Ref: O'Toole, A.J., Abdi, H., Deffenbacher, K.A., Valentin, D. (1995). A perceptual learning theory on the information in faces. In T. Valentine, (Ed.) *Cognitive and Computational Aspects of Face Recognition.* London: Routledge. pp. 159–182.

Chapter 8

A Perceptual Learning Theory of the Information in Faces

Alice J. O'Toole[*] , Hervé Abdi[†*] ,
Kenneth A. Deffenbacher[**],
and Dominique Valentin[*]

[*]**The University of Texas at Dallas,**
[†] **Université de Bourgogne à Dijon (France),**
[**]**The University of Nebraska at Ohmaha**

The study of human face processing has advanced considerably in recent years, from consisting of a collection of isolated empirical facts and anecdotal observations, to a relatively coherent view of the complexity and diversity of the problems tackled by a human observer when confronted with a face. This rapid progress can be traced to the proposal of comprehensive theories of face processing (cf. Bruce & Young, 1986; Ellis, 1975, 1986; Hay & Young, 1982), which have provided a theoretical framework for investigating human face processing within functional subsystems. These models have had much to say about the kinds of tasks subserved by the human face processing system (e.g., naming faces, extracting visual categorical information such as sex and age, etc.), and about the coordination of processing among these tasks (e.g., Young, McWeeny, Hay, & Ellis, 1986). They have also provided important constraints for making sense of neuropsychological data on patients with various face processing deficits (e.g., Bruyer, 1986). Despite the success of these models in guiding research efforts into many aspects of human face processing, they have provided somewhat less guidance in understanding the immensely complicated problems solved by the perceptual system in extracting and representing the richness of the perceptual information available in human faces. In recent years, it has been primarily from computational models that the difficulty of this problem and its importance to understanding human face processing abilities has come to be appreciated.

In the present chapter, we concentrate entirely on the problem of quantifying and representing the information in human faces. We propose a quantifiable theory of the perceptual information in faces and propose a simple statistical/neural network model to simulate the learning of this information. We believe that perceptual learning provides a useful analogy for the problem of selecting and "learning" the information in faces that is most useful for performing a given task. In many ways, acquiring expertise in processing human faces seems similar to learning many of the kinds of stimuli for which we typically call on perceptual learning as a theoretical construct. For example, how do we listen to music and extract features that enable us to accurately distinguish among different composers? Of primary importance is experience. Both the amount and diversity of experience we have with music constrain the kinds of distinctions we can make. While many people can distinguish a previously unheard Mozart piece from a previously unheard Prokofiev piece after hearing only a few different examples of Mozart and Prokofiev, the problem of distinguishing Mozart and Haydn pieces may require a great deal more experience and considerably more sophisticated distinctions. Second, explicit verbal instructions in learning seem to be little-used and of little use. Someone is more likely to tell you that Prokofiev is "gentler" or "rounder" than Stravinsky, than they are to give you a list of objective musical features for distinguishing the two. Also analogous to the face processing problem is the recognition/naming dissociation—the perceived familiarity of a piece of music is a compelling experience that occurs frequently even when we are unable to recall anything else about it, such as who wrote it, or where we heard it.

What seems to make faces similar to the kinds of stimuli to which we apply perceptual learning theory is the elusiveness of a feature list with which they can be described in a precise and globally agreed upon language. This intuition is supported by empirical evidence indicating that verbalizing a detailed physical description of faces can actually impair later recognition of the described faces (Schooler & Engstler-Schooler, 1990). As with music, human observers seem to be surprisingly comfortable applying abstract language to convey information about faces. We often describe faces using language like "mean-looking" or 'perky", which oddly enough, most of us seem to find helpful for distinguishing among faces, and which some studies have found to be beneficial for recognizing faces (e.g., Bower & Karlin, 1974).

While we believe a perceptual learning theory is applicable to learning faces in many respects, what makes faces very different from these other kinds of stimuli is the sheer quantity of experience we have with them and the strong importance they play in social interaction. In these ways, learning human faces is perhaps comparable to some aspects of natural language learning. We will develop this analogy shortly in the context of learning same- versus other-race faces.

The perceptual learning or statistical structure theory we propose represents faces using "features" derived from the statistical structure of a set of learned faces. With perceptual learning, the information most useful for distinguishing among faces within a learned set emerges as an optimal code. We propose to model this learning process with a computational autoassociative memory that operates on image-based codings of faces. This "memory" implements principal components analysis, which is a statistical procedure used for expressing a set of correlated variables in terms of a (smaller) set of uncorrelated variables (Hotelling, 1933). Further, the model we propose can be viewed in standard neural network terms as a parallel distributed processing system (McClelland & Rumelhart, 1988). It is important to note that we do not view this model as a replacement for current face processing models (Bruce & Young, 1986; Ellis, 1986; Hay & Young, 1982), but rather, as a perceptual "front-end"[1] to these more comprehensive systems. We would argue, however, that the nature of this front-end has strong implications for the efficiency and accuracy with which different face processing tasks can be performed. In fact, we believe that many robust empirical findings concerning faces are due, at least in part, to difficulties that can be understood in terms of perceptual constraints on the problem.

As a psychological model of processing the perceptual information in faces, autoassociative memories have several appealing properties that have to do primarily with the distributed nature of the storage mechanism in the model. Since faces share the same storage space, the representations of similar faces can interfere with each other in relatively natural ways. This is another way of saying that the memory is context sensitive and so its performance will depend on similarity relationships within the entire set of faces on which it is trained. At the level of individual faces, this model makes interesting predictions about the distinctiveness of individual faces. At the level of sets of faces, the model's ability to act as a statistical analysis tool that operates on physical codings of sets of faces allows for some interesting explorations of the effects of the heterogeneity of the faces learned on the model's recognition and classification abilities.

Before proceeding, it is perhaps worth mentioning the goals of such a model. Any psychological model of the information in faces should meet the following criteria. First, it should be adequate to support the diversity of face processing tasks that humans achieve. Additionally, with a psychologically

[1] A term we borrow from Bruce (1988).

relevant, quantifiable model of the information in faces, it should be possible to predict the quality of information available for any given task (e.g., sex classification or recognition). While these criteria are very far from being met by any current model, including the one we propose, we believe that much can be learned by exploring the extent to which informational or perceptual constraints alone can account for some well-known phenomena associated with human face processing. Hence, one goal of testing this model will be to determine where cognitive or semantic factors must be postulated to account for these phenomena.

This chapter is organized as follows. First, we outline a sample of approaches that have been used for specifying the information in faces. Second, we give a brief definition of the autoassociative neural network model. In the next section we demonstrate, first, that the computational model is capable of solving some useful face processing tasks. We then review some recent studies suggesting its potential psychological relevance. Finally, we discuss the relationship of a perceptual learning representation to the approaches discussed in the representational issues section.

1. Representational issues

An overview of the psychological and computational literature on face processing reveals a variety of attempts to "specify" the kinds of physical information in human faces. Unfortunately, while several of these approaches are related, it is often very difficult to make concrete comparisons between them (and sometimes even within an approach) due to differences in the way definitions have been operationalized or to differences in the kinds of data they have been used to describe. In this section of the chapter, we outline a sample of the approaches that are commonly found in the literature and point out common threads in these approaches. While concrete comparisons are not possible in many cases, it seems unreasonable to ignore the important ways in which the approaches may be tapping similar kinds of coding principles. With that caution in mind, our primary purpose in this endeavor is to lay a foundation for comparing the proposed perceptual learning approach to these other well-studied approaches.

In the psychological literature, the most frequently encountered distinction made concerning the kinds of information in faces is a qualitative one drawn between feature-based and configural information. As noted by Bartlett and Searcy (1993), this is perhaps better described as a family of distinctions that have been referred to variously as component (piecemeal) versus configural (e.g., Carey & Diamond, 1977), global versus local (Navon, 1977), and isolated features versus second-order relational information (Rhodes, Brake, & Atkinson, 1993). Bruce (1988) defines a feature as "a discrete component part of a face such as a nose or a chin", whereas configural information refers to the "spatial interrelationship of facial features" (p. 38). The primary practical difficulty encountered in testing the relative importance of feature-based versus configural information in face processing concerns the problem of selectively varying the two sources of information. Thus, while it seems possible to selectively alter facial configuration, it is not clear that it is possible to selectively vary feature-based information. As noted by Sergent (1984), changes in the features of faces, such as switching the noses of two faces, necessarily change some properties of the configuration. In fact, Rhodes et al. (1993) note that even simple configuration changes can change the dimensions of what may be plausibly considered isolated features (e.g., moving the mouth up or down in a face changes a feature like upper lip length).

In practice, the manipulation of configural information has been operationally defined in experiments in a wide variety of ways. For example, Young, Hellawell, and Hay (1987) distorted configural information in composite faces[2] by horizontally misaligning the top and bottom halves of the faces. Additionally, inversion of the eyes and mouth in an upright face, the primary manipulation in the Thatcher illusion (Thompson, 1980), is generally considered a disruption of configural information. (See Stevenage, this volume). While both manipulations disrupt the spatial configuration of the features, intuitively, they seem to be very different kinds of manipulations. Additionally, both entail some change to the facial features. Aligning and misaligning the top and bottom halves of faces change the shape of individual features in the center of the faces – like the nose and ears. Likewise, the inversion of the mouth and eyes in an upright "Thatcherized face" changes the form of the eyes/mouth to the point of grotesqueness. Selective configuration manipulations are most closely approximated in studies that move features relative to one another (e.g., Bartlett & Searcy, 1993; Sergent, 1984). Interestingly, what seems to operationally bind all three of these manipulations together are the measurable differences in human performance with these faces inverted, as opposed to upright. We will discuss some aspects of the effects of inversion in the final section of this paper. What has been gained by using a configural/feature-based dichotomy is the understanding that for a human observer, the face is clearly more that the sum of its parts. This has provided insight into the nature of the perceptual unit comprised by a face and has been useful for linking face studies to the larger literature on processing visual features.

[2] Faces made by combining two faces, in this case, the top half of one face and the bottom half of a second face.

A second approach to describing the information in faces that has been explored in psychological studies is a quantitative analysis of the spectral information in face images. As imaged on the retina, faces are two-dimensional spatial patterns of light intensities that can be measured using standard Fourier analysis. The appeal of quantifying the information in faces in this way is two-fold. First, converging evidence in neurophysiology and psychology is consistent with the notion that the visual system analyzes input at several spatial resolution scales (cf. Shapley, Caelli, Grossberg, Morgan, & Rentschler, 1989, for a thorough review). Thus, spatial frequency preprocessing of faces is consistent with what is known about early visual processing. Like the principal components analysis model we propose, spatial frequency analysis represents an image as a weighted combination of basis functions – specifically, trigonometric (sine and cosine) functions of different frequencies, amplitudes, and phases. High frequencies carry finely detailed information, whereas low frequencies carry coarse, shape-based information. This continuum of low to high frequency information is in some ways related to the feature-based versus configural dichotomy in that the low frequency information tends to capture global form information, whereas the high frequency information tends to capture local information.

A second advantage of the spatial frequency analysis is that it is an objective physical measure and hence can be assessed directly in individual faces and varied selectively. While this information must be specified with respect to faces rather than in visual angle dimensions (i.e., cycles per face rather than cycles per degree of visual angle), within this context it can serve as a useful tool for objectively quantifying the information in faces. Though most early work using a spatial frequency quantification of faces was aimed at discovering the minimal spectral information for recognizing faces (e.g., Ginsburg, 1978; Harmon, 1973), the primary contribution resulting from studies that have varied spatial frequency content in faces has been the realization that different kinds of information may be optimal for different face processing tasks. For example, Sergent (1986) showed that human observer performance in the tasks of face identification, male/female categorization, and a semantic categorization (whether the familiar face was a professor, graduate student, etc.) interacted with spatial frequency content and visual hemisphere field of presentation, thus indicating the necessity of considering the task when evaluating the importance of different kinds of information in faces.

The spatial frequency approach to quantifying the information in faces forms a bridge between the qualitative feature-based versus configural distinction and the focus on the functional value of information that is common in computational modeling approaches. We review computational modeling approaches only briefly here since very recent, thorough reviews of both non-connectionist (Samal & Iyengar, 1992) and connectionist (Valentin, Abdi, O'Toole, & Cottrell, in press) models are available. Typically, computational models have been designed to solve only a single face processing task. For example, the neural network models of Brunelli and Poggio (1992) and Golomb, Lawrence, and Sejnowski (1991), as well as the discriminant analysis model of Burton, Bruce, and Dench (1993)[3] were designed to classify faces by sex. In the models of Burton, et al. (1993) and Brunelli and Poggio (1992), the features used were those that seemed likely to be informative for the task. In fact, one focus of the Burton et al., study was to apply the discriminant analysis for evaluating the utility of individual features for this purpose. Using both feature-based and configural codings derived from two- and three-dimensional face representations yielded reasonable levels of accuracy. Unfortunately, however, in all of their coding attempts, model misclassifications of individual faces by sex were unrelated to human misclassifications, indicating that the information used by humans may be quite different from that used by the model. Likewise, Brunelli and Poggio (1992) used measures of the dimensions of faces, including thickness of eyebrow, breadth of the face, and six chin radii (i.e., the length of lines drawn at varying angles from the center point of the mouth to the chin contour). These features were input to a hyper basis function network that learned to classify faces by sex. Even though these representations have been useful for accomplishing the specific task for which they have been designed, the facial representations they employ are perhaps not optimal for other tasks, such as face recognition.

One problem with pre-selected feature sets is that they often discard important information about the texture and internal shape contours of the face. The autoassociative model we propose and several other related computational models (e.g., Cottrell & Fleming, 1990; Golomb, Lawrence, & Sejnowski, 1991; Sirovich & Kirby, 1987; Turk & Pentland, 1991) have used a normalized, pixel-based coding of faces. This is unabashedly a "kitchen-sink" approach to the problem that has both advantages and disadvantages. The primary advantage is that no information is discarded a priori. Thus, geometric representations are coded implicitly, but in addition, detailed texture and shape information are preserved. For faces we believe that this kind of pixel-based code is a reasonable approach for two reasons, a theoretical one and a practical one. From a theoretical point of view, when we consider face recognition by comparison to object recognition, it becomes evident that we mean different things by "recognition". The processes implied by

[3] While presented by these authors as a statistical discriminant model, the analysis is formally equivalent to a perceptron type of neural network (Rosenblatt, 1958; cf. Abdi, 1994 for equivalence proof).

face recognition are operating at a different level of the processing hierarchy. Specifically, in most object recognition applications, the goal of the task is to identify a subset of pixels in an image as an instance of a particular object, a chair, for instance. In standard cognitive psychology terminology, this task is a basic level category classification (Rosch & Mervis, 1975). Little importance is placed on the chair being a particular chair, one you know or have sat in previously, for example. For faces, the identification of a subset of pixels in the image as a face represents only the first step of the process. Additionally, you would like to know if the face is one you know. To approach the person to begin a conversation, often you would like to know the age, sex, and even current mood of the person. To accomplish these latter tasks, internal shading and texture information is likely to be very useful. In fact, if the object recognition task were aimed at this level of information, a reevaluation of the coding schemes generally used for object recognition would be in order. So, for example, if your task were to be the identification, not of a car, but of your car, from among a parking lot full of similar models, you would need to consider subtle textural information, including "dings" and the dirt layer texture, and so on.

The primary disadvantage of a pixel-based approach is that it does not create a translation or view-invariant representation. From a practical point of view, however, good algorithms exist for finding a face in an image (Turk & Pentland, 1991), and the computational problem of scaling and aligning faces (necessary for processing faces in the present approach) is easy to solve. Using a pixel code, therefore, allows modelers to concentrate on the problems of recognition and visual categorization that make faces a qualitatively different kind of visual stimulus from the other kinds of objects with which we interact.

2. Autoassociative Model Definition

In this section we will give a very brief definition of the autoassociative memory model and will show how faces can be described as a weighted sum of the eigenvectors extracted from the autoassociative matrix. A very detailed presentation of this model and its application to face processing can be found elsewhere (e.g., Valentin, Abdi, & O'Toole, in press; for a tutorial presentation see Abdi, in press).

An autoassociative memory matrix is constructed as the sum of outer-product (i.e., cross product) matrices for a set of stimuli coded as vectors:

$$\mathbf{A} = \sum_i \mathbf{f}_i \mathbf{f}_i^T \tag{1}$$

where $\mathbf{f}_i$is the i-th face, coded as a pixel vector consisting of the concatenation of the rows of the face image, and where the faces are assumed to be normalized vectors (i.e., $\mathbf{f}_i^T\mathbf{f}_i = 1$). Simply stated, **A** contains a measure of the covariance of all possible pairs of pixels in the set of learned faces. Recall of the i-th face from this matrix is achieved as follows:

$$\widehat{\mathbf{f}}_i = \mathbf{A}\mathbf{f}_i \tag{2}$$

where is the system estimate of $\mathbf{f}_i$. The quality of the output face estimate is measured by comparing the "retrieved" (reconstructed) image with the original image, using the cosine (i.e., normalized correlation) of the angle between the vectors $\widehat{\mathbf{f}}_i$ and $\mathbf{f}_i$.

Like any positive semi-definite matrix, the matrix **A** can be expressed as a weighted sum of the outerproducts of its eigenvectors :

$$\mathbf{A} = \sum_i \lambda_i \mathbf{e}_i \mathbf{e}_i^T \tag{3}$$

where λ_i is the i-th eigenvalue and $\mathbf{e}_i$ is the i-th eigenvector of **A**. Retrieval of a face vector from this matrix can be illustrated by rewriting Equation 2 and substituting Equation 3 for **A** as follows:

$$\widehat{\mathbf{f}}_i = \lambda_1(\mathbf{f}_i \cdot \mathbf{e}_1) + \lambda_2(\mathbf{f}_i \cdot \mathbf{e}_2) + \cdots + \lambda_n \tag{4}$$

where the weights are $(\mathbf{f}_i \cdot \mathbf{e}_j)$, which is equal to the dot product between the i-th face vector and the j-th eigenvector, and where n is the rank of the matrix. In other words, each retrieved face can be represented by a weighted sum of eigenvectors, and thus, individual faces are made by putting together these "eigen-images" in different weighted combinations.

The storage capacity of an autoassociative matrix can be improved by applying error correction in the form of the Widrow-Hoff or delta rule (Duda & Hart, 1973) during learning of the faces. Error correction can be implemented iteratively as follows:

$$\mathbf{A}_{[t+1]} = \mathbf{A}_{[t]} + \eta(\mathbf{f}_i - \mathbf{A}_{[t]}\mathbf{f}_i)\mathbf{f}_i^T \tag{5}$$

where η is a learning rate parameter. Simply seen, the matrix **A** is updated at time $t+1$ by calculating the "error", or difference between the actual face and the model estimate at time t, $(\mathbf{f}_i - \mathbf{A}_{[t]}\mathbf{f}_i)$, and reteaching this "difference" vector to the model *via* the outerproduct rule (Equation 1). This process is repeated for all faces over many iterations. The learning parameter η can be set to be very small or can decrease exponentially such that finer and finer changes are made to the matrix over time. From the principal component point of view, the effect of this error correction is equivalent to dropping the eigenvalues from Equations 4 or 3 (cf. Abdi, 1994).

3. Demonstrations of the Model's Ability to Perform Useful Face Processing Tasks

3.1. *Recognition.* How can this model be applied to the problem of distinguishing learned from new faces? In general, we have begun by training a model with a large number of full-face images. The model can then be tested by "recalling" both learned and novel faces using Equation 2. We can then evaluate the "quality of the representation" for any given learned or new face by computing the cosine between the original and model-reconstructed faces. The higher this cosine, the more faithful the autoassociative memory's representation of the face. This cosine measure is a sort of model resonance or "feeling of familiarity" with the face. To test the model's ability to recognize faces, we would like to show that the model's "feeling of familiarity" is higher for learned than for new faces. We have used signal detection theory (SDT) methodology for this purpose. With this procedure, we define the signal+noise distribution as the OLD or learned faces and the noise distribution as the NEW or unlearned faces. Using the cosine between the original and reconstructed faces, the procedure assigns each face to the category of OLD or NEW as follows. If the cosine for a given face exceeds some criterion, the face is assigned to the OLD face category, otherwise, the face is assigned to the NEW face category. Faces, then, can be categorized as hits, false alarms, misses, and correct rejections. To select the criterion we generally use the "ideal observer" criterion, midway between the mean of the cosines for the OLD and NEW faces. A d' is then calculated in the standard manner.[4] A complete ROC curve can be calculated simply by sliding the criterion along the cosine histogram (cf. O'Toole, Deffenbacher, Abdi, & Bartlett, 1991).

On the average, cosines for the learned faces exceed those for the novel faces, indicating that the model, within capacity limits, can distinguish the learned from the novel faces (O'Toole, Millward, & Anderson, 1988; O'Toole, Deffenbacher, et al., 1991; and O'Toole, Abdi, Deffenbacher, & Valentin, 1993). We will discuss the relationship between model and human recognition performance in the section considering the application of the model to psychological issues.

3.2. *Visually-derived Semantic Categorization.* Variations of the present model have been used for the categorization of faces along the visually-derived semantic dimensions of race and sex. O'Toole, Abdi, Deffenbacher, & Bartlett (1991) have shown that when a heterogeneous set of male and female, Japanese and Caucasian faces is learned by the model, information about the race and sex of faces can be found in a single or small subset of eigenvectors. As noted previously, a face can be represented by the set of weights needed to

[4] See Turk & Pentland (1991) for another recognition algorithm.

combine the eigenvectors to reconstruct it. Figure 1 shows the weight profiles of the faces divided by race. As can be seen, the weight on the second eigenvector appears to provide good information about the race of a face. We tested this formally by taking the mean of the mean weights for the Japanese and Caucasian faces and using this grand mean as a criterion. Race membership predictions were made by assigning faces with weights exceeding the criterion to one race and faces with weights less than this criterion to the other race. Using only the weights for the second eigenvector yielded correct race predictions for 88.6% of the faces.[5]

One advantage of this approach is that the present model is not trained explicitly to classify by race, but rather is trained to recognize faces. Hence, the classification information is a natural part of the information used in the recognition task.

4. Demonstrations of the Feasibility of Applying the Model to Psychological Issues

4.1. *The other-race effect.* It is well-known that people are better at recognizing faces of their own race than faces of other races. While a variety of explanations have been proposed, a very simple one can be framed in terms of perceptual learning. By this account, long-term repeated exposure to the many faces of one race allows the perceptual system to make effective use of subtle variations in the form and configuration of the facial features of the "same-race" faces (i.e., those learned). Unfortunately, other-race faces are not well-characterized by these highly specialized features, and so we are less accurate at recognizing such faces. This account of the other-race effect is not unlike what is known about learning one's own native language. With a great deal of exposure to a single language, people become adept at processing the features of the language that are most useful for distinguishing between speech sounds in that language. This occurs at the cost of loosing an ability to distinguish speech sounds that are important in other languages, but are not particularly useful in one's own language.

Using Japanese and Caucasian faces, we (O'Toole et al., 1991) simulated a "face history" by training a neural network to recognize a large number of faces of one race, (a "majority" race), and a lesser number of faces of another race, (a "minority" race). We found that the model "perceived" or

[5] Note that for visually-based classifications (e.g., sex) it is possible to achieve 100% correct categorizations of the learned faces by combining all eigenvector weights. Performance on new faces, while less than perfect, is well-above what can be achieved with a single eigenvector (cf. Abdi, Valentin, & O'Toole, in preparation). Additionally, while using the weights on all eigenvectors allows for perfect performance on the learned faces, only the eigenvectors with relatively larger eigenvalues contribute to the model's ability to generalize sex information to unlearned faces (Abdi, et al., in press).

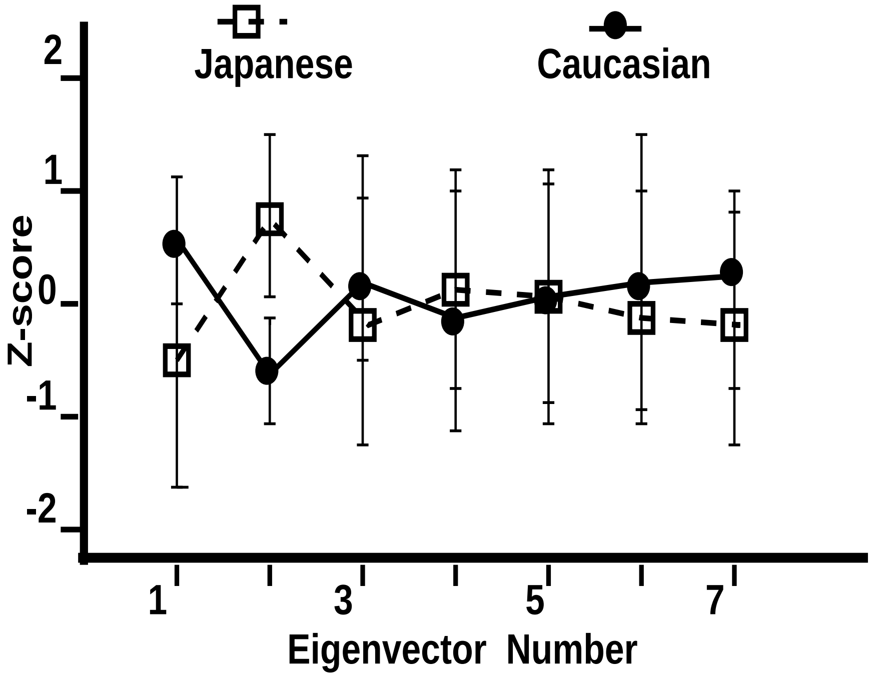

FIGURE 1. Mean coefficient z-score profiles of Caucasian and Japanese faces. Error bars show the standard deviations of the z-scores. The best race separation is achieved with the second eigenvector.

represented novel faces from the minority race with codings that were more similar to one another than the codings with which it represented novel majority faces. This is reminiscent of the oft-noted feeling that other-race faces "all look alike" and is consistent with Bruce's (1988) and Shepherd's (1981) suggestion that other-race faces are less recognizable because they are perceived as more similar to one another. Both authors suggested that the higher perceived interface similarity for other-race faces is due to the fact that the dimensions of the similarity space are determined mostly by same-race faces. We also found that the model recognized majority faces more accurately than minority faces in an episodic memory task.[6] This is the classic"other-race effect" phenomenon.

[6] Both Japanese and Caucasian faces served alternately as the majority and minority race faces for tests yielding conclusions 1. and 2. Due to a shortage of Japanese faces for the episodic recognition task, this was tested with only Caucasian faces as the majority race faces.

4.2. *Typicality and recognizability.* The relationship between rated typicality and recognizability of faces has been demonstrated by a number of investigators (e.g., Light, Kayra-Stuart, & Hollander, 1979). This relationship has been interpreted in terms of the existence of a facial prototype, with typical faces being less well recognized than unusual faces. Recent findings by Vokey & Read (1992), however, indicate that rated typicality is a more complicated concept than had been thought previously. Applying a principal components analysis[7] to faces rated by human subjects for typicality, memorability (i.e., "one that the observer thought would be easy to remember"), familiarity (i.e., "a face that they believe they may have seen around campus"), attractiveness, and likability, they show that the rated typicality of faces is composed of two orthogonal components:

1. a general familiarity component consisting of a positive manifold of typicality, familiarity, attractiveness, and likability,
2. a memorability component showing typicality inversely related to the rated memorability of a face.

This suggests that human observers are basing their typicality ratings on two independent aspects of the faces, dissociable via their independent relationship to attractiveness, likability, and familiarity, on the one hand, and to memorability on the other hand.

While the data of Vokey and Read (1992) are robust and replicable (O'Toole, Deffenbacher, Valentin, & Abdi, 1994), little is understood about what makes a face typical versus atypical with respect to the two orthogonal components. Recently, we (O'Toole, et al., 1994) have extended the two-component typicality results of Vokey & Read (1992) by adding a variable derived from the autoassociative memory to the coding for each face in the principal components analysis. Specifically, we added to the human rating and recognition data the model's "feeling of familiarity" measure (cosine) for each face. We then applied principal components analysis to the combined performance, rating, and model data for the faces. The results indicated a separation of the multidimensional space into performance and rating subspaces. The rating subspace replicated the typicality component results found by Vokey and Read (1992). The performance axes were interpretable as a criterion ("indictability") axis and an accuracy axis. The cosine measure taken from the autoassociative memory loaded more strongly on the accuracy axis than did any of the observer ratings. In other words, the model measure related more strongly to human performance accuracy than did any of the human ratings. Additionally, by looking at particular faces that contributed strongly to the different typicality components, the model

[7] This principal components analysis was applied to subject judgments, not to a representation of the stimuli. Note also, that a Varimax rotation was applied to the space after the principal components analysis.

gave insight into the reason for the separation of the typicality components in the human data. Faces that contributed strongly to the memorability component were characterized by a distinctive localized feature such as an unusual mouth expression, a grimace, for example, whereas faces important for the familiarity component were characterized by more global deviations, such as unusual face shapes.

These results suggest the importance of considering faces as perceptual stimuli that provide observers with very rich, elaborate information that they use quite effectively, but which they cannot capture very well in discrete verbal ratings.

4.3. *Recognition and the Perception of Visually-Derived Semantic Information.* Interestingly, a good likeness of a face can be captured using only a subset of the eigenvectors, those with larger eigenvalues (cf. Sirovich and Kirby, 1987). While this representation is optimal in a least squares error sense for approximating the face, we have noted that in eliminating eigenvectors from the reconstructed faces, the likeness or general perceptual quality of the face decreases more by eliminating ranges of eigenvectors with smaller eigenvalues than by eliminating the"more important" eigenvectors, those with larger eigenvalues (O'Toole, et al., 1993). For example, Figure 2 displays an original face and the appearance of the face produced by eliminating different ranges of eigenvectors.

This observation, in combination with the results indicating the importance of the eigenvectors with larger eigenvalues for determining visual category information from faces (O'Toole, Abdi, et al., 1991), motivated us to examine the importance of different ranges of eigenvectors for recognition and sex categorization. O'Toole, et al., (1993) trained the model with a large number of male and female young adult Caucasian faces and reconstructed both the learned faces and a second set of faces not learned by the model, while varying the range of eigenvectors used in the reconstruction. The model was tested for recognition using SDT methodology. The *d*' for discriminating learned and new faces across this range appears in Figure 3 and shows that the discriminability of the image information provided by the model peaks bimodally, with the most useful information found in the eigenvectors with smaller eigenvalues. This indicates that the least squares error minimization strategy is perhaps not the best one for the purposes of recognition.

We then carried out a test of the model's ability to classify faces by sex across the eigenvectors. We did this for each eigenvector by computing point biserial correlations between the eigenvector weight for each of the faces and the face sex.[8] Figure 4 shows the cumulative proportion of explained

[8] As defined by 0 and 1 for male and female faces, respectively.

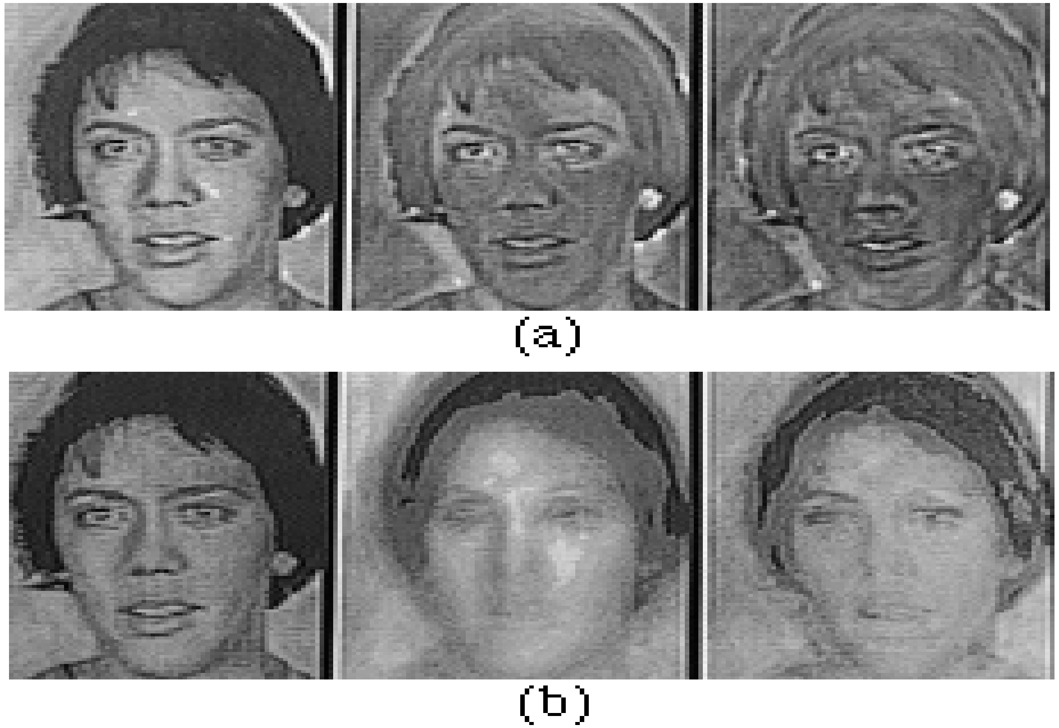

FIGURE 2. From left to right: a.) the original face; reconstructions using b.) the first 20 eigenvectors; c.) the first 40 eigenvectors; d.) all but the first 20 eigenvectors; e.) all but the first 40 eigenvectors.

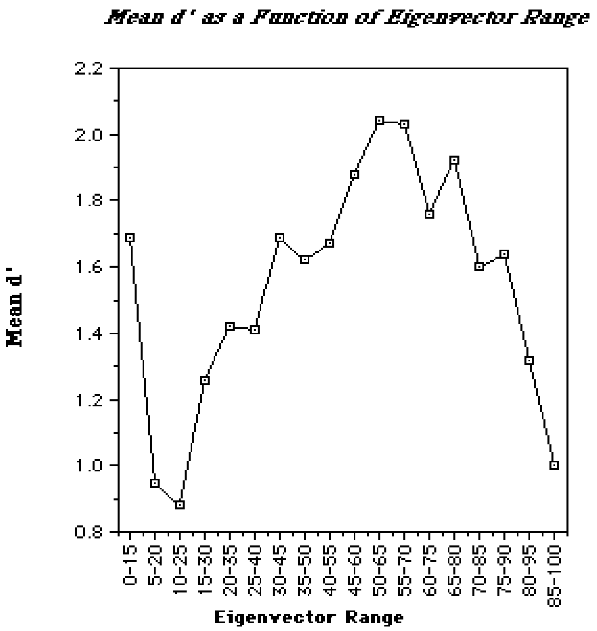

FIGURE 3. Mean d' peaks in 2 separate ranges of eigenvectors, with best performance in a range of eigenvectors with smaller eigenvalues.

variance for sex classification across the eigenvectors.[9] As can be seen, the best information for predicting face sex is found in the eigenvectors with the largest eigenvalues. The second eigenvector was particularly useful ($r =$ .66) as is demonstrated in Figure 5 (O'Toole, et al., 1993). From left to right, the figure shows the first eigenvector, the second eigenvector, the first eigenvector plus the second eigenvector, and the first eigenvector minus the second eigenvector. Adding the second eigenvector to the first produces a masculine looking face, whereas subtracting the second eigenvector from the first produces a feminine looking face. This is particularly striking in that the second eigenvector, at first glance, reveals little information that would appear to be relevant to the sex of the face.

Combined, the data from the recognition and sex categorization tasks show that the model contains information for both tasks, but that the optimal information for each task is found in different ranges of eigenvectors. For the visually-derived semantic classification by sex, the eigenvectors with

[9] Only eigenvector weights that correlated significantly (significance of r test, $p < .05$) are included. One-hundred percent of the variance would be explained if all eigenvectors were included.

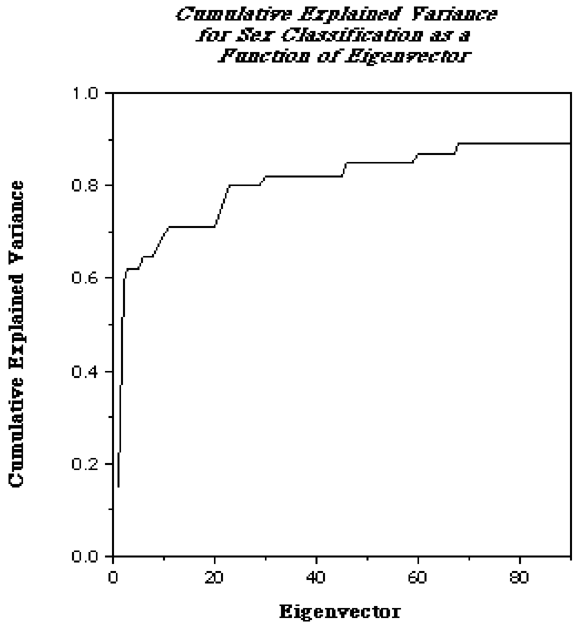

FIGURE 4. Weights on the eigenvectors with larger eigenvalues account for most of the explained variance in sex classification

larger eigenvalues provide the best information. For the recognition task, the eigenvectors with smaller eigenvalues provide the most useful information.

5. Representational Issues Revisited

In this section, we present some ways in which the statistical structure/perceptual learning theory of the information in faces can be related to all three representational approaches discussed previously:

1. the configural versus feature-based distinction
2. the spatial frequency approach
3. the functional emphasis of computational models.

We stress, again, that while concrete comparisons are not possible, it seems unreasonable to ignore important ways in which the approaches may be related. We do this with the goal of building links that may enable a more coherent view of representational issues found in the literature.

To begin, the informational components of the perceptual learning/statistical structure theory are eigenvectors. Each eigenvector can be measured in terms

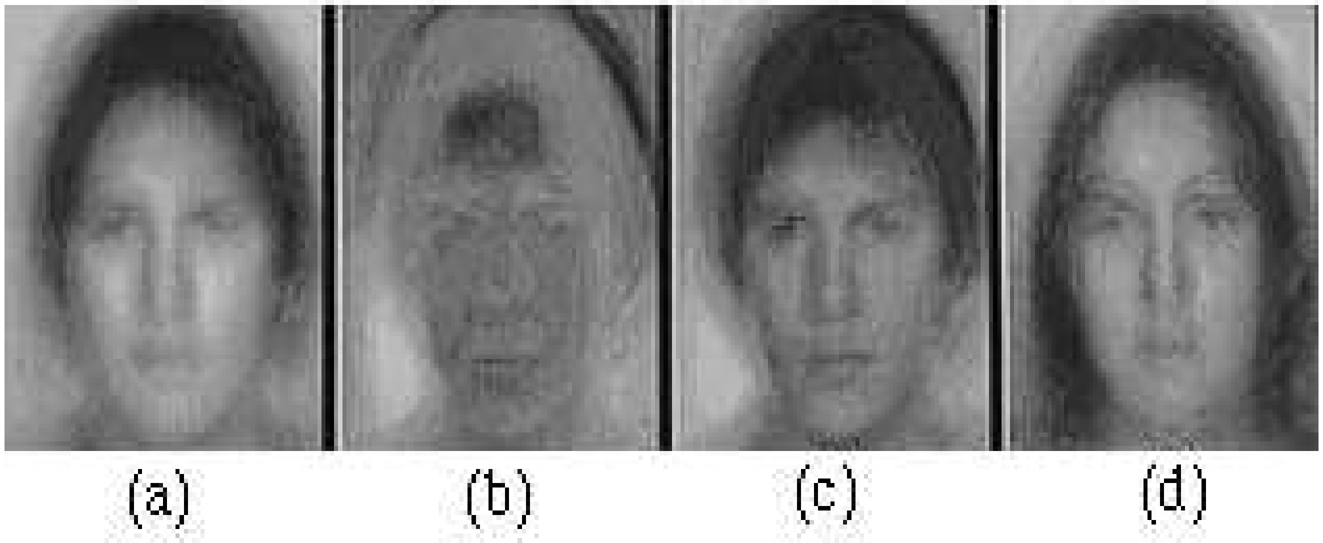

FIGURE 5. From left to right: a.) first eigenvector; b.) second eigenvector; c.) $a + .4b$; d.) $a - .4b$. Positive weights of the second eigenvector give rise to faces that appear masculine, whereas negative weights give faces that are feminine in appearance.

of its "importance" to the general face representation system as the proportion of variance it explains in the pixel-by-pixel autoassociative matrix. A primary aspect of this concept is that "importance" for any given eigenvector can come from two sources, which can be described in intuitive terms as follows. First, not too surprisingly, eigenvectors that are useful for making distinctions relevant for many faces will explain relatively larger proportions of variance than eigenvectors useful for making distinctions relevant for only a few faces. Major visually-based categorical distinctions such as sex, race, and age are examples of distinctions relevant for many/all faces. This is likely to be one reason that we have found the eigenvectors most related to the sex and race of a face to have relatively large eigenvalues.

A second source of "importance", which is perhaps less obvious, concerns the relative intensity of different kinds of information in faces. By contrast to the first source, which is related to the number of faces for which a particular eigenvector is important, this source is related to properties of the face image per se, and becomes evident when the computational problem is viewed in terms of a spatial frequency description of faces. The statistical structure of natural images, of which faces are an example, is such that low spatial frequency components generally have higher amplitudes than high spatial frequency components.[10] An analogy to auditory stimuli may prove helpful for understanding the point. The amplitude of a spatial frequency component is like the "loudness" of a frequency component in an auditory stimulus. For face images, the "loudest" components tend to be the lower frequency, shape-based properties of faces rather than the higher frequency details. In a principal components analysis, therefore, low frequencies will generally account for more variance than high frequencies and hence will be associated with eigenvectors with larger eigenvalues (i.e., be more important). In fact, a face reconstructed with different subsets or ranges of eigenvectors from a principal components analysis (cf. Figure 2) appears to vary systematically in spatial frequency content. While we have not confirmed this formally, principal components analysis on faces would appear to naturally implement some aspects of spatial frequency filtering—primarily as a function of the relative intensities of the different frequency components.

To what extent does the relative amplitude relationship of different frequencies impact performance on face processing tasks? Returning to questions concerning the optimality of different information for different tasks, some previous findings become more clear. The higher amplitude of lower

[10] This "energy" differential between high and low frequency information in faces is one of several factors outlined by Sergent (1989) as potentially accounting for diverging conclusions on human studies of the importance of spatial scale information in face processing.

frequency components[11] should make the global configural information easier to detect and also detectable/resolvable from a further distance. Additionally, as Sergent (1986) has illustrated, using psychological experiments varying the spatial frequency content of faces, and as the principal components model tends to confirm computationally, these low frequency, higher amplitude components seem to be particularly useful for visually-based categorizations such as sex and race categorizations. It is possible that the difference in the intensity of these components explains some of the reasons why sex classification can be done more quickly than other semantic categorizations and identifications (Sergent, 1986).

The particularly useful nature of the high amplitude, low frequency, global information for making sex and race classifications is likely to be a second reason why we have found eigenvectors useful for these classifications to have relatively large eigenvalues. Presumably, however, the number and intensity factors are independent. Intuitively, we might imagine these factors dissociating in a problem like classification by age, where low intensity, high frequency, texture information such as wrinkling might actually be useful for large scale distinctions among many faces. We have not yet explored this problem.

The perceptual learning and spatial frequency approaches are different, however, in the sense that spatial frequency analysis measures the information in a single face image, whereas principal components analysis measures the information in a set of face images. The representation of faces that emerges is sensitive, therefore, to the model's face history. This is a useful component of any computational model interested in simulating human face processing phenomena like the effects of face typicality and the "other-race effect". A good example of how the properties of the set of faces affect model reconstructions of faces can be found in Valentin, Abdi, & O'Toole (in press, a,b).

A final common thread we wish to explore concerns the puzzle of human observers' difficulties in representing and processing inverted faces. Psychologists have been fascinated with the effects of face inversion since Yin's original paper in 1969. Perhaps the principal lesson that has been learned from these effects is that as amazingly accurate and flexible as our abilities with faces seem to be, the representation of faces we employ is not without limits. From a computational point of view, the surprise has been that the inversion transformation, technically-speaking, discards no information.

From a perceptual point of view, Rock (1974) has argued that the key to understanding inversion effects with faces is two-fold. First, faces are highly

[11] Even in light of the modulation transfer function of the eye (cf. Cornsweet, 1970), the intensity difference between the lowest and highest frequency information in faces is very large (unpublished observations).

complex and similar to one another, containing many small nuances that are important for distinguishing among similar faces. The second factor is the large differential in experience that human observers have with upright rather than inverted faces—faces are typically mono-oriented in space. The combination of these two factors is in many ways typical of the problems we have in acquiring the subtle features of some classical perceptual learning stimuli. Returning to our original example of distinguishing the music of two composers, the complexity of musical phrasing and interrelation of subcomponents in music creates a perceptual experience that is difficult to "subdivide" into simple parts. Likewise, a phoneme in language, while serving as a perceptual unit of sorts (perhaps not unlike the eyes, nose, and mouth of a face), is strongly affected both perceptually and computationally by contextual factors. It is for this reason that the phonemic unit has proved somewhat less useful than expected for quantifying speech streams. Both faces and these auditory stimuli are: 1.) highly complex; and 2.) mono-oriented with respect to a (some) physical dimension(s)—specifically, the x and y dimensions of space for faces and the time dimension for language and music. Inversion of a face might be considered similar to inverting time in a speech stream or musical composition; that is, playing a tape of a sentence or Mozart piece backwards. Clearly, all of the physical information remains present in this kind of a transformation, (e.g., Fourier spectrum). We are, nonetheless, completely unable to identify words or musical phrases with such a transformation.

The inversion of faces constitutes a much less extreme transformation than the inversion of time in an auditory stimulus due to the fact that faces can be inverted in space naturally, whereas music and speech cannot be inverted naturally in time (at least not at less than the speed of light! cf. Einstein, 1918). In fact, most of us have some limited experience with upside down faces and have much more experience with recognizing less complicated inverted objects (cups, chairs, etc.). Additionally, all of us have experience in observing objects undergoing spatial inversion (watching an object being turned upside down).

The face representation used in the perceptual learning model would be likely to perform very badly on inverted faces. Just how badly would depend primarily on the proportion of inverted to upright faces the model learns. From a perceptual learning point of view, this suggests that the problem we have recognizing inverted faces is similar to the problem we have recognizing other-race faces. The experience differential with upright versus inverted faces yields a representation optimal for coding upright faces in ways that make them optimally distinctive. Inverted faces, like other-race faces, should appear more similar to each other than upright faces—this may be related to the fact that a face and its "Thatcherization" appear very dissimilar upright, but much more similar inverted (Bartlett & Searcy, 1993).

Despite coding principles in the perceptual learning model that might predict some aspects of the effects

of inversion, we do not expect that it will provide a complete account of the various inversion phenomena. In particular, the model does not have general knowledge about the statistical structure of non-face objects, nor does it have access to general procedures that we seem to be able to call on successfully for general object recognition, (e.g., mental rotation, Shepard & Metzler, 1971). Understanding how these more general-purpose object recognition tools work may eventually be useful, perhaps even necessary, for understanding face inversion phenomena.

In summary, we believe that many problems in face processing can be understood, at least in part, at the level of the perceptual constraints on face processing. This indicates the importance in psychological and computational models of taking into account the kinds of perceptual problems posed by the statistical structure of faces as visual stimuli. The model we propose is far from answering many important questions about the kinds of information we derive from faces. It falls particularly short in giving insight into how we accomplish view transformations and in how the representation of an unfamiliar face changes with additional and more diverse experience with the face, over time and through motion, for instance. While the former may reasonably be attacked by enriching the quality of three-dimensional information available to the model, the later will likely require much more sophisticated modeling techniques than the ones we are currently employing.

References

[1] Abdi, H. (1994a). A primer of neural networks. *Journal of Biological Systems, 2,* 247–281.

[2] Abdi, H. (1994b). *Les réseaux de neurones.* Grenoble: Presses Universitaires de Grenoble.

[3] Abdi, H., Valentin, D. & O'Toole, A. J. (in press). More about the difference between men and women: Evidence from linear neural networks and principal components approach. *Perception* .

[4] Bartlett, J. C. & Searcy, J. (1993). Inversion and configuration of faces. *Cognitive Psychology, 25,* 281–316.

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

[6] Bruce, V. (1988). Recognizing faces. Hillsdale, NJ: Erlbaum. Bruce, V. & Young, A.W. (1986). Understanding face recognition. *British Journal of Psychology, 77,* 305-327.

[7] Brunelli, R. & Poggio, T. (1992). Hyperbf networks for gender classification. In *Proc. DARPA Image Understanding Workshop,* San Mateo: Morgan Kaufmann, pp. 311–314.

[8] Bruyer, R. (1986). *The neuropsychology of face perception and facial expression.* Hillsdale, N.J.: Erlbaum Associates.

[9] Burton, A. M., Bruce, V. & Dench, N. (1993). What's the difference between men and women? Evidence from facial measurement. *Perception, 22*, 153–176.

[10] Carey, S., & Diamond, R. (1977). From piecemeal to configurational representation of faces. *Science, 195*, 312–314.

[11] Cornsweet, T. (1970). *Visual perception.* New York: Academic Press.

[12] Cottrell, G.W. & Fleming, M.K. (1990) Face recognition using unsupervised feature extraction. *Proceedings of the International Neural Networks Conference.* Kluwer Dordrecht, pp. 322–335.

[13] Duda, R. O., & Hart, P.E. (1973). *Pattern classification.* New York: John Wiley & Sons.

[14] Einstein, A. (1918). *Mein weltbild.* Berlin : Ullstein Verlag.

[15] Ellis, H. D. (1975). Recognizing faces. *British Journal of Psychology, 66,* 409–426.

[16] Ellis, H. D. (1986). Processes underlying face recognition. In R. Bruyer (Ed.), *The neuropsychology of face perception and facial expression.* Hillsdale, NJ: Erlbaum.

[17] Ginsburg, A. (1978). *Visual information processing based on spatial filters constrained by biological data. PhD. Thesis,* University of Cambridge (Published as AFAMRL Technical Report TR-78-129).

[18] Golomb, B. A., Lawrence, D. T., & Sejnowski, T. J. (1991). sexnet: A neural network identifies sex from human faces. In R. Lippmann, J. Moody & D. S. Touretsky (Eds.). *Advances in Neural Information Processing Systems 3*, San Mateo, CA: Morgan Kaufmann.

[19] Harmon, L. D. (1973). The recognition of human faces. *Scientific American, 227,* 71–82.

[20] Hay, D. C. & Young, A. W. (1982). The human face. In A.W. Ellis (Ed.) *Normality and pathology in cognitive functions.* New York: Academic Press.

[21] Hotelling, H. (1933). Analysis of a complete set of variables into principal components. *Journal of Educational Psychology, 24,* 417–441.

[22] Kleiner, K. A. (1987). Amplitude and phase spectra as indices of infant's pattern preferences. *Infant Behaviour and Development, 10*, 49–59.

[23] Light, L. L., Kayra-Stuart, F., & Hollander, S. (1979). Recognition memory for typical and unusual faces. *Journal of Experimental Psychology: Human Learning and Memory, 5,* 212–228.

[24] McClelland, J. & Rumelhart, D. (1988). *Explorations in parallel distributed processing.* Cambridge, MA: MIT Press.

[25] Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology, 9,* 353–383.

[26] O'Toole, A. J., Abdi, H., Deffenbacher, K. A., & Bartlett, J. (1991). Classifying faces by race and sex using an autoassociative memory trained for recognition. *Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Lawrence Erlbaum, pp. 847–851.

[27] O'Toole, A. J., Abdi, H., Deffenbacher, K. A., & Valentin, D. (1993). A low dimensional representation of faces in the higher dimensions of the space. *Journal of the Optical Society of America A, 10,* 405–411.

[28] O'Toole, A. J., Deffenbacher, K. A., Abdi, H., & Bartlett, J. A. (1991). Simulating the "other-race effect" as a problem in perceptual learning. *Connection Science Journal of Neural Computing, Artificial Intelligence, and Cognitive Research, 3,* 163–178.

[29] O'Toole, A. J., Deffenbacher, K. A., Valentin, D., & Abdi, H. (1994). Structural aspects of face recognition and the other-race effect. *Memory and Cognition, 22,* 208–224.

[30] O'Toole, A. J., Millward, R. B., & Anderson, J. A. (1988). A physical system approach to recognition memory for spatially transformed faces. *Neural Networks, 1,* 179–199.

[31] Rhodes, G., Brake, S., and Atkinson, A.P. (1993). What's lost in inverted faces? *Cognition, 47,* 25–57.

[32] Rock, I. (1974). The perception of disoriented figures. *Scientific American, 230,* 78–85.

[33] Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review, 65,* 386–408.

[34] Rosch, E. & Mervis, C. B. (1975). Family resemblance studies in the internal structure of categories. *Cognitive Psychology, 7,* 573–605.

[35] Samal, A., & Iyengar, P.A. (1992). Automatic recognition and analysis of human faces and facial expression: A survey. *Pattern Recognition, 25,* 65–77.

[36] Schooler, J. W. & Engstler-Schooler, T. Y. (1990). Verbal overshadowing of visual memories: Some things are better left unsaid. *Cognitive Psychology, 22,* 36–71.

[37] Sergent, J. (1984). An investigation into the component and configural processes underlying face recognition. *British Journal of Psychology, 75,* 221–242.

[38] Sergent, J. (1986). Microgenesis of face perception. In H.D. Ellis, M.A. Jeeves, F. Newcombe, & A. Young (Eds.) *Aspects of face processing.* Dordrecht: Nijhoff.

[39] Sergent, J. (1989). Structural processing of faces. In A.W. Young, H. D. Ellis (Eds.) *Handbook of research in face processing.* Amsterdam: Elsevier.

[40] Shephard, R. M. & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science, 171,* 701–703.

[41] Shepherd, J. (1981). Social factors in face recognition. In G. Davies, H. Ellis, & J. Shepherd (Eds.) *Perceiving and remembering faces.* London: Academic Press.

[42] Sirovich, L. and Kirby, M. (1987). Low-dimensional procedure for the characterization of human face. *Journal of the Optical Society of America A, 4,* 519–518.

[43] Shapley, R., Caelli, T., Grossberg, S. Morgan, M. & Rentschler, I. (1990). Computational theories of visual perception. In L. Spillman, J. S. Werner (Eds.) *Visual perception: The neurophysiological foundations.* San Diego: Academic Press.

[44] Thompson, P. (1980). Margaret Thatcher: A new illusion. *Perception, 9,* 483–484.

[45] Turk, M. & Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience, 3,* 71–86.

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

[47] Valentin, D., Abdi, H., & O'Toole, A.J. (1994a). Categorization and identification of human face images by a neural network: A review of linear autoassociative and principal components approaches. *Journal of Biological Systems, 2*, 413–429.

[48] Valentin, D., Abdi, H., O'Toole, A. J., & Cottrell, G. (1994b). Connectionist models of face processing: A survey. *Pattern Recognition, 27*, 1209–1230.

[49] Vokey, J.R., & Read, J.D. (1992). Familiarity, memorability, and the effect of typicality on the recognition of faces. Memory and Cognition, 20, 291–302.

[50] Young, A., W., Hellawell, D., & Hay, D. C. (1987). Configural information in face perception. *Perception, 16,* 747–759.

[51] Young, A. W., McWeeny, K. H., Hay, D. C., & Ellis, A. W. (1986). Access to identity specific semantic codes from unfamiliar faces. *Quarterly Journal of Experimental Psychology, 38A,* 271–295.