A GENERALIZED APPROACH FOR CONNECTIONIST AUTO-ASSOCIATIVE MEMORIES: INTERPRETATION, IMPLICATION & ILLUSTRATION FOR FACE PROCESSING

HERVÉ ABDI *

Université de Bourgogne & University of Texas at Dallas

XI.1. Introduction

Recent years have seen an increasing number of papers in Psychology that attempt to model cognitive and perceptual processes using associative memory models (*e.g.,* Hinton & Anderson, 1981; Kohonen, 1984; Rumelhart & McClelland, 1986). Basically, these models use large sets of neuron-like units. The units are linked by connections of modifiable intensity (*e.g.,* synapses). Learning or information storage occurs by modification of the connections. The major advantages of associative models over more traditional information processing and artificial intelligence models are that they make use of computations that are potentially completely parallel and that they employ distributed rather than localized storage of data. Parallel computations are simply those that may, in principle, be done simultaneously and therefore do not depend upon the outcome of the other computations. Distributed storage refers to that fact that discrete locations of memory do not code for individual pieces of data, but rather form parts of the coding of many stimuli.

The workings of both of these features have provided useful insights into some psychological and computational issues in cognition and perception that have proven quite difficult to model with logic-based information processing models.

Along with the psychological studies that have appeared recently, there have been many papers that have attempted to analyze the various learning

* The author wishes to thank Alice O'Toole and Raymond Bruyer for help and comments on previous drafts of this paper and Jim Anderson for support. This paper has been written during a visiting professorship in Brown University made possible by a Fullbright scholarship (1986–1988).

Correspondence about this paper should be adressed to: Hervé Abdi, The University of Texas at Dallas, Program in Cognition, MS:GR4.1., Richardson, TX75083-0688, USA. email: herve@utdallas.edu .

Ref: Abdi, H. (1988). A generalized approach for connectionist auto-associative memories: Interpretation, implication and illustration for face processing. In J. Demongeot, T. Hervé, V. Rialle, & C. Roche (Eds.), *Artificial intelligence and cognitive sciences.* Manchester: Manchester University Press. *pps. 149–165.*

rules used with associative systems. Primarily, these analyses frame learning in an associative network as a process that tries to satisfy a series of constraints. In general, the point is to show that the given learning rule minimizes an error function and that it is able to deal effectively with the usual problems of error minimization such as the avoidance of local minima.

The analysis of energy minimization has traditionally been applied to the problem of characterizing the properties of physical systems. This is, however, a profoundly different point of view than that normally taken by psychologists who have tended more to characterize memory in terms specific to the content and coding of the input. Thus, while this type of physical system analysis has been useful, there is a second, equally useful analysis, that has the potential to provide some interesting common ground between associative models and more traditional ways of modeling psychological data. What is proposed is an analysis of the properties of an associative memory as a *data base*. The aim of this anaylsis is twofold. First, it should provide a way of quantifying similarity relationships among stimuli processed by an associative memory. Second, and perhaps more importantly, it should define stimulus similarity in terms of the properties of the storage matrix formed during the learning process.

Despite the obvious and strong analogies to traditional multidimensional scaling methods of representing distance relationships among stimuli, data from associative memory models have seldom been viewed or presented in this context. We propose to show here a generalization of associative memory that will make this analogy more evident.

Specifically, the generalization will provide a division of the original memory in such a way as to partition out two sets of weights that correspond to the importance of individual stimuli in a set and individual components or "features" of the stimuli. Most models of the distance relations among stimuli have adopted either a Euclidean or a city-block metric for the space; our generalization will provide distance relations in a weighted Euclidean space. These weights will be shown to have meaningful interpretations for the associative model, the multidimensional model, and for some kinds of psychological data.

From a psychological point of view, the analysis proposed is potentially quite useful for a number of reasons. First, it provides a quantitative model of how the components of inter- and intra-stimulus similarity contribute to the behaviour of the associative system. This is relevant in that one of the most interesting and successful uses of associative systems in Psychology has been to model the formation of concepts and prototypes from exemplar data (Knapp & Anderson, 1984). Generally, these psychological data come from sorting or categorization tasks through which perceived distance measures may be derived and the stimuli may be plotted in a low dimensional space that represents a best-fit to the data. It has been shown theoretically (Abdi,

1987) that most of the current models of concept formation can be mimicked by connectionist models. The model described in this paper allows to derive the straightforward derivation representations for auto-associative memories.

In the current connectionist models, all the neuron-like units are of equal importance, and completely independent. Similarly, all the stimuli are of equal importance. This assumption may be quite unrealistic in some cases. This paper describes a generalization for a class of connectionist models: the linear auto-associators, or auto-associative memories. In these models, a set of stimuli is stored in a memory. When presented with a degraded stimulus, an auto-associative memory is able to reconstruct the original stimulus. Auto-associative memories can be interpreted as content-adressable memories (Hopfield, 1982), or categorizer (Anderson *et al.* 1977). The generalization allows stimuli and units to be of differential importance and being non-independant.

XI.2. The "classical" auto-associative model

The "classical" auto-associative model appears frequently in the network model literature (*e.g.,* Anderson, Silverstein, Ritz & Jones, 1977, Hinton & Anderson, 1981; Kohonen, 1984). Its basic features are briefly reviewed here inasmuch as they will be useful in the following sections.

Stimuli are represented by column vectors whose components code the value of the features used to describe the stimuli, or equivalently, the components of the vector give the input values for the basic units (*e.g.,* neurons) of the neural network. Thus, the i^{th} stimulus described by J features will be represented by a $J \times 1$ vector: $\mathbf{f}_i$. It is generally assumed for convenience that the vectors are normalized (*i.e.,* $\mathbf{f}_i^T \mathbf{f}_i = 1$). The set of I stimuli is stored in a $I \times J$ matrix $\mathbf{F}$ in which the i^{th} row stands for $\mathbf{f}_i^T$. The storage of $\mathbf{f}_i$ in the memory is given by the auto-correlation matrix $\mathbf{A}_i = \mathbf{f}_i\, \mathbf{f}_i^T$. The retrieval of the i^{th} exemplar is obtained by computing $\widehat{\mathbf{f}_i} = \mathbf{A}_i \mathbf{f}_i = \mathbf{f}_i \mathbf{f}_i^T \mathbf{f}_i = \mathbf{f}_i$. To store I stimuli in the same memory $\mathbf{A}$, suffice to "integrate" the different matrices $\mathbf{A}_i$:

$$\mathbf{A} = \sum \mathbf{A}_i = \sum \mathbf{f}_i \mathbf{f}_i^T = \mathbf{F}^T \mathbf{F}\ . \tag{1}$$

The retrieval of the i^{th} exemplar stored in $\mathbf{A}$ is given by:

$$\widehat{\mathbf{f}_i} = \mathbf{A}\mathbf{f}_i = \sum_{k=1}^{I} \mathbf{f}_k {\mathbf{f}_k}^T \mathbf{f}_i = \mathbf{f}_i + \sum_{k \neq i} \cos(\mathbf{f}_k\ , \mathbf{f}_i)\mathbf{f}_k \tag{2}$$

Note that when $\mathbf{f}_i$ is orthogonal to $\mathbf{f}_{i'}$ for all i different from i' (*i.e.,* $\mathbf{f}_i^T \mathbf{f}_{i'} = 0, \forall\ i \neq i'$), then $\widehat{\mathbf{f}_i} = \mathbf{f}_i$. In general, the quality of the retrieval is estimated by comparing $\widehat{\mathbf{f}_i}$ with $\mathbf{f}_i$. A popular measure is $\cos(\widehat{\mathbf{f}_i}\, \mathbf{f}_i)$.

XI.3. Generalized auto-associative model

Let $\mathbf{M}$ be a $I \times I$ matrix of constraints on the stimuli. The matrix $\mathbf{M}$ can express a set of *a priori* or *a posteriori* constraints brought to bear on the stimuli. A particular case is to weight differentially the stimuli (*i.e.,* to give different masses), in this case, the matrix $\mathbf{M}$ is a diagonal matrix.

Along the same lines, let $\mathbf{W}$ be a $J \times J$ matrix of constraints on the features, which can represent the *a priori* constraints build-in by the "wiring" of the network, or some *a posteriori* constraints. A particular case is to weight differentially the features used to describe the stimuli. In this case, $\mathbf{W}$ is a diagonal matrix. In this paper, $\mathbf{W}$ and $\mathbf{M}$ are assumed to be *positive definite* matrices.

The effect of the two contraint matrices may be interpreted as a "*deformation*" or "*filtering*" of the stimuli set. Thus, $\mathbf{F}$ is transformed in $\widetilde{\mathbf{F}} = \mathbf{M}^{\frac{1}{2}}\mathbf{F}\mathbf{W}^{\frac{1}{2}}$. Equivalently, $\mathbf{W}$ can be interpreted as a "re-coding" of the original stimuli $\mathbf{f}_i$ in $\widetilde{\mathbf{f}}_i = \mathbf{W}^{\frac{1}{2}}\mathbf{f}_i$ *prior to storage into the memory.* Similarly, $\mathbf{M}$ represents the interference between the stimuli, when or after they have been stored in the memory. Different varieties of interference can be thought of, but only a differential weigthing of the stimuli seems straightforward at the moment. Consequently, in the following section, we assume that $\mathbf{M}$ is a diagonal matrix where m_i stands for the i^{th} diagonal element of $\mathbf{M}$.

For convenience (but without lost of generality) the vectors $\mathbf{f}_i$ are normalized in the metric defined by $\mathbf{W}$ (*i.e.,* $\mathbf{f}_i^T\mathbf{W}\mathbf{f}_i = 1$). The cosine in the metric defined by $\mathbf{W}$ between two vectors is defined as:

$$\begin{aligned}\cos_{\mathbf{W}}(\mathbf{f}_i,\mathbf{f}_k) &= \frac{\mathbf{f}_i^T\mathbf{W}\mathbf{f}_k}{(\mathbf{f}_i^T\mathbf{W}\mathbf{f}_i)^{\frac{1}{2}}\,(\mathbf{f}_k{}^T\mathbf{W}\mathbf{f}_k)^{\frac{1}{2}}} \\ &= \frac{\mathbf{f}_i^T\mathbf{W}\mathbf{f}_k}{\|\mathbf{f}_i\|_{\mathbf{W}}\ \|\mathbf{f}_k\|_{\mathbf{W}}} \\ &= \mathbf{f}_i^T\mathbf{W}\mathbf{f}_k\ . \end{aligned} \tag{3}$$

Given these notations, then the "generalized" auto-associative memory is given by:

$$\widetilde{\mathbf{A}} = \widetilde{\mathbf{F}}^T\widetilde{\mathbf{F}} = \mathbf{W}^{\frac{1}{2}}\mathbf{F}^T\mathbf{M}\mathbf{F}\mathbf{W}^{\frac{1}{2}} = \sum_{i=1}^{I} m_i\widetilde{\mathbf{f}}_i\widetilde{\mathbf{f}}_i^T\ . \tag{4}$$

The retrieval from the auto-associative memory is:

$$\begin{aligned}\widehat{\mathbf{f}}_i &= \widetilde{\mathbf{A}}\widetilde{\mathbf{f}}_i \\ &= \widetilde{\mathbf{A}}\mathbf{W}^{\frac{1}{2}}\mathbf{f}_i \end{aligned} \tag{5}$$

$$= \mathbf{W}^{\frac{1}{2}} \left(\sum_{k=1}^{I} m_k \mathbf{f}_k \mathbf{f}_k{}^T \right) \mathbf{W} \mathbf{f}_i$$

$$= m_i \mathbf{W}^{\frac{1}{2}} \mathbf{f}_i \mathbf{f}_i^T \mathbf{W} \mathbf{f}_i + \mathbf{W}^{\frac{1}{2}} \sum_{k \neq i}^{I-1} m_k \mathbf{f}_k \mathbf{f}_k{}^T \mathbf{W} \mathbf{f}_i$$

$$= \mathbf{W}^{\frac{1}{2}} \left(m_i \mathbf{f}_i + \sum_{k \neq i} m_k \cos_{\mathbf{W}}(\mathbf{f}_i, \mathbf{f}_k) \mathbf{f}_k \right)$$

$$= m_i \widetilde{\mathbf{f}}_i + \sum_{k \neq i} m_k \cos_{\mathbf{W}}(\mathbf{f}_i, \mathbf{f}_k) \widetilde{\mathbf{f}}_k \ . \tag{6}$$

Note that when $\mathbf{f}_i \perp \mathbf{f}_{i'}$ for all $i \neq i'$ (*i.e.,* $\mathbf{f}_i^T \mathbf{W} \mathbf{f}_{i'} = 0$) then $\widehat{\mathbf{f}}_i = m_i \mathbf{W} \mathbf{f}_i = m_i \widetilde{\mathbf{f}}_i$. Consequently $\cos(\widehat{\mathbf{f}}_i, \mathbf{f}_i) < 1$ as well as $\cos_{\mathbf{W}}(\widehat{\mathbf{f}}_i, \mathbf{f}_i) < 1$ *even when* $\mathbf{f}_i$ *is the only stimulus stored in the memory.* However, $\cos(\widehat{\mathbf{f}}_i \, , \, \widetilde{\mathbf{f}}_i) = 1$ when $\mathbf{f}_i$ is the only stimulus stored in the memory, or when all the stimuli stored are orthogonal. This suggests that the quality of the storage should be evaluated with $\cos(\widehat{\mathbf{f}}_i, \widetilde{\mathbf{f}}_i)$.

The term m_i corresponds to a general amplification of the stimulus, by the matrix. In order to evaluate its intensity, suffice to compute the ratio $\widehat{\mathbf{f}}_i \, / \, \widetilde{\mathbf{f}}_i$.

In sum, the generalized memory acts as a filter that deforms the input to take into account the constraints imposed by the $\mathbf{W}$ matrix, when one stimulus is stored.

XI.4. Maximal responses of an auto-associative memory

When the stimulus set is composed of mutually orthogonal stimuli (in the metric defined by $\mathbf{W}$), then the $\widetilde{\mathbf{f}}_i$'s are stored perfectly since the cosine between $\widehat{\mathbf{f}}_i$ and $\widetilde{\mathbf{f}}_i$ is equal to 1. In the more realistic case when the stimuli are not orthogonal, the memory will not give a perfect response to a stored stimulus. Actually, the memory will be prone to "*false recognitions*". In particular, the memory will reconstitute perfectly some stimuli that have not been stored. These stimuli, denoted $\mathbf{u}_i$, are defined by the equation:

$$\widetilde{\mathbf{A}} \widetilde{\mathbf{u}}_i = \lambda_i \widetilde{\mathbf{u}}_i \qquad \text{with: } \widetilde{\mathbf{u}}_i{}^T \widetilde{\mathbf{u}}_i = 1. \tag{7}$$

which classically defines the eigenvectors of $\widetilde{\mathbf{A}}$ (with λ denoting the eigenvalue). From $\widetilde{\mathbf{u}}_i$ it is possible to define the "*generalized eigenvectors*" of $\mathbf{A}$ in the metric $\mathbf{W}$ by the equation:

$$\mathbf{u}_i = \mathbf{W}^{-\frac{1}{2}} \widetilde{\mathbf{u}}_i \qquad \text{where } \mathbf{u}_i{}^T \mathbf{W} \mathbf{u}_i = 1 \tag{8}$$

XI.5. More on the eigenvectors of A and Ã

XI.5.1. *Eigenvectors as "Prototypes".* The eigenvectors $\mathbf{u}_i$ correspond to the maximal responses of the system, under the constraints expressed by $\mathbf{W}$ and $\mathbf{M}$; hence they are "prototypes" spontaneously extracted by the memory as a simple consequence of the storage mecanism. It should be emphasized that the generalized eigenvectors $\mathbf{u}_i$ are different from the standard eigenvectors of the matrix $\mathbf{A}$. Since the matrix $\widetilde{\mathbf{A}}$ is symmetric, the eigenvectors are orthogonal, and the eigenvalues are non-negative. The matrix $\mathbf{A}$ can be reconstituted as:

$$\mathbf{A} = \sum \lambda_i \mathbf{u}_i {\mathbf{u}_i}^T = \mathbf{U\Lambda U}^T \ . \tag{9}$$

(with $\mathbf{\Lambda}$: diagonal matrix of the eigenvalues). The similarity with equation (**4**) is obvious, the eigenvalues λ_i correspond to the m_i and express the importance or the mass of the $\mathbf{u}_i$ in the reconstitution of $\mathbf{A}$.

XI.5.2. *Eigenvectors as "global or macro or holistic features".* The matrix $\mathbf{A}$ can be optimally reconstructed (in a least-square sense) by a small number of eigenvectors. Along the same lines, $\mathbf{F}$ can be reconstituted from the eigenvectors $\mathbf{u}_i$. The reconstitution corresponds to the *generalized singular value decomposition* of $\mathbf{F}$ under the constraints imposed by $\mathbf{M}$ and $\mathbf{W}$:

$$\mathbf{F} = \mathbf{V\Lambda}^{\frac{1}{2}}\mathbf{U}^T \qquad \text{where } \mathbf{V}^T\mathbf{MV} = \mathbf{U}^T\mathbf{WU} = \mathbf{I} \tag{10}$$

The matrix $\mathbf{V}$ contains the generalized eigenvectors of the matrix $\mathbf{FF}^T$ under the constraints given by $\mathbf{M}$ (*i.e.,* $\widetilde{\mathbf{V}} = \mathbf{M}^{\frac{1}{2}}\mathbf{V}$, with $\widetilde{\mathbf{V}}$ being the eigenvectors of $\widetilde{\mathbf{F}}\widetilde{\mathbf{F}}^T$). As each $\mathbf{f}_i$ may be reconstituted by a weighted sum of $\mathbf{u}_i$'s, the $\mathbf{u}_i$'s can be interpreted as "*global* or *macro* or *holistic* features" to be distinguished from the *local* or *micro* features used to describe the stimuli.

XI.5.3. *Eigenvectors and generalized principal component analysis.* It has been pointed out previously (Anderson et al., 1977) that the eigenvectors of the matrix $\mathbf{A}$ may be interpreted as the principal components of the features set (*i.e.,* the features are projected on the "stimuli eigenvectors"). This procedure would correspond to a "*Q-Analysis*" in the factorial analysis terminology. With a generalized auto-associative memory, a generalized principal component analysis is performed by the $\mathbf{u}_i$'s eigenvectors. The projections of the features are given by:

$$\mathbf{P} = \mathbf{U\Lambda}^{\frac{1}{2}} \ . \tag{11}$$

The eigenvalues correspond to the variance of the projections of the features of the axis. Along the same lines, the $\mathbf{V}$ matrix performs a principal component analysis of the stimuli set. The projections of the stimuli are given by:

$$\mathbf{Q} = \mathbf{V\Lambda}^{\frac{1}{2}} \tag{12}$$

As a consequence, the distance between stimuli represented as points in the weighted Euclidean space of the features is "optimally" decomposed along the eigenvectors stored in $\mathbf{V}$. Precisely, the distance between stimuli is obtained as:

$$\begin{aligned} d^2_{\mathbf{W}}(i,i') &= (\mathbf{f}_i - \mathbf{f}_{i'})^T \mathbf{W} (\mathbf{f}_i - \mathbf{f}_{i'}) \\ &= (\mathbf{q}_i - \mathbf{q}_{i'})^T \mathbf{\Lambda} (\mathbf{q}_i - \mathbf{q}_{i'}) = \sum_{j=1}^{K^*} \lambda_j (q_{i,j} - q_{i',j})^2 \end{aligned} \tag{13}$$

where K^* is the number of non-zero eigenvalues of $\mathbf{V}$ or $\mathbf{U}$ ($K^* \leq \min\{I, J\}$), and $q_{i,j}$ is the coordinate of the projection of the i^{th} stimulus on the j^{th} component. This suggests the use of generalized principal component analysis to display the stimuli as they are "perceived" by the auto-associative memory.

XI.5.4. *Widrow-hoff Learning procedure (alias Delta rule).* In order to increase the performance of the system, some learning procedure may be used. The most popular is the Widrow-Hoff procedure (also called Delta-Rule). Basically, the difference between the input and the output is used to correct the matrix $\widetilde{\mathbf{A}}$ in a step by step manner, as shown by the following equation:

$$\widetilde{\mathbf{A}}_{t+1} = \widetilde{\mathbf{A}}_t + \eta(\widetilde{\mathbf{f}}_i - \widehat{\mathbf{f}}_i)\widetilde{\mathbf{f}}_i^T \tag{14}$$

where $\widetilde{\mathbf{A}}_t$ denotes the matrix $\widetilde{\mathbf{A}}$ at time t, η an arbitrary positive constant, and i being randomly chosen. It can be shown (*cf.* Kohonen *et al.*, 1981, Kohonen, 1984) that that procedure converges to

$$\widetilde{\mathbf{A}}^* = \widetilde{\mathbf{F}}^+ \widetilde{\mathbf{F}} = \widetilde{\mathbf{U}}^T \widetilde{\mathbf{U}} \tag{15}$$

where $\widetilde{\mathbf{F}}^+$ denotes the pseudo-inverse of $\widetilde{\mathbf{F}}$.

XI.6. A neural model for face perception

In this section, we present some connectionist approaches to the problem of face processing. The first subsection exposes the selection of features for describing faces. The second subsection deals with the tranfer of learning for faces transformed by spatial filtering. In the third subsection, we show that an auto-associative memory will spontaneously extract face prototypes as a consequence of the storage of different faces.

XI.6.1. **Features for faces micro *vs.* elaborated features.** A first approach to code face information in long term memory is to propose a coding schema in term of *elaborated-features.* According to this coding scheme, a face will be decomposed into nose, mouth, chin, hair, etc. Each of these elaborated features will have one level among several possible levels (a nose can be straight, aquiline...). The recognition of a face will be equivalent to identify the correct level of the elaborated features.

Such an approach is indeed intuitively appealling. Leonardo da Vinci supported it in his *Tratado della Picturia.* This approach is also the basis for some well-known methods of reconstruction of faces by eyewitnesses (*e.g.,* Photofit or Identikit). However, although we are able, if needed, to describe a face using elaborated-features, psychological evidence suggests strongly that we do not use such a coding scheme to store information in long term memory:

1. Eyewitnesses find it hard to describe unknown faces with such a coding scheme (Loftus, 1979, Yarmey, 1979), even when they are asked to do so with faces that are present (Klatsky & Forest, 1984).
2. The reconstruction of faces using systems like Photofit or Identikit is very poor *even when the face to be reconstructed is available.*
3. Attempts to improve performance of eyewitnesses by improving their ability to code a face in elaborated features failed or even lead to a *decrement* in performance (Woodhead, Simmonds & Baddeley, 1979).
4. There is no correlation between the ability to describe a face in terms of elaborated features and the ability to recognize faces (Goldstein, Johnston & Chance, 1979).
5. By contrast, holistic coding activities increase memory task performance. Specifically, subjects who rate faces for subjective qualities (honesty, likeability, intelligence...) perform better on recognition tasks than subjects who rate faces for elaborated features. Even though the difference is small, it is however reliable (Patterson & Baddeley, 1975; Bower & Karlin, 1974). In general this effect is interpreted within the "levels of processing" framework. It is worth noting that subjects find meaningful such questions about the holistic properties of faces and show a clear agreement among subjects (Abdi, 1986).
6. Recognition of familiar faces is resistant to change in the elaborated features (glasses, change in hair style, absence or presence of facial hair; *e.g.,* , Galper & Hochburg, 1971; Pittenger & Shaw, 1975), or to modifications of the elaborated features as involved for example in aging (Seamon, 1980a,b; Bahrick *et al.*, 1975).
7. Transformations of faces that increase the saliency of the elaborated features (e.g. caricatures, drawing) actually *decrease* subjects' performance in perceptual and mnemonic tasks (Loftus & Bell, 1975; Hagen & Perkins, 1983; Tversky & Baratz, 1985).
8. The configuration of elaborated features is a better predictor than the elaborated features themselves in reaction time studies (Sergent, 1986a). Moreover, performance can be affected by transformations that do not affect the elaborated features, such as mirror transformation (McKelvie, 1986) or spatial filtering (Millward & O'Toole, 1986; Sergent, 1986b)

	Test **Normal**	**Test** **Low** **frequencies**	**Test** **High** **frequencies**
Learning Normal	.82	.73	.62
Learning Low frequencies	.66	.71	.54
Learning High frequencies	.61	.60	.77

TABLE 1. Probability of target detection (human subjects).

Thus, it seems preferable to code the faces in terms of "*macro features*" rather than in terms of elaborated features. A primary advantage of a low level code such as pixels or lines is that it avoids coding dilemmas such as deciding "what type of chin a given person has", or even, "is the type of chin a relevant feature for a face". Moreover, the necessity of assuming the existence of pre-processing systems able to extract elaborated features can be eliminated.

Such an approach has already be shown feasible and fruitful. O'Toole, Millward, and Anderson (1987) used pixels and lines in a series of computer simulations of the recognition transfer between spatially-filtered versions of the same faces. These simulations were done in conjunction with a series of experiments that tested the same type of recognition tranfer with human subjects.

In O'Toole *et al.* (1987), observers viewed low-pass (L), high-pass (H), or normal (N) unfiltered faces in a learning phase. Afterward, in a test phase the subjects were tested in a two-alternative forced choice recognition task. They were asked to find which of two faces was presented in the learning phase. These pairs of test faces, were presented either as L, H, or N faces, thus, all possible transfer cases were observed. The results of O'Toole *et al.* experiment (1987) are presented in Table 1.

O'Toole *et al.* also simulated the transfer using a connectionist model. Faces were digitized and represented by 900-pixel vectors. The communication between units was limited to a neigborhood of five units. To improve storage, the Widrow-Hoff procedure was used for five trials per stimulus. Test trials were simulated by pairing each face vector learned by the model with a new one. Each face-vector was then recalled, and the cosine between input and output was computed. The system was said to have "recognized" the face in the pair with the highest cosine. The results are given in Table 2. The general pattern of results of the experiment and the simulation is highly similar as indicated by the large correlation between the two sets of results [$r(7) = .88$]. Specifically, both human subjects and simulations show a clear

	Test Normal	Test Low frequencies	Test High frequencies
Learning Normal	.90	.80	.80
Learning Low frequencies	.80	.85	.70
Learning High frequencies	.60	.55	1.00

TABLE 2. Probability of target detection (computer simulation).

specificity encoding effect (*cf.* Tulving, 1982): the transfer is optimal when encoding and test are performed under the same condition.

O'Toole *et al.* demonstrated that a connectionist model can simulate the behavior of human subjects with a good accuracy. Recently, Ellis, Young, Flude & Hay (1987) have suggested that some repetition priming effect of face recognition may be best explained by connectionist models. In the following section it is shown that a auto-associative memory may exhibit spontaneously a *cognitive* behavior, namely the extraction of face-prototypes from a set of faces.

XI.6.2. **Face prototype abstraction by an auto-associative memory.**

XI.6.2.1. *Material.* Photographs of thirty-two Brown University undergraduates (16 males and 16 females) were used as stimuli. All the models wore white drapes to hide their clothing. None had beard or wore glasses in the photographs. In order to increase the "ecological validity" of the experiment, the photographs were not scaled nor were the poses (*i.e.,* some faces are photographed full view, some other are almost 3/4, however, none are profile). Some exemplars are displayed in Figure 1. This lack of control is thought to show the resistance to noise of connectionist memories. The photographs were digitized to give a $71 \times 71 = 5041$ pixel-vector whose components give the light intensity of the pixels (from 0 to 256).

XI.6.2.2. *Procedure.* The faces were given a unitary weight. The matrix $\mathbf{F}$ was rescaled to be a correspondence-matrix (i.e. the sum of its elements is one, and elements are positive or null). The luminance of the screen was maintained constant for each face, as a consequence the sum of each row of the matrix $\mathbf{F}$ is constant and equal to $f_{i.} = 1/32$. The units (i.e. the pixels or the neurons corresponding to the perceptive field) were weighted using an "informational" schema. The general idea was that the importance of a neuron should be inversely proportional to its use or to its probability of firing. The matrix $\mathbf{W}$ was a diagonal matrix with diagonal elements

FIGURE 1. A sample of the faces stored in the auto-associative memory

$$w_{j,j} = 1/\sum_{i=1}^{I} f_{i,j} \tag{16}$$

The matrix $\widetilde{\mathbf{A}}$ was constructed following equation (**4**).

XI.6.2.3. *Results.* According to the predictions of the theoretical section, the eigenvectors of the matrix $\widetilde{\mathbf{A}}$ correspond to the "optimal stimuli". As Figure 2 illustrates, they are remarkably *face-like*. The first[1] eigenvector expresses what is common to the set of faces and could be interpreted as a "face detector" (as opposed to anything being non-face like). The second third and forth vectors correspond to "prototypical faces" which, in a sense, are created by the memory from the original faces but are not actual faces. The following vectors can be interpreted as the "noise" of the system (i.e

[1]Actually, what we call here the first eigenvector is the second. The first eigenvector is, in fact, a *trivial vector* who expresses only the fact that the rows of the matrix **F** sum to one (*cf.* Lancaster & Tismenetsky, 1985).

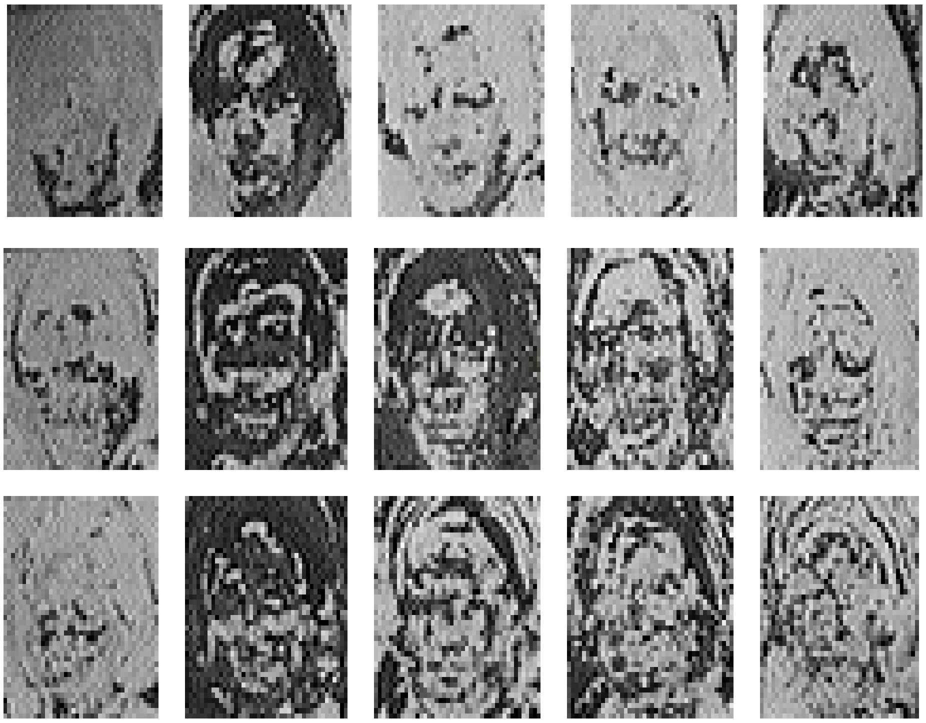

FIGURE 2. The first 15 eigenvectors of the auto-associative memory.
The corresponding eigenvalues (λ) are:
.020 .018 .012 .011 .010 .008 .008 .008 .008 .007 .007 .007 .007 .006 .006 .006.
The "percentage of inertia" ($\tau = \lambda / \sum \lambda$) are:
10.8 8.6 6.1 4.8 4.6 4.1 3.9 3.7 3.5 3.4 3.3 3.2 2.9 2.8 2.7.

the fact that the faces are not precisely scaled, that the poses are different, etc.).

XI.6.2.4. *Correspondence analysis interpretation of the eigenvectors.* As pointed out previously, the eigenvectors of the matrix $\widetilde{\mathbf{A}}$ can be interpreted as the generalized principal components of $\mathbf{F}$. With the particular set of weights that we have used, the eigenvectors perform a *correspondence analysis* of the faces (*cf.* , Benzécri, 1973; Greenacre, 1984). As a consequence, it is possible to draw a graph of the faces as they are "perceived" by the system by projecting the faces onto the eigenvectors. This procedure is equivalent

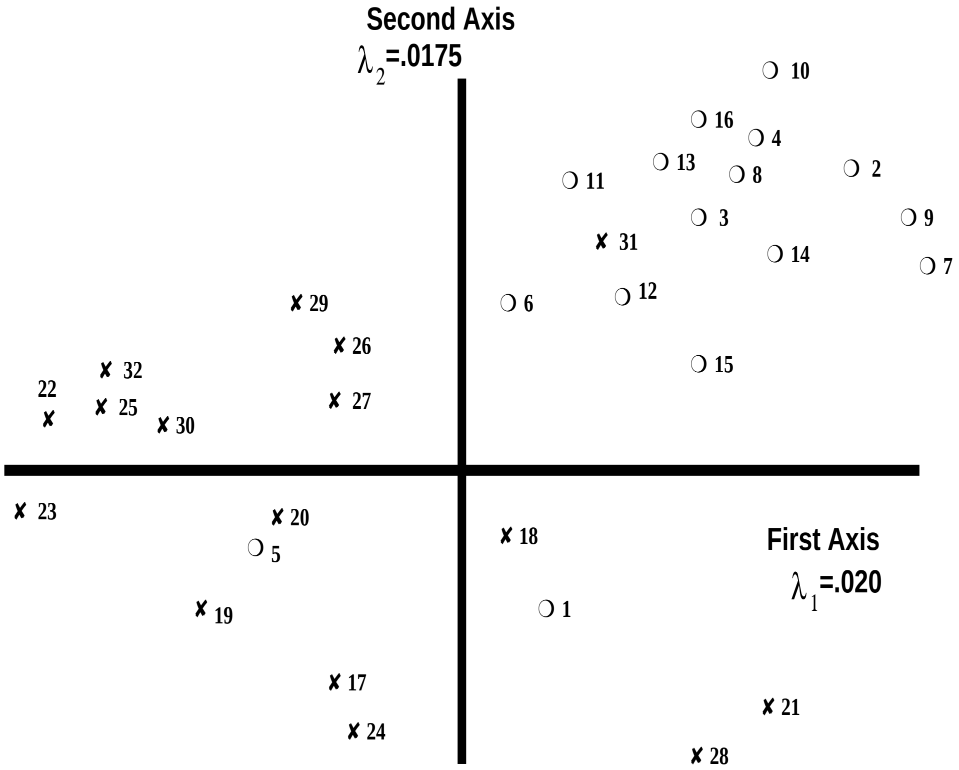

FIGURE 3. Correspondence analysis of the faces: Axes 1 & 2. $\lambda_1 = .0204$, $\tau_1 = 10.82$, $\lambda_2 = .0175$, $\tau_1 = 8.63$. Male faces are preceded by $\times$, female faces by o.

to correspondence analysis.[2] Figure 3 displays the positions of the faces in the space defined by the first two eigenvectors. As can be seen, the major opposition is between male and female, which is coherent with the empirical results obtained for humans.

XI.7. DISCUSSION

These first results clearly constitute an encouraging first step toward a comprehensive connectionist model for face perception. Some developments of this class of model looks particularly promising. A first approach will be to use these systems to simulate *"face-ogen"* (or *"prosopo-gen"*) abstraction (*cf.* Bruce, 1986; Bruyer, 1987), a process identical for faces to Morton's logogens (1969). In particular, connectionist models present spontaneously

[2]Recall that the strong property of correspondence analysis is the so-called "dual-representation": it is possible to represent in a same space both rows and columns of the data matrix.

a "priming" behavior: the recall of one face will prime the recall of similar faces stored in the memory (*cf.* Anderson & Mozer, 1981).

A second approach will aim toward the modeling of context effects (*cf.* Tiberghien, 1986). In order to do so, some extraneaous information (*i.e.,* visual, semantic, or some description of the situation) should be added to the code for faces. This can be done by embedding the units coding the faces within a larger set of units coding the extraneous information. Then, the model will present some *madeleine de Proust* behavior: a context associated with only one face will be able to reconstitute the face (*cf.* Knapp & Anderson, 1984).

A third approach will deal with the development of facial expressions. A prediction is that, if one face is stored with different expressions, then the first eigenvector should represent an expressionless face and the successive eigenvectors should correspond to different "faceless" expressions. Theses predictions will be dealt with in forthcoming studies.

However, the current model is only partial. Specifically, the faces should be pre-scaled even if approximately and they should be coded in micro-features before being stored in the memory. Moreover, some cognitive processes such as conscious search, decisionnel processes, imagery, etc. do not easily fit into that framework. Despite its shortcomings, the current model is able to give an account of some important cognitive processes such as tranfer of learning, prototype extraction, etc. and would be able also to mimic most of the current models of face perception (see Bruyer, 1987, for a recent review of these models).

References

Abdi, H. (1986). Faces, prototypes and additive tree representations, in H.D. Ellis, M.A. Jeeves, F. Newcombe, A. Young (Eds.), *Aspects of face processing*, Dordrecht: Nijhoff.

Abdi, H. (1987). Do we really need a contingency model for concept formation? *British Journal of Psychology*, **78**, 113–125.

Anderson, J.A., Mozer, M.C. (1981). Categorization and selective neurons. In G.E. Hinton & J.A. Anderson (Eds.), *Parallel models of associative memory*, Hillsade: Erlbaum.

Anderson, J.A., Rosenfeld, E. (1987). *Neurocomputing: some important papers*, Cambridge: MIT press.

Anderson, J.A., Silverstein, J.W., Ritz, S.A., & Jones, R.S. (1977). Distinctive features, categorical perception, and probability learning: some applications of a neural model, *Psychological Review*, **84**, 413–451.

Bahrick, H.P. (1983). Memory for people, in J. Harris (Ed.), *Everyday memory, actions and absentmindness,* London: Academic Press.

Bahrick, H.P., Bahrick, P.O., Wittlinger, R.P. (1975). Fifty-years of memory for names and faces: a cross-sectional approach, *Journal of Experimental psychology: General*, **104**, 54–75.

Benzècri, J.P. (1973). *L'analyse des données,* Paris: Dunod.

Bower, G.H., Karlin, M.B. (1974). Depth of processing of faces and recognition memory, *Journal of Experimental Psychology,* **103**, 751–757.

Bruce, V. (1986). Recognising familiar faces, in H.D. Ellis, M.A. Jeeves, F. Newcombe, A. Young (Eds.), *Aspects of face processing*, Dordrecht: Nijhoff.

Bruyer, R. (1987). *Les mécanismes de reconnaissance des visages,* Grenoble: P.U.G.

Davis, G.M., Ellis, H.D., Sheperd, J.W. (1978). Face recognition accuracy as a function of mode of presentation, *Journal of Applied Psychology,* **62**, 180–187.

Da Vinci, L. (1882). *Trattato della Picturia*, Vienna: H. Luwig.

Ellis, A.W., Young, A.W., Flude B.M., Hay, D.C. (1987). Repetition priming of face recognition, *Quaterly Journal of Experimental Psychology*, **39a**, 193-210.

Galper, R.E., Hochberg, J. (1971). Recognition memory for photographs of faces, *American Journal of Psychology*, **84**, 351–359.

Goldstein, A.G., Johnson, K.S., Chance, J.E. (1979). Does fluency of face description implies superior face recognition? *Bulletin of the Psychonomic Society*, **13**, 15–18.

Greenacre, M.J. (1984). *Correspondence analysis*, London, Academic Press.

Hinton, G.E., Anderson, J.A. (1984). *Parallel models of associative memory*, Hillsdale: Erlbaum.

Hopfield, J.J. (1982). Neural networks and physical system with emergent collective computational abilities. *Proceedings of the National Academy of Science, USA*, **79**, 6871–6874.

Hopfield, J.J. (1984). Neurons with graded responses have collective computational abilities, *Proceeding of the national academy of Sciences, USA*, **81**, 3088–2558.

Klatzky, R.L., Forrest, F.H. (1984). Recognizing familiar and unfamiliar faces, *Memory & Cognition*, **12**, 60-70.

Kohonen, T. (1977). *Associative memory: A system theoretical approach*, Berlin: Springer Verlag.

Kohonen, T. (1984). *Self organization and associative memory*, Berlin: Springer Verlag.

Lancaster, P., Tismenetsky, M. (1985). *The theory of Matrices*, New York: Academic Press.

Loftus, E. (1979). *Eyewitness testimony*, Cambridge (MA): C.U.P.

McClelland, J.L., Rumelhart, D.E. & Hinton, G.E. (1986). The appeal of parallel distributed processing, in D.E. Rumelhart & J.L. McClelland (Eds.), *Parallel distributed Processing*, Cambridge: MIT Press.

Millward, R.B., O'Toole, A.J. (1986). Recognition memory transfer between spatial-frequency analyzed faces, in H.D. Ellis, M.A. Jeeves, F. Newcombe, A. Young (Eds.), *Aspects of face processing*, Dordrecht: Nijhoff.

Morton, J. (1969). Interaction of information in word recognition, *Psychological Review,* **76**, 165-178.

O'Toole, A.J., Millward, R.B., Anderson, J.A. (1987). *A physical system approach to recognition memory for spatially transform faces*, unpublished manuscript: Brown University.

Patterson, K.E., Baddeley, A.D. (1975). When face recognition fails, *Journal of Experimental Psychology: Human Learning & Memory*, **3**, 406–417.

Pittenger, J.B., Shaw, R.E. (1975). Aging faces as viscal-elastic events, *Journal of Experimental Psychology: Human Perception & Performance*, **104**, 374–382.

Rumelhart, D.E., McClelland, J.L. (1986). *Parallel distributed processing*, Cambridge: MIT Press.

Seamon, J.G. (1980a). *Dynamic facial recognition: preliminary studies and theory*, unpublished manuscript.

Seamon, J.G. (1980b). *Memory & Cognition*, New York, O.U.P.

Sergent, J. (1986a). An investigation into component and configural processes underlying face perception, *British Journal of Psychology*, **75**, 221–242.

Sergent (1986b). Microgenesis of face perception, in H.D. Ellis, M.A. Jeeves, F. Newcombe, A. Young (Eds.), *Aspects of face processing*, Dordrecht: Nijhoff.

Tiberghien, G. (1986). Context effects in recognition memory of faces: some theoretical problems, in H.D. Ellis, M.A. Jeeves, F. Newcombe, A. Young (Eds.), *Aspects of face processing*, Dordrecht: Nijhoff.

Woodhead, M.M., Baddeley, A.D., Simmonds, D.C.V. (1979). On training people to recognise faces, *Ergonomics*, **22**, 333-343.

Yarmey, A.D. (1979). *The psychology of eyewitness testimony*, New York: Free Press.