components and the spatial frequencies approaches provide a somewhat similar and complementary analysis of the information contained in facial patterns. The work by O'Toole *et al.* (1993) and Sergent (1986a) shows that most of the information in faces as measured by traditional image analysis techniques (*i.e.*, energy spec-