Pergamon

PII: S0042-6989(97)00010-2

Vision Res., Vol. 37, No. 14, pp. 1961–1973, 1997

Printed in Great Britain
0042-6989/97 $17.00 + 0.00

Gender Discrimination and Prediction on the Basis of Facial Metric Information

JEAN-MARC FELLOUS*†

Received 9 February 1995; in revised form 29 May 1996

Horizontal and vertical facial measurements are statistically independent. Discriminant analysis shows that five of such normalized distances explain over 95% of the gender differences of "training" samples and predict the gender of 90% novel test faces exhibiting various facial expressions. The robustness of the method and its results are assessed. It is argued that these distances (termed fiducial) are compatible with those found experimentally by psychophysical and neurophysiological studies. In consequence, partial explanations for the effects observed in these experiments can be found in the intrinsic statistical nature of the facial stimuli used. © 1997 Elsevier Science Ltd.

Face Sex Discrimination Prediction

INTRODUCTION

Since the pioneering work of Darwin (Darwin, 1872), much scientific interest has been devoted to the understanding of information displayed by the face. It is striking to note how easy it is for a human (or a monkey) to quickly and reliably extract information such as gender (Bruce *et al.*, 1993), facial expression or identity from a single picture of a face. However, the face is not an "easy" visual object. Very subtle differences in shading, shape and depth profile can have tremendous influences on the perception of the face. One is, for example, able to discriminate between a "real" and a "fake" smile even though differences are minute (Ekman, 1992; Ekman & Davidson, 1993). It is of evident social value to be able to extract facial information, either dynamically, in the flow of social interactions, or statically, from pictures or brief visual exposures. Unfortunately, the precise nature of the facial traits that carry such information is still poorly understood. However, psychophysical experiments have indicated that, at least in the case of gender assessments, certain facial traits were more important than others, hence leading the way for a systematic and quantitative analysis of this information channel. We briefly review some of the most recent of such experiments below.

College students were requested to qualitatively assess the facial features of several male and female faces (Meerdink *et al.*, 1990). Twelve features, such as face width (narrow/wide), or nose size (small/large), were included. It was found that both male and female subjects had similar assessments of the male and female faces, but that those assessment patterns differed for male and female stimuli. Statistical analysis reveals that among metric features (i.e., not considering hair or eye colors for example), subjects were found to assess gender on the basis of face shape (face width and face length), mouth size, cheek position and eye size. In addition, judgments of male faces relied on eye spacing and a combination of nose size and eyebrow shape while female faces relied on nose size in isolation and the compound eye–eyebrow. Even though the metric properties used in this study were limited in number, and qualitatively assessed (small/large), the results clearly suggest that facial gender discrimination may be achieved on the basis of some "rules" shaped by experience and evolution, which in turn may be based on objective precise metric differences between male and female faces.

In a subsequent experiment, it was proposed that gender discrimination was achieved on the basis of certain facial feature arrangements, such as width of the chin or relative vertical distances between the eyes and the eyebrows (Brown & Perrett, 1993). This study was performed on the basis of average faces obtained from 16 female faces and 16 male faces. Pictures were carefully registered with respect to chosen reference points using appropriate deformations, and separate averages were computed for male and female faces. Stimuli were then created by sectioning the average faces into five horizontal bands containing respectively, eyebrows, eyes, nose, mouth, and lower jaw. These bands were then reassembled to form a new face, using some features from the male average, and others from the female average, the relative vertical distances between features being controlled. Subjects were then asked to identify the gender of the resulting faces. The results show that all

*Center for Neural Engineering, University of Southern California, Los Angeles, CA 90089-2520, U.S.A.

†Address all correspondence to present address: Brandeis University, Volen Center for Complex Systems, Waltham, MA 02254-9110, U.S.A. [*Email* fellous@volen.ccs.brandeis.edu].

features, except for the nose, carried some gender information, with possibly more weight given to the eye region and the chin, compatible with others showing that face identification mainly relies on internal features related to the eye–eyebrow complex and the mouth, but not the nose (Haig, 1986; Hosie *et al.*, 1988). This interesting study, however, presents some shortcomings. As in other studies, the processes of registration and normalization used in this study introduce a significant amount of distortion whose influence on the subjects' performances is difficult to evaluate. Similarly, the process of averaging might have artificially introduced (or removed) important texture features. Even though the averages "look" male or female, it is difficult to evaluate how distant the "maleness" or "femaleness" of these averages are from the samples from which they have been constructed. *A priori*, results from studies using averages or arbitrarily normalized faces rely on both distance and texture information whose respective contributions are again difficult to assess. In addition, the assembly of face parts was done manually and might have introduced additional perturbations to the geometry of the resulting stimuli faces. Finally, horizontal distances were not accounted for by this study.

In a separate study, Bruce *et al.* (1993) found that 13 subjects could achieve over 95% correct classification on a set of 180 pictures of male and female faces in which "easy features" such as hair style or earrings have been removed. In an attempt to explain these results, further experiments were conducted and showed that local information taken in isolation (such as eyes or nose) have few significant effects on the overall performance. However, in apparent contradiction with the Brown & Perrett (1993) study, masking the nose seemed to impair male recognition, while masking the eye region had a larger effect on female recognition. Even though individual features were masked locally, their perception, and hence their role in gender classification, might depend on the shape and positioning of other features such as hair, which was not present and whose contribution, in any case, is extremely difficult to assess.

Such studies are undoubtedly important to understand the nature of the facial information used to assess gender. However, they are inherently limited by the uncertainty about the underlying visual processes subserving face processing in general. It is not clear, for example, how much the gender assessment system depends on the face identity or face detection subsystems, should such subsystems be separable. Moreover, it is not clear whether a male and a female face differ in the same way that two male faces differ, for example. More quantitative studies might be useful to overcome such difficulties.

In an attempt to study the facial metric differences between males and females, Ferrario *et al.* (1993) extracted 22 facial points from 57 male and 51 female frontal faces controlled for size and position. Two hundred and thirty-one point-to-point distances were then extracted and their mean computed separately for male and female faces. The ratios of these mean distances were computed and subsequently analyzed. The results show that most of the salient metric differences involve points that are close to being vertically aligned. A large number of these distances involve the tip of the chin (pogonion) and are larger in men than in women. The next set of significant differences involves distances between horizontally aligned points, located exclusively on the right side of the face (larger in men than in women). Together, these results suggest that male faces are wider and longer than female faces. Unfortunately, as noted by the authors, significant conclusions could only be reached for overall face configurations (length, width), but not for particular distances in isolation. It is possible that because the method used to analyze the facial measurements (inspection of the mean ratios) relies on first-order statistics, it did not allow for individual dimensions to appear in the analysis. Could second-order statistics possibly have allowed for an account of their contributions? Other reasons for such a conclusion might be that although size and position were controlled, slight undesirable variations are always possible, and were not compensated for (i.e., normalized). Additionally, facial expressions were not reported to have been controlled, the subjects having been instructed to assume a natural and normal posture. From our own experience, female subjects have the tendency to smile (even slightly) when requested to take such a pose, possibly biasing the results and contributing to the appearance of the pogonion-related distances as gender discriminating. Point-to-point distances are more subject to noise than strictly vertical or horizontal distances, requiring a larger set of samples and more sophisticated statistical considerations. Finally, the problem of gender prediction was not addressed in this study.

The results of the independent and more detailed study of Burton *et al.* (1993) show that indeed such explanations might be reasonable. They extracted 73 points from a database of 91 male and 88 female neutral frontal pictures. Using discriminant analysis (second-order statistics), they found that 12 out of 18 point-to-point distances were able to account for 85% of the discrimination on the "training" set (set of pictures used to derive the discriminant function). The most important distances were in decreasing order, eyebrow thickness, width of the nose, width of the mouth, eye-to-eyebrow distance, forehead height, and distance between the inner corners of the eyebrows. In a second analysis of the data, 30 distance ratios and 30 angles were extracted, and their individual significance on an independent means t-test were calculated. Of these 60 measures, eight angles and five ratios reached criterion (significance >0.01) and were included in a new discriminant analysis. Finally, 5 of these 13 dimensions allowed for 73% discrimination. The four most significant variables were ratios. In decreasing order of importance these were the ratio of the mouth width to "angel brows"*; the ratio of the mouth width to

*"Angel brows" are measured by the end points of the nasal philtral ridge on the upper lip (Fig. 1).

the distance between the mouth and the nose; the ratio of the nose inflection* to the distance between the inner ends of the eyebrows; and the ratio of the mid-eye to nose base distance to the nose inflection. Other statistical studies were conducted by these authors on the basis of profile views, line drawings and 3-D information extracted from these same data, but they will not be discussed here as we wish to focus on gender classification of frontal pictures only. We note, however, that facial measurements taken from frontals contain implicit information from the 3-D profile of the face.

This very interesting study presents some shortcomings.

All distances were normalized with respect to the horizontal distance between the eyes. It is not clear whether such a normalization, originally chosen for its popularity, is actually valid for all distances, especially when they are considered point-to-point (i.e., not necessarily horizontal). In particular, normalizing vertical distances by a horizontal distance assumes that such distances are, in general, correlated, which does not appear obvious. Wide faces (on average) are probably not necessarily long.

As noted by the authors, the choices of the distances were largely arbitrary, and it is not known whether the results they obtained could not be improved should other distances be used. Moreover, for the chosen set of distances, the robustness of the findings was not assessed. In particular, some of the distances used to derive the discriminant functions were extremely small (mid-eye to upper base of the nose, "angel brows" and nose inflection, for example), and the influence of manual mispositioning (which according to our experience is unavoidable given the wide range of possible faces and picture quality) on the overall robustness of the results is not clear.

As pointed out by the authors, their study on frontal pictures did not yield "impressive" results. The classification rate was relatively low on the training set (85%) and could be expected to be even lower on a test set. The distances used to derive the discriminant functions were only partially related to psychophysical observations and the cases of misclassification did not correlate satisfactorily with human performance. Overall, they concluded that such a statistical analysis of the images might not be intrinsically adequate nor sufficient, and that the mathematics involved were possibly too simple. We will in the remainder of this document, suggest that there is nothing fundamentally wrong with such an approach, and that the metric features of visual stimuli can indeed account for a significant amount of the perceptual effects observed on human subjects.

If the body of psychophysical data on face perception is substantial, relatively little is known of its underlying neural mechanisms in general, or when related to gender assessment in particular. Recent neurophysiological work has pointed to the possibility for some "face selective" cells in the Superior Temporal Polysensory area (STP) and in the Inferotemporal cortex (IT) to code for certain metric properties of the face such as round vs elongated shape (Young & Yamane, 1992), or a combination of facial metric horizontal and vertical measurements (Yamane *et al.*, 1988). Unfortunately, the stimulus set did not include female faces, hence it did not allow for their assessments in gender discrimination.

The purpose of the present work is to investigate whether it is possible, on the sole basis of the statistics of a well chosen subset of some facial measurements to achieve efficient gender discrimination and prediction. Our study can be in many ways considered as an extension of the Burton *et al.* (1993) exploratory research which did not, unfortunately, yield satisfactory results. We will show here that by choosing a different set of measurements (suggested by neurophysiological experiments) the method used (and its results) is to a great extent adequate and successful by not only giving close to human performance in classification (over 95% compared to the 85% of the abovementioned study), but also by being robust with respect to the data, by holding on a test set (not just on a training set) and by being compatible with a significant amount of psychophysical and neurophysiological data, showing that indeed, the statistical structure of faces can explain and predict many aspects pertaining to their perception. Such an argument could be made for object perception in general, but it would be beyond the scope of this paper to document such a claim.

METHOD

A set of 109 pictures (256×256 pixels, 255 gray levels) were used in this study. They were subdivided into two groups.

The first set of pictures ("training set") consisted of 52 pictures acquired from 26 males and 26 females of both the ARPA/ARL FERET database and pictures taken in our laboratory. Together, these images consisted of 47 Caucasians and 5 Asians. All pictures were frontals (two ears visible) and subjects were requested to display a neutral facial expression.

Forty points were then individually extracted for each picture, and their pixel coordinates recorded (Fig. 1). These points were chosen to minimize the amount of error in their extraction. Twenty-nine of these points are identical to those chosen in other studies (Ferrario *et al.*, 1993; Cunningham, 1986) that have shown that the error of extraction of their location due to the use of different operators or different extraction sessions was negligible. The remaining 11 points (part of the Burton *et al.* (1993) study) presented the same properties.† For this study, all points were extracted by the same operator using an original computer program written by the author. We will call these facial loci "fiducial points" hereafter.

*The inflexion of the nose is the vertical distance between the upper base of the nose and the most prominent point of the nose bone (see Fig. 1).

†Assessments of the reliability of extraction of these 40 points across operators and across sessions was assessed in our laboratory on 15 subjects and 10 pictures (five males and five females), and resulted in less than 3 pixel errors on average (data not shown).

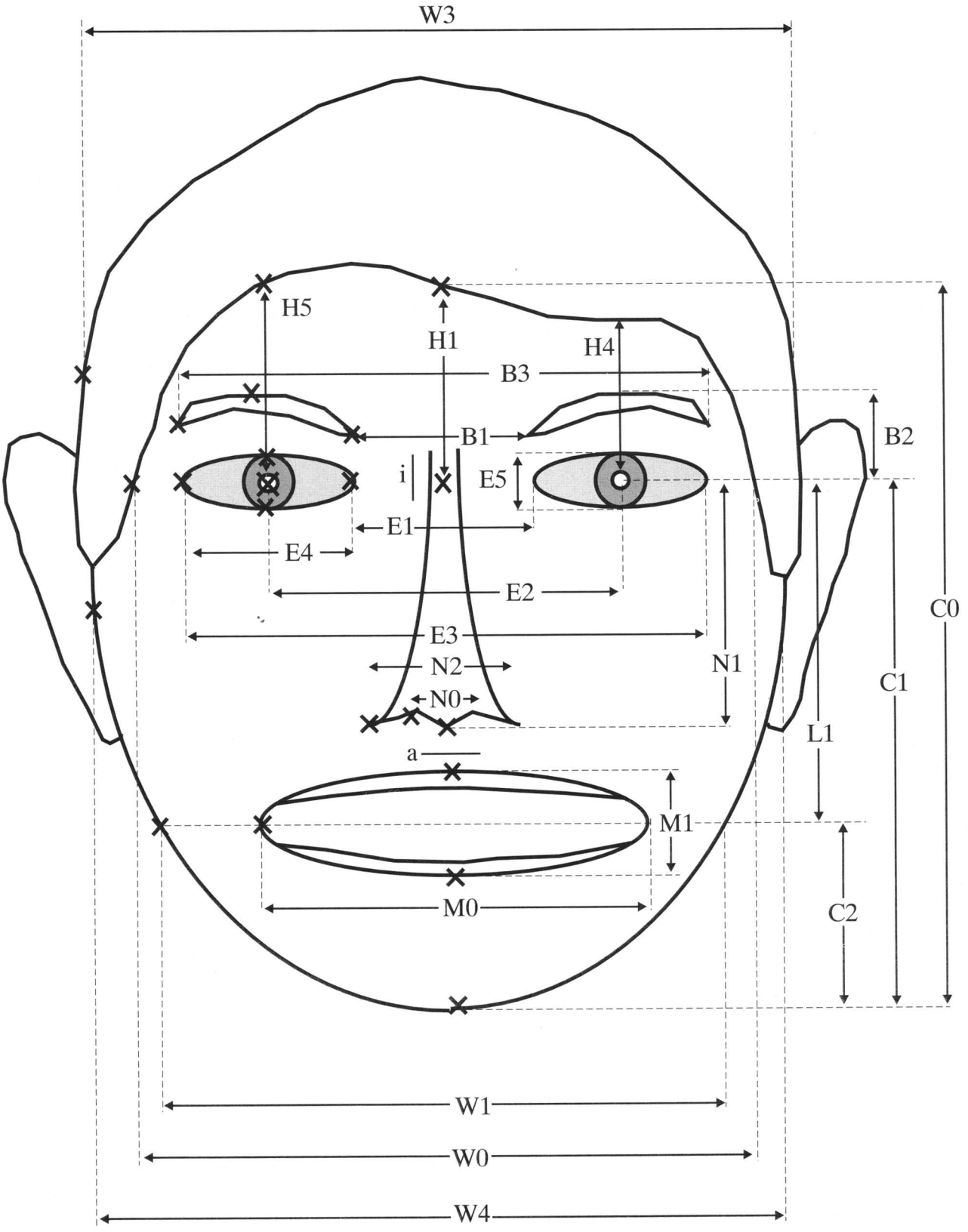

FIGURE 1. Fiducial points (only right side represented) and fiducial distances. Notations compatible with Young & Yamane (1992). "i" indicates the nose inflection, "a" indicates the "angel brow".

One the basis of these points, 24 horizontal and vertical distances (fiducial distances) were then calculated (Fig. 1). These distances were chosen on the basis of the Young & Yamane (1992) and Yamane *et al.* (1988) studies which suggested their possible coding by temporal lobe face cells. The relative independence of horizontal and vertical distances was then tested (see Results).

Horizontal distances were normalized with respect to the interpupillary distance, whereas vertical distances

were normalized with respect to the distance of the eye-midpoint to the philtral ridges midpoint (see below and Fig. 1). The 22 remaining normalized distances were then subjected to discriminant analysis. Five of the original 22 dimensions were found to satisfactorily account for gender discrimination and were kept for further analysis. On their basis, discriminant functions were derived, and gender discrimination was assessed on the training set.

A second set of 57 frontal pictures (26 female, 31 male faces) exhibiting various facial expressions was then submitted to the same fiducial distances extraction and was used as a test set to assess gender prediction.

RESULTS

We first will compute and study the correlation patterns between all the 24 dimensions extracted. We will next perform a discriminant analysis. After normalizing our measurements, we will use the correlation matrix to find the dimensions which contribute the most to gender discrimination. Using these dimensions, we will derive the discriminant functions, assess their performance on the training set, a noisy set of training data and a novel test set of pictures.

Independence of horizontal and vertical distances: normalization

Researchers working on issues related to face analysis are confronted with the problem of eliminating possible bias introduced by the data collection (picture taking) procedure. In our case, since we are solely interested in metric information, we have to control for possible metric variation across faces due to the overall size of the face within the picture (i.e., distance from the subject to the camera).

In order to compensate for such variations we start this study by looking for appropriate ways to normalize the raw fiducial distances extracted from the pictures. Most studies chose to align arbitrarily chosen facial points (such as the eyes) with each other, across the data set. However, most of these techniques do not account for both vertical and horizontal dimensions and assume that horizontal normalizations, for example, are sufficient for the purpose of their subsequent processing. There is some evidence that horizontal and vertical deformations of fiducial distances (on familiar faces) are perceived independently by subjects, suggesting that both vertical and horizontal assessments are perceptually somewhat independent (Hosie *et al.*, 1988). Other studies on facial metrics found that most differences between male and female faces could be illustrated by distances which are close to being vertical or horizontal (Ferrario *et al.*, 1993).

As can be noted from the correlation matrix of our 24 fiducial dimensions (Fig. 2), hair-related vertical measurements (H1, H4, H5, C0) are strongly correlated with each other, and with the exception of C0, uncorrelated with any other dimension (below 0.5 threshold). As expected, E5 is uncorrelated with all the other variables, since it measures the extent to which the eyes are open. It

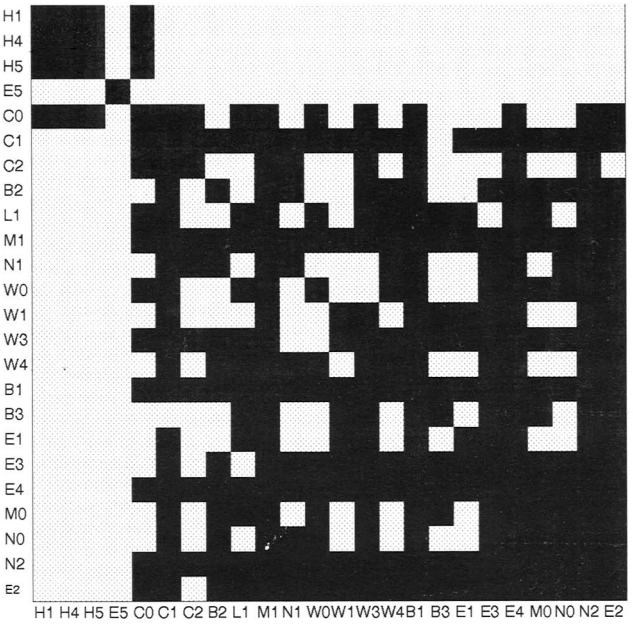

FIGURE 2. Fiducial distances correlation matrix. Black represents values greater than 0.5 (high correlation), gray represents values below 0.5. Abbreviations as in Fig 1. H1, H4, H5 and C0 are strongly correlated with each other. E5 is uncorrelated with all the other dimensions. The average correlation of horizontal dimensions and vertical dimensions is 0.65, the average correlation between horizontal and vertical dimensions is 0.55.

is interesting to note that the three groups of horizontal measurements related respectively to the eyes (E1, E2, E3), the nose (N0, N2) and the mouth (M0) are strongly correlated with each other.

A numerical analysis shows that the average correlation of horizontal distances with each other (average of the sub-matrix delimited by W0 and E2) and the average correlation of horizontal distances with each other (average of the sub-matrix delimited by C0 and N1) are both 0.65, whereas the average of the correlation between horizontal and vertical distances is lower (0.55).

To summarize, this analysis indicates that hair-related dimensions together with E5 constitute a group of measurements which can be regarded as independent from the others, and that horizontal and vertical dimensions correlate, on average, slightly more with each other than between each other.

As another way of studying the relationship between these dimensions we chose to investigate their correlation with their principal components. In the 24 dimensional space constituted by the fiducial distances, each sample face is represented by a data point. We performed Principle Component Analysis (PCA) on these data points in this space to find the axes of maximal variance. These axes are given by the eigenvectors of the correlation matrix computed on the basis of the 24 fiducial dimensions and 52 samples. Each corresponding eigenvalue (because the eigenvectors are normalized) indicates the amount of variance accounted for by its associated eigenvector.

The first four axes were found to explain 75% of the total variance. According to both the Kaiser criterion and

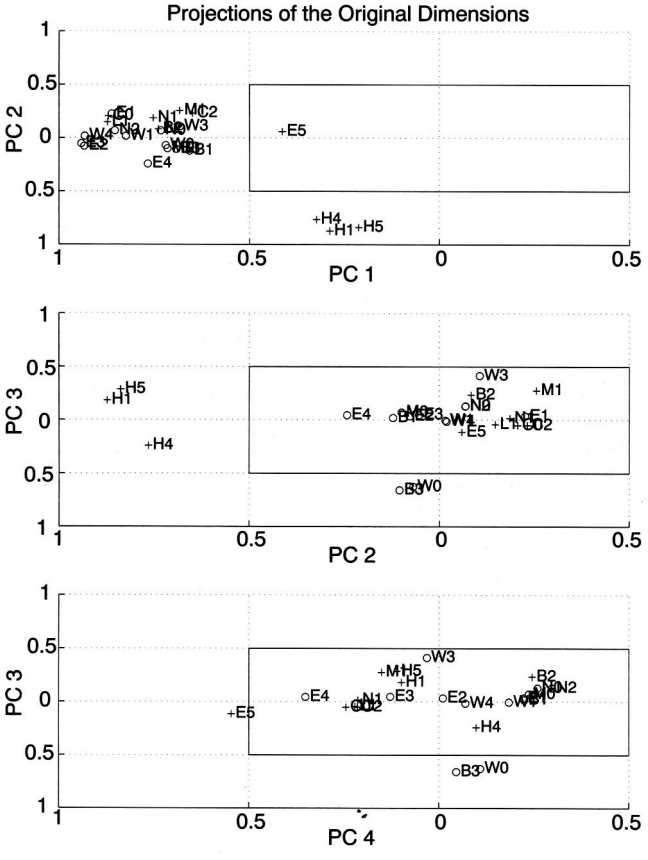

FIGURE 3. Projections of the original 24 fiducial dimensions onto their four first principal components. Circles (○) indicate horizontal dimensions, crosses (+) indicate vertical dimensions. Horizontal dimensions are best described by the first principal component, hair-related dimensions by the second, and B3 and W0 by the third. The fourth component can be used to separate most of the horizontal and vertical dimensions.

a Scree test performed on the eigenvalues, these four components can be chosen as an adequate subspace in which the original data could be represented without losing their overall structure. In general, it is not easy to interpret the principal components. However, it is possible to qualitatively assess the relationship between the original dimensions (fiducial distances) and the principal components (axis of maximal variances) by projecting the former onto the latter. The original 24 dimensions were therefore projected onto the first four principal components. If an original fiducial dimension is strongly correlated with a given principal component, its projection should be non-zero (in absolute value greater than an a priori chosen threshold), while its projection onto the other components should be negligible (in absolute value smaller than an a priori chosen threshold). We chose 0.5 to be the threshold value to make such a decision.

As can be seen from Fig. 3, all of the horizontal dimensions (indicated by circles) were best correlated with the first principal component, whereas all the vertical dimensions related to the hair were correlated with the second component. This result is compatible with our observations of the correlation matrix. Since the eigenvectors form an orthonormal basis, we also conclude that hair-related dimensions account for a portion of the variance which is distinct from the one accounted for by the other dimensions. Furthermore, we note from Fig. 3 that two horizontal dimensions (B3 and W0) appear to correlate with the third principal component. Again, because the eigenvectors are orthogonal to each other, these results suggest that these two dimensions account for a separate portion of the overall variance of the data set. Finally, projections onto the fourth principal component indicate that the latter can be used to separate horizontal from vertical dimensions (projections are essentially negative for vertical distances and positive for horizontal distances) (Fig. 3).

Altogether, these results show that the correlation between all variables and the variance of the data set are explained by groups of dimensions which either belong to vertical or horizontal measurement, but not both, suggesting that these two subclasses of dimensions can be regarded as independent to a first degree of approximation. Consequently, we chose to normalize vertical and horizontal distances separately.

All the faces being frontals, and in an attempt to be compatible with the Burton *et al.* (1993) study, we chose the interpupillary distance (E2) as a normalizing factor of horizontal distances. We chose the distance between the eye midpoint and the philtral ridges midpoint (N1) as the normalizing factor of vertical distances (Fig. 1). We note that our choices are different from the ones of Cunningham (1986) who also chose to normalize separately vertical and horizontal distances. In his study, vertical normalization was achieved on the basis of face length, defined as the distance between the hairline and the chin. This distance is not strictly facial in the sense that it introduces the undesirable (and to a large extent arbitrary) influence of hairstyle and the extent of facial expressions (all smiles) which were not controlled for. Horizontal distances were normalized with respect to several distances: the upper part of the face was normalized with respect to cheek bones' distances (their distance 2, our W4) whereas lower parts of the face were normalized with respect to the width of the face at the mouth level (their distance 3, our W1). In contrast, we chose to normalize all horizontal distances with respect to a single measurement which could be reliably extracted (E2).

In addition, to compensate for possible slight planar tilts and slight depth rotations of the head, we used an average of analogous fiducial points and distances whenever possible. For example, the eye mid-point was computed as the center of mass of the three points: left and right pupils and manually extracted eye mid-point.

Model construction

A qualitative inspection of the projection of the original data (face points) onto the four first principal components shows that the latter do not allow for a discrimination between male and female faces (Fig. 4). A similar observation could also be made by considering

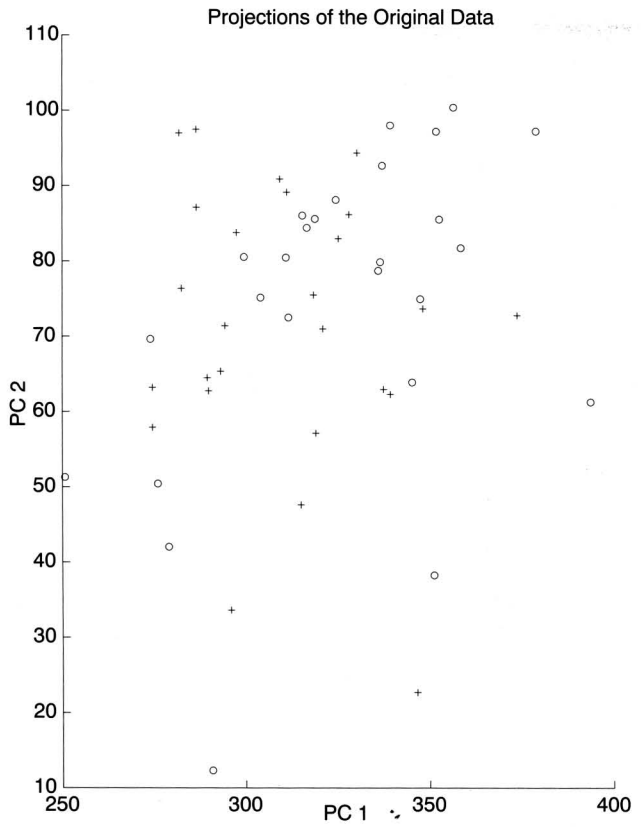

FIGURE 4. Projections of the data on their first two principal components showing that it is not possible to separate male (indicated by +) from female (indicated by ○) data points. Projections on the two other components yielded similar plots (not shown).

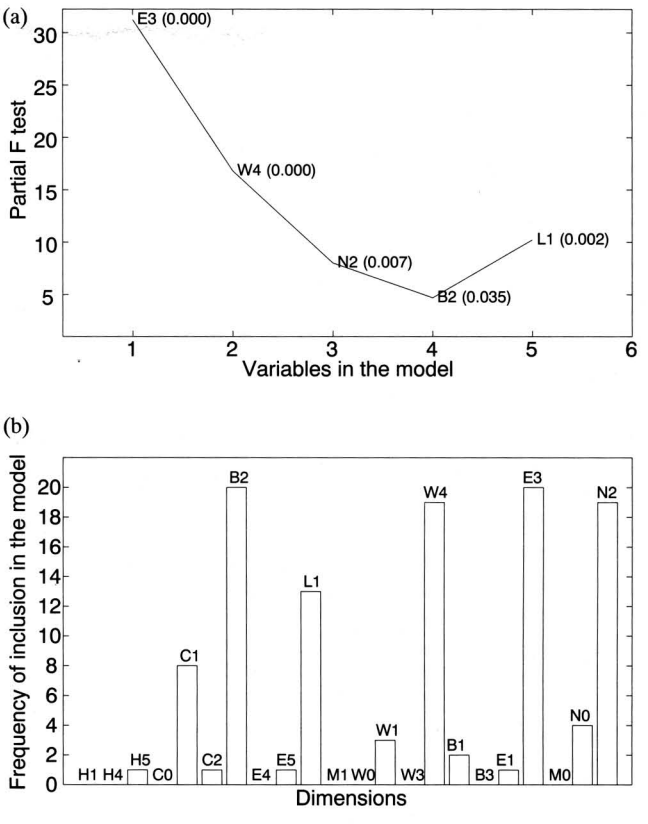

FIGURE 5. (a) Summary of the results of the iterative procedure used to extract normalized dimensions that account for most of the discrimination between male and female fiducial distances. The confidence level (P) is indicated in parentheses. (b) Histogram of the frequency of occurrence in the model of the normalized dimensions extracted on the basis of noisy data. This histogram has been derived from 20 model constructions (and noise injection).

each original dimension individually. This result shows that the axis of maximal variance of the original 24 dimensional space is unable to account for gender differences. In other words, no principal components describe a linear combination of the original 24 dimensions allowing for gender discrimination.

In the space of 22 normalized dimensions (two dimensions are used for normalizing), faces are again represented by a single point. Owing to the inherent structure of the face, it is probable that faces occupy a relatively small portion of this space. Moreover, it is possible that male and female faces, should they bear any metric differences, are located in distinguishable regions of this space. If so, only a few of the original 22 dimensions are necessary to allow for the discrimination between male and female faces in this space.

We next studied whether a subspace of the original dimensional space could be found that allows for such a discrimination. The purpose of the second part of this study was to determine which of the original dimensions allowed for gender discrimination (according to a linear separation criterion) and their individual relative contribution to the overall discrimination. The procedure described below is analogous to the procedure termed "stepwise inclusion of variables" in the discriminant analysis literature, which was also used by the Burton *et al.* (1993) study.

We separated the faces of the training set into two groups according to their gender. We then computed the F statistics for each of the original normalized dimensions (univariate F), a measure of the ratio between the total and average within-group variances accounted for by this dimension alone. The larger the value of F, the more significant is the difference between the means of the two groups, the greater is the dimension accounting for the discrimination. The dimension with the largest F value was then included in the model. The procedure was then repeated by computing the partial F values of the remaining dimensions until the F value became smaller than a chosen threshold [1.5 in our case, which is slightly more stringent than 1 used in the Burton *et al.* (1993) study]. In accordance with the terminology of discriminant analysis, we called the variables of largest F value, *model-variables*. Those variables were the ones chosen for further analysis of the data. All other variables were ignored. Figure 5(a) illustrates the results of this procedure.

The graph shows that the procedure stops after the inclusion of five variables: E3 (distance between the outer most corner of the eyes), W4 (distance between the two cheek bones), N2 (width of the nose), B2 (distance between the eyes and the eyebrows) and L1 (distance

between the eyes and the mouth). All F values were significant ($P < 0.04$). We noted that "easy features" such as distances related to the hair (H1, H4, H5, W3, C0 and to a lesser extent W0), which a priori could be of help for gender discrimination, did not appear in this analysis. This indicated that the original database included a wide variation of hairstyles for males as well as females.

We next assessed the robustness of this finding by introducing noise in the original data. Noise of a chosen amplitude (measured in pixels on the two classes of dimensions) was added to the original measurements, and the procedure mentioned above was repeated several times. The frequency of occurrence of each variable in the model was then recorded. Figure 5(b) shows the result of 20 iterations of the procedure for noise amplitudes of 3.0 pixels on the vertical dimensions and 2.5 pixels on the horizontal dimensions. The graph shows that the dimensions which occur most often in the model are still B2, L1 W4, E3 and N2, suggesting their robustness with respect to possible errors in the initial extraction of the fiducial points. We also noted that W1 and C1 appeared as potential candidates for inclusion in the model, even though their contributions seemed weaker. The occurrence of N0 might be a side effect of N2's.

The set of distances mentioned above (E3, W4, N2, B2 and L1) constitute a model (or subspace) on which further data analysis will be performed. The dimensions will be referred to as model dimensions.

Canonical correlation analysis

Our previous study showed that five of the original dimensions might be sufficient to discriminate between the two groups of data (male and female measurements). We now wished to determine whether a single linear combination of these variables (projection in a one-dimensional subspace) could account for the group differences, and if such a linear combination existed, what the relative contributions of the original five variables to the discrimination would be. In other words, we wanted to determine the vector of coefficient v which maximized the ratio:

$$\frac{v'Bv}{v'Vv}$$

where B is the between group sum of squares matrix, V is the total sum of squares matrix, both rescaled according to the individual standard deviations of each dimension (we used correlation rather than covariance). The prime operation corresponds to transposition. The two groups considered here are male and female measurements. Results from canonical correlation analysis show that v is non-unique, and corresponds to the generalized eigenvectors satisfying:

$$(B - \lambda V)v = 0$$

Our analysis showed that the largest eigenvalue was several orders of magnitude higher than the second, therefore confirming that it is indeed possible to represent the data (and discriminate between the two groups) in a one-dimensional space constituted by the first eigenvector. Figure 6 shows the projection of the data onto the subspace constituted by the first two eigenvectors. Unlike their projections in the principal component space (Fig. 4), a clear discrimination could be made between male and female faces. Figure 6 also shows the relative contribution of the model variables to gender discrimination. For example, the larger the values E3 and B2 (in this order), the more likely is the person to be a female, whereas the larger the values N2, L1 and W4 (in this order), the more likely is the face to be male. A linear function of the model variables can be computed which is such that, given the five measurements of an unknown face, positive values will indicate a high likelihood for the person to be of one chosen gender, whereas negative values will characterize the other gender. Figure 7 illustrates this point on the 52 original training samples, two of which are misclassified. Figure 7 shows the same data with the addition of noise, simulating 5×52 samples, as previously described.

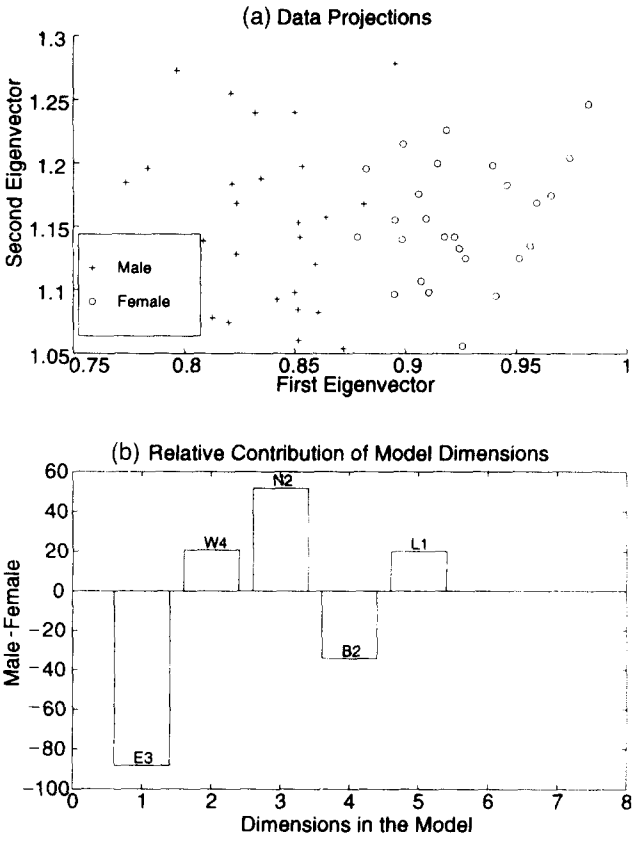

FIGURE 6. (a) Projections of the training data onto the first two eigenvectors, indicating that the first eigenvector alone can be used for gender discrimination. (b) Relative contribution of the model dimensions to the discrimination between male and female faces. See text for details.

Generalization: gender prediction

To conclude this study, we assessed the potential of this discriminant function to predict the gender of new pictures, coded by their fiducial distances.

We constituted a test set of 26 female and 31 male frontal faces (11 of these pictures were from persons who

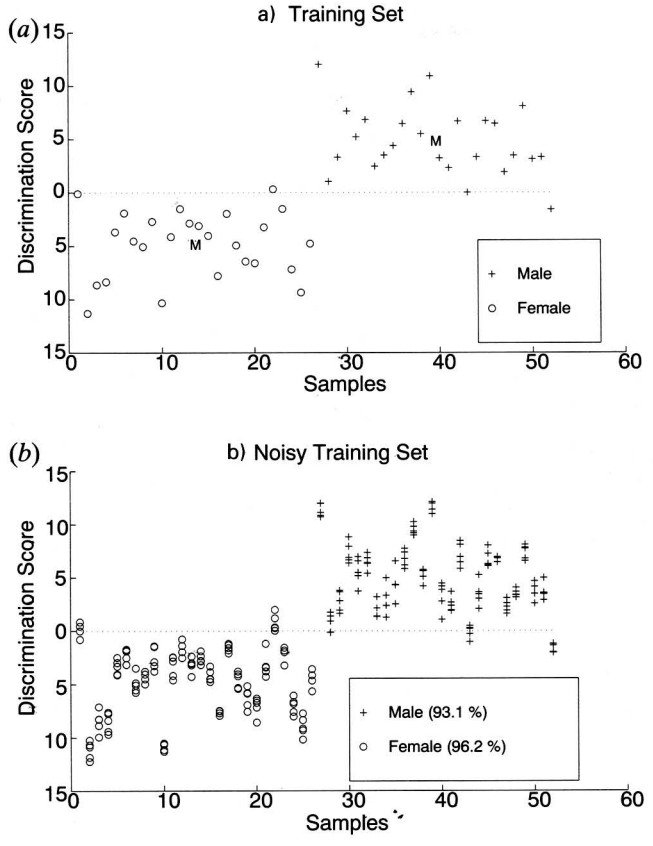

FIGURE 7. Plot of the discrimination scores for the training data set. Positive values should be males, negative values should be females. Mean values of the measurements are indicated by the letter "M" and are plotted for both male and female. (a) Training set. (b) Training set with added noise. Five noise injections are represented. 93.1% of the discrimination is correct for males, 96.2% of the discrimination is correct for females.

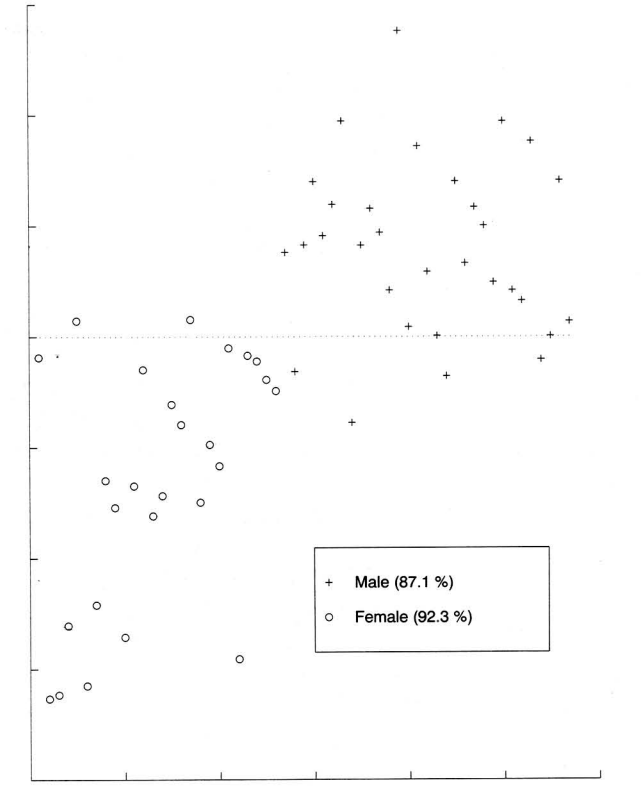

FIGURE 8. Plot of the discrimination scores for 57 novel test pictures. The model achieves 87.1% discrimination for males and 92.3% for females.

had a different picture in the training set). Together these images included 54 Caucasians and three Asians. Subjects were asked to freely exhibit some of the following facial expressions: neutral, angry, surprise and smile. Smiling faces were a posteriori subdivided into two groups according to whether teeth were apparent (large smile) or not (small smile) (Table 1); neutral faces were taken from people who had no pictures in the training set.

Figure 8 shows the result of the classification, using the most stringent criterion (either male or female). The classification yielded 87.1% correct for males, and 92.3% correct for female. A consideration of the mean values of the model dimensions and their standard deviations showed that the worst misclassified male case was due to a large E3 (Asian, small smile) while the worst misclassified female case was due to a small B2 (Caucasian, angry).

TABLE 1. Composition of the test set

	Female	Male
Neutral	4	6
Angry	2	5
Surprise	2	4
Small smile	6	7
Large smile	12	9
Total	26	31

DISCUSSION

Summary

Our results show that considering either all 24 fiducial dimensions together or in isolation does not allow for gender discrimination (Fig. 4). However, projecting the data onto a subspace of five (independent) dimensions (Fig. 6) allows for the derivation of a single dimension onto which projections of the data yield acceptable results for gender discrimination and prediction (Figs 7 and 8). Facial dimensions are either vertical or horizontal and are normalized independently.

Discriminant analysis (Figs 5 and 6) suggests that "femaleness" relies on large distances between external eye corners (E3), a measure of overall eye extent, large distance between the eyes and the eyebrows (B2), a small nose (N2), a narrow (small W4) and roundish (small L1) face (see Fig. 1). "Maleness" relies on the presence of a large nostril-to-nostril width (N2), wide cheek bones (W4), lengthy face (L1), small extent of the eyes (E3) and small distances between the eyebrows and the eyes (B2). C1 and W1 are to a lesser extent involved in gender discrimination (Fig. 5), but were not included in our model. On the basis of these five dimensions it was

possible to achieve over 95% gender discrimination on the training set (Fig. 7) and about 90% discrimination on a test set including various facial expressions (Fig. 8). These performances are, to date, and to our knowledge, the highest amongst gender discrimination studies solely based on 2-D facial metric information.

Other statistical studies of facial measurements

Our study assesses to what extent it is possible to extract gender information from frontal pictures of faces on the sole basis of facial metric information. It refines the pioneering studies of Ferrario *et al.* (1993) and Burton *et al.* (1993), and in addition, shows that such extraction also allows for the prediction of the gender of a newly presented face.

The methods we have used somewhat differ from these two studies. Firstly, partly on the basis of the Ferrario *et al.* (1993) results, and on neurophysiological data on "face cells" (Young & Yamane, 1992; Yamane *et al.*, 1988), we have chosen to consider horizontal and vertical distances (Fig. 1), rather than point-to-point distances as in the Burton *et al.* (1993) and Ferrario *et al.* (1993) studies. Indeed, considering 234 arbitrary point-to-point mean facial distances shows that most of the gender discriminating distances were either vertical or horizontal (Ferrario *et al.*, 1993). Moreover, face-selective cells seem to correlate their firing rate with linear combinations of horizontal or vertical facial distances, possibly indicating that they are important for the coding of a face in general. Our choice of such distances is further justified by our finding that the particular horizontal and vertical distances we have chosen can be considered to a large extent to be independent of one another.

Secondly, because our distances were measured along fixed directions (vertical and horizontal), and because of their relative independence, we normalized them independently. The Ferrario *et al.* (1993) study did not normalize their data (but computed ratios rather than raw distances) and Burton *et al.* (1993) normalized all point-to-point distances with respect to a single (horizontal) one, therefore implicitly assuming that such a distance is correlated with all others in the same manner, which is not necessarily true.

Thirdly, unlike the Ferrario *et al.* (1993) study, but like the Burton *et al.* (1993) study, we used second-order, rather than first-order statistics, therefore accounting for the variance of the measurements, rather than simply their average.

The nature of the Ferrario *et al.* (1993) data and their analysis did not allow for definite conclusions regarding individual distances, but gave meaningful information about the general facial shape differences between male (elongated) and female (squared) faces, which our study also showed. Moreover, using different techniques, and only male faces, others have observed that most of the male face variability was captured by dimensions related to the amount of hair and overall shape of the face (elongated vs round/squarish) (Young & Yamane, 1992), indicating that facial shape might also account for intra-gender discriminations. In addition, Ferrario *et al.* (1993) pointed to the lower third (pogonion related) of the face as a possible major locus for gender differences. Our study did not, however, find that the mouth region bore any more statistical differences than the upper parts of the face. We would like to suggest that their results might be due the fact that the stimuli were uncontrolled for facial expressions (which were "normal") which primarily occur in the lower third of the face, possibly biasing the results toward smiling faces, which, from our experience, are more likely to occur for female subjects requested to present a normal facial expression. Our results can be considered as a subset of the 42 gender discriminating point-to-point measurements derived by Ferrario *et al.* (1993). For example, our eyebrow-related distances are to be compared to their 7–10 and 11–12 distances, measured between the supraorbital foramen and the eye medial canthus, and which were larger in women than in men. These distances are the most similar (almost vertical) to our distance B2.* We found, however, that a much smaller number (5 compared to 42) of such distances are sufficient to achieve acceptable gender discrimination.

Using a very different technique, Brunelli & Poggio (1993a) achieved the same type of performance for gender discrimination and prediction. This study extracted 18 point-to-point facial dimensions, of which eight were also part of our study (even though they were not normalized) from 21 male and 21 female pictures, controlled for pose, size and expression. They show that gender discrimination can be achieved on the basis of eyebrow thickness (not included in our study), upper lips thickness (not included in our study), pupil to eyebrow separation (B2, a model dimension) and lower lips thickness (not included in our study), in descending order of importance. They also show that distances analogous to W4, W1, L1, N2 and C1 do not significantly contribute to the discrimination. E3 was not measured.

If it is very difficult to assess lip thickness in general (how is it defined?); it is equally difficult to control in subjects. A slight lip pressure might not appear significant to the naked eye (as far as facial expression is concerned, for example), but might change the value of the measurement significantly. Moreover, it is clear that gender discrimination is achieved by humans irrespective of the facial expression displayed, which is likely to introduce severe perturbations of lip thickness measurements. Unfortunately, gender determination was only assessed on the basis of neutral pictures. Furthermore, their analysis relies on 21 pictures of each gender, 20 of which are used as a training set, one as test (in a randomized fashion). It is not clear to us how this procedure might rely on possible measurement biases (lip related, for example) which were achieved automatically, since no significance measurements were provided for

*It is interesting to note that low (small B2) and heavy eyebrows have long been taught to be a prominent feature of male faces when drawn or painted [(Laidman, 1979) cited in Nakdimen (1984)].

either the feature extraction or the gender estimation procedures. It is, therefore, possible that discarding lip information might have significantly changed the nature of their results, and introduced such dimensions as W4, N2 or L1. Lip thickness (both upper and lower lips) also appears to contribute significantly to gender discrimination in the Burton *et al.* (1993) study. Similarly, eyebrow thickness was also reported to be a significant measure of gender differences in both studies (Burton *et al.*, 1993; Brunelli & Poggio, 1993a). It was not reported to what extent the database was controlled for females having plucked their eyebrows, which would have naturally biased the results. Moreover, the precise criteria used to determine the "beginning" and the "end" of the eyebrows or lips were not indicated, making assessment of the reliability of such small measurements difficult. Finally, as for lip thickness, eyebrow thickness (depending on how it is measured) might vary across facial expressions, possibly making gender estimation dependent on facial expression. In contrast, our study relied on "manually" extracted fiducial features and included a large test set, comprising novel images of more than 40 people who were not members of the training set and who exhibited various facial expressions.

Unlike both our study and that by Burton *et al.* (1993), chin information in the Brunelli & Poggio (1993a) study did not appear to significantly influence gender discrimination, in comparison to the above-mentioned dimensions (lips thickness, eyebrow thickness...). However, their "typical" male and female faces (their Fig. 4) differ notably with respect to the chin. The male face presents a large and flat chin, while the female face presents an elongated chin, compatible with our observations related to W4 and W1. This apparent discrepancy was not discussed. We also note that hair information was not included in either Burton *et al.* (1993) or Brunelli & Poggio (1993a) studies, while it was in ours, showing no significance, and therefore validating their choice.

It is interesting to note that N2 (analogous to their nose width at base), W4 (face width at cheek) and B2 (eye to eyebrow distance, normalized with respect to the interpupillary distance) were also derived by the Burton *et al.* (1993) exploratory study. It is our prediction that L1 and E3, had they been included in their analysis, would have also probably appeared as significant measurements, possibly overshadowing the other nine discriminative measurements they derived, and improving the overall classification performance.

We should finally mention that other statistical methods and data have been used to perform gender discrimination. Most of these methods are based on the analysis of pixel-based templates rather than feature-like distances (Brunelli & Poggio, 1993b) and are discussed in Burton *et al.* (1993). We would like to suggest that approaches such as the one presented in this study are suitable for improving general computer-based automatic face recognition methods (Samal & Iyengar, 1992; Valentin *et al.*, 1994; Wiskott *et al.*, 1995) by allowing for an *a priori* classification of pictures into several subclasses such as male and female. In such subclasses, better recognition performance can be achieved (by preventing male faces being compared to females faces, for example).

Psychophysical experiments on the influence of facial features on gender discrimination

Even though hair-related measurements were included as possible candidates for gender discrimination, they did not appear to be statistically significant (Fig. 5), nor did they correlate with any other strictly facial dimensions (Fig. 2). Our study, therefore, does not rely on "easy" features such as typically feminine hairstyles, for example, showing that some strictly facial features are sufficient to achieve gender discrimination, in possible agreement with studies performed on infants. Our results are also compatible and complementary to others derived from adult assessments of gender (Brown & Perrett, 1993; Bruce *et al.*, 1993; Meerdink *et al.*, 1990; Roberts & Bruce, 1988; Hosie *et al.*, 1988; Haig, 1986), using different techniques.

On the basis of average faces, Brown & Perrett (1993) found that the vertical positioning of the eye–eyebrow complex, the eyebrows alone (related to B2), the whole jaw (related to L1) and the chin (related to C1) accounted for most of the gender discrimination assessed by college students, on the basis of exposures to these features in isolation or in relation to one another. Unfortunately, it is not clear from this study what the respective contributions of facial metric information, texture information and averaging effects are. Our results, however, suggest that such regions (B2, L1 and to a lesser extent C1) appear to play a significant role in gender discrimination, should only metric information be used. Moreover, in our study, assessment of gender was achieved on the basis of individual samples and not on averages. Consequently, our method for gender discrimination and prediction does not rely on averaging effects, which might have been the case in the Brown & Perrett (1993) psychophysical study. Horizontal dimensions were found to be important for face recognition (Hosie *et al.*, 1988), and there are few reasons to believe that they do not play a role in gender discrimination. Unfortunately, Brown & Perrett (1993) did not account for such dimensions. Finally, our results predict that repeating this study with varying horizontal dimensions might reveal the contributions of N2, W4, E3 and possibly W1 to gender discrimination by subjects.

Bruce *et al.* (1993) and Roberts & Bruce (1988) have shown, for their part, that gender discrimination was to a certain extent sensitive to the masking of the eye and nose regions. They noted that masking the nose region had a greater effect on male assessments, while masking the eyes had a greater effect on female assessments. These results are compatible with ours. The nose region is captured by L1 and N2 which are larger in male faces. The eye region is captured by E3 and B2, which are larger in female faces. Moreover, our study predicts that masking the lateral edges at the level of the cheeks (captured by W4) would have an effect on gender

assessment, and that this effect would be slightly more prominent for males than for females (see Fig. 6).

Relying on qualitative assessment of facial features, Meerdink *et al.* (1990) found that female faces were discriminated on the basis of metric features such as the size of the nose (Enlow, 1982) (of which N2 is a measure) and the compound eye–eyebrow (of which B2 and E3 are measures). Male faces were characterized by judgment related to eye spacing (E1) and a combination of nose size and eyebrow shape (of which B2 is an indication). Our results support and augment their conclusions. We were indeed able to refine their results by quantitatively assessing the relative importance of the various metric features, and characterizing their individual influence on gender discrimination from a purely statistical perspective.

However, if all the model dimensions derived in our study appear to have significantly influenced subjects in their discrimination decision, others do not stem from this statistical analysis. Mouth size, for example, has been found to be used for gender discrimination (Meerdink *et al.*, 1990; Brown & Perrett, 1993), but does not appear in our analysis. We would like to suggest that mouth size effects may be due to facial expression displays which were not entirely controlled (Meerdink *et al.*, 1990), or possibly due to texture and averaging effects (Brown & Perrett, 1993) which were intentionally not considered in our study. Similarly, the nose region has not been found to play a significant role in gender discrimination (Brown & Perrett, 1993) whereas it was suggested to be a factor by our results and others (Nakdimen, 1984; Roberts & Bruce, 1988; Bruce *et al.*, 1993; Enlow, 1982; Burton *et al.*, 1993). We would like to suggest that it is because the Brown & Perrett (1993) study did not account for horizontal dimensions that N2 did not appear significant. We also agree with the Roberts & Bruce (1988) and Bruce *et al.* (1993) conclusions that it is the relationship between the nose and other facial features, rather than the nose alone, which carries information about the gender of a face. Indeed, we offered a possible quantification of their statement by showing that all five normalized model dimensions (of which N2) are needed to achieve a good discrimination. None of them appeared intrinsically sufficiently sexually dimorphic to generate over 90% discrimination.

Our results therefore indicate that the reason why certain facial features appear to play a role in gender discrimination assessed by psychophysical experiments might be found in the natural statistics of the faces which, we argue, subjects are sensitive to, due to phylogenic or ontogenic factors. We note with Burton *et al.* (1993) that it would be of interest to study the correlation of such results with human performances, in particular in the cases of weak or incorrect classification. Such a study is left for future work.

Finally, we note that many psychophysical studies have indicated the differential importance of individual facial regions in attractiveness judgments (Cunningham, 1986; Meerdink *et al.*, 1990; Langlois *et al.*, 1994; Perrett *et al.*, 1994; Etcoff, 1994). Interestingly, most of their conclusions are compatible with ours which are derived simply on the basis of gender. Our results suggest, therefore, that attractiveness might be assessed by subjects of a given gender on the basis of the facial attributes which statistically characterize the faces of the opposite gender (Fellous, 1995). Our position occupies a middle ground between gestaltist and componential theories: we suggest that a subset of features (our model dimensions) are, when taken together, sufficient to explain most of the differences between male and female fiducial distances, and that it is possibly on the basis of their particular average values (Fig. 7) that attractiveness judgment is attributed. It is the case that the closer the male (female) model measurements from their average (indicated by M on Fig. 7) are, the higher is the probability for a correct classification as "male" (female). However, there are model fiducial distance values which yield even better scores than the average values [for males (females), values that yield a score greater (smaller) than average], and it is our prediction that particular faces exhibiting such properties would be assessed as more attractive than the average (Perrett *et al.*, 1994), should only metric properties be used for such an assessment.

Neurophysiological data on face perception

It is interesting to note that the activity of some face-selective neurons correlates to some extent with some metric aspects of the face. Yamane *et al.* (1988), for example, found cells in the inferotemporal cortex of the monkey that correlate with some fiducial distances, especially when considered together. Among these distances, a few are related to the hairstyle, whereas other are purely facial. Interestingly, they report five neurons which correlate with linear combinations of hair-related and face-dependent measurements, the latter being E1, L1, W1 and B2. Our results therefore suggest that the reason why L1, B2, and to a lesser extent, W1, might appear in their analysis is related to the intrinsic structure of the data, since L1 and B2 are part of the model derived statistically from the data, and since W1 shows a significant amount of discriminatory power on data in which noise has been introduced. This result suggests that some face-selective cells might be sensitive to statistically significant gender features, as a way of coding for the face in general. In other words, face-selective cells would encode face information such as gender, and would use this information to achieve face identification, for example.

In a second study, Young & Yamane (1992) suggested that AIT face-selective cell populations possibly code for physical properties (fiducial distances) of the face. During a discrimination task, they found that most of the recorded cells correlated their responses with distances relating the hairline to other facial points (H3, H4,...), but were not reported to correlate with any other distances such as the ones we derived in this study. We would suggest that such a result is expected, since all

faces presented to the monkeys were male faces. It is possible that, within this particular data set, hairline-related distances appear to be the fiducial distances that bear the maximal amount of discriminatory power, and are therefore coded by the neurons, for the particular purpose of the task the monkey was involved in. According to the results of our study, one might expect to find neurons correlating with distances such as L1, W4, B2, N2 or E3, should the data set include female faces with hairstyles that do not, alone, allow for their discrimination with male faces. More generally, it is expected that a statistical study such as the one conducted here, could provide valuable insights as to what could be coded by face-selective cells (for example, fiducial distances), depending on the data set, and the experimental paradigm.

CONCLUSION

This study demonstrated to what extent gender discrimination can be achieved on the sole basis of facial metric information. We derived five normalized dimensions which allow for over 95% correct gender discrimination on the training set, compatible with the actual performance of human subjects (Bruce *et al.*, 1993). Classification yielded 90% correct gender prediction on a test set including various facial expressions.

We argued that most of the recent psychophysical experiments on gender have derived results that are compatible with ours, and can be understood in the context of the statistical analysis presented above. If it is clear that gender perception does not rely solely on facial metric information, humans seem to perceive gender according to facial features (such as the distances we used) which statistically contribute the most to the difference between male and female faces. We suggest that gender perception processes are therefore strongly related to the inherent structure of the visual stimuli, and that they take advantage of it to achieve fast and reliable gender recognition. Evidence is presented for the possible correlation of these features with the activity of face-selective cells in the temporal lobes. Finally, we argued that male and female attractiveness depend in part on facial metric information, making attractiveness judgments rely on dimensional features which are often seen and which characterize the gender of the stimulus.

REFERENCES

Brown, E. & Perrett, D. (1993). What gives a face its gender? *Perception, 22*, 829–840.

Bruce, V. A., Burton, M., Hanna, E., Healey, P., Mason, O., Coombes, A., Fright, R. & Linney, A. (1993). Sex discrimination: how well do we tell the difference between male and female faces? *Perception, 22*, 131–152.

Brunelli, R. & Poggio, T. (1993a) Caricatural effects in automated face perception. *Biological Cybernetics, 69*, 235–241.

Brunelli, R. & Poggio, T. (1993b) Face recognition: features versus templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 1510*, 71–86.

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

Cunningham, M. R. (1986). Measuring the physical in physical attractiveness: quasi-experiments on the sociobiology of female facial beauty. *Journal of Personality and Social Psychology, 505*, 925–935.

Darwin, C. (1872). *The expression of emotions in man and animals*. London: Julian Friedman (published in 1979).

Ekman, P. (1992). Facial expressions of emotion: an old controversy and new findings. *Philosophical Transactions of the Royal Society of London B, 335*, 63–69.

Ekman, P. & Davidson, R. J. (1993). Voluntary smiling changes regional brain activity. *Psychological Science, 45*, 342–345.

Enlow, D. (1982). *Handbook of facial growth*. Philadelphia: W. H. Saunders.

Etcoff, N. L. (1994). Beauty and the beholder. *Nature, 368*, 186–187.

Fellous, J.-M. (1995). A neural code for face representation in the temporal lobes: from V1 receptive fields to IT "face" selective cells. Ph.D. Thesis, University of Southern California, Los Angeles, September 1995.

Ferrario, V., Sforza, C., Pizzini, G., Vogel, G. & Miani, A. (1993). Sexual dimorphism in the human face assessed by Euclidian distance matrix analysis. *Journal of Anatomy, 183*, 593–600.

Haig, N. (1986). Exploring recognition with interchanged facial features. *Perception, 15*, 235–247.

Hosie, J. A., Ellis, H. D. & Haig, N. D. (1988). The effect of feature displacement on the perception of well-known faces. *Perception, 174*, 461–474.

Laidman, H. (1979). *Figures/faces: a sketcher's handbook*. New York: Viking Press.

Langlois, J. H., Roggman, L. A. & Musselman, L. (1994). What is average and what is not average about attractive faces? *Psychological Science, 54*, 214–220.

Meerdink, J. E., Garbin, C. P. & Leger, D. W. (1990). Cross-gender perceptions of facial attributes and their relation to attractiveness: do we see them differently than they see us? *Perception and Psychophysics, 483*, 227–233.

Nakdimen, K. A. (1984). The physiognomic basis for sexual stereotyping. *American Journal of Psychiatry, 1414*, 499–503.

Perrett, D. I., May, K. A. & Yoshikawa, S. (1994). Facial shape and judgements of female attractiveness. *Nature, 368*, 239–242.

Roberts, T. & Bruce, V. (1988). Feature saliency in judging the sex and familiarity of faces. *Perception, 174*, 475–481.

Samal, A. & Iyengar, P. A. (1992). Automatic recognition and analysis of human faces and facial expressions: a survey. *Pattern Recognition, 251*, 67–77.

Valentin, D., Abdi, H., O'Toole, A. & Cottrell, G. W. (1994). Connectionist models of face processing: a survey. *Pattern Recognition, 279*, 1209–1230.

Wiskott, L., Fellous, J.-M., Kruger, N. & von der Malsburg, C. (1995). Face recognition and gender determination. Proceedings of the International Workshop on Automatic Face and Gesture Recognition-Zurich, June 1995.

Yamane, S., Kaji, S. & Kawano, K. (1988). What facial features activate face neurons in the inferotemporal cortex of the monkey? *Experimental Brain Research, 73*, 209–214.

Young, M. & Yamane, S. (1992). Sparse population coding of faces in the inferotemporal cortex. *Science, 256*, 1327–1331.

Acknowledgements—Portions of this study used the FERET (Face Recognition Technology) database of facial images collected under the ARPA/ARL FERET program. The author would like to thank József Fiser, Dr. M. Arbib, Dr. I. Biederman and Dr. C. von der Malsburg for helpful discussions and Dr. M. Burton and an anonymous reviewer for their helpful comments. This study was supported in part by a grant from the US Army Research Laboratory (01/93K-0109) and an AFSOR grant (F49620-93-1-0109) to Dr C. von der Malsburg.