INVESTIGATING THE SPACE OF CHERNOFF FACES

Robert J.K. Jacob

Computer Science and Systems Branch
Naval Research Laboratory
Washington, D.C.

and

Department of Statistics
George Washington University
Washington, D.C.

Since Herman Chernoff first proposed the use of faces to represent multidimensional data points (1971), investigators have applied this powerful graphical method to a variety of problems. With the method, each data point in a multidimensional space is depicted as a cartoon face. Variation in each of the coordinates of the data is represented by variation in one characteristic of some feature of the face. For example, one coordinate of the data might be represented by the curvature of the mouth, another by the size of the eyes, and so on. The overall multidimensional value of a datum is then represented by the expression on a single face. The strength of this method lies in the ability of an observer to integrate the elements of a face into a unified mental construct. With most other representations, it is more difficult for an observer to combine the coordinates of a data point to form a single mental impression.

Observers of Chernoff faces have, however, voiced one principal complaint: The perceived meaning of the resulting

©1983 by Academic Press, Inc.
All rights of reproduction in any form reserved
ISBN 0-12-589320-5

face plots is dependent on the way that the coordinates of the data are assigned to the features of the face (e.g., Bruckner 1978, Fienberg 1979, Kleiner and Hartigan 1981). Each user of Chernoff faces claims to have developed an intuitively satisfying way to make this assignment for his own problem, but only a few investigators have attempted to find more general solutions. Chernoff and Rizvi (1975) provided an empirical measurement of the importance of the choice of this assignment. Tobler (1976) and Harmon (1977, 1978, 1981) both proposed criteria that an optimal data-to-face transformation should meet but did not provide methods for constructing transformations that meet their criteria. Kleiner and Hartigan (1981) proposed a method, based on clustering, for assigning data coordinates to the construction parameters of two new displays, designed specifically for their method. The same assignment procedure could usefully be applied to other displays including faces (Jacob 1981).

The basic problem is to find a transformation from the space of the data to be plotted to the space of possible face parameter vectors such that an observer's perception of the relationships among the resulting faces corresponds to the spatial configuration of the original data points. For example, given a set of data points in a Euclidean space, the assignment of data coordinates to face parameters should be chosen so that the similarity between any two plotted faces perceived by an observer reflects the Euclidean distance between the corresponding data points. The problem is that it is difficult to measure the spatial configuration formed by an observer's perception of similarities between faces. In principle, a multidimensional scaling procedure could be used to construct a description of an observer's "mental configuration" of a set of stimuli from a collection of measurements of similarities between all pairs of the stimuli (Torgerson 1952, 1958). However, the face plots are

principally useful for data of high dimensionality, and, for such, a scaling procedure would require an impossibly large number of these similarity measurements.

A very simple approach is to look for "good" or "bad" regions of the space of face parameter vectors with respect to Euclidean data. A "good" region is one in which observers' perceptions of similarities between faces correspond well to the Euclidean distances between the points represented by the faces. Some good or bad regions may be found in an ad-hoc fashion, and then the bad ones improved by trial and error. Specifically, wherever uniform changes in a single face parameter produce widely differing changes in perceptions, a "bad" region is found. As Chernoff's (1973) geologist notes, a principal problem area is the face outline. It carries too much perceptual significance, partly because the placement of all the other features depends on the outline and partly because the parameter axes used to describe it do not map onto straight and orthogonal perceptual configurations. That is, as one moves along a straight line parallel to an axis in the space of face parameter vectors, there will be some points at which the perceptual effect of the move changes in kind, rather than degree. This problem was reduced for the outline by creating a single new parameter that controls a scaling function of the ratio of the eccentricities of the two ellipses in the outline and replacing the two original eccentricity parameters with it. These and other similar ad-hoc modifications resulted in a plotting program (Jacob 1976b) in which the outline is less prominent, the parameters that describe it are more nearly orthogonal, and the ill-behaved outer fringes of the space are avoided. As an example, Mezzich and Worthington (1978) used Chernoff's original face program in a careful comparative study of a number of multivariate methods. Faces did not yield particularly good

performance in the comparison; but, in examining their faces, the outline seems to have been especially significant. Many of their data points lay in the "bad" region of the space of face parameter vectors, where small parameter variations cause disproportionate changes in the outline. Re-plotting their data with the new program provides a better representation of the underlying clusters.

I. AN EMPIRICAL METHOD FOR EVALUATING REGIONS OF THE SPACE OF FACE PARAMETER VECTORS

Beyond making ad-hoc improvements in the face parameters, a more general question arises: Is there some empirical way to identify "good" or "bad" regions of the space of face parameter vectors? This is, in a sense, the reverse of multidimensional scaling. Instead of using subjects' similarity judgments to infer the spatial configuration of the stimuli, this procedure starts by hypothesizing a particular configuration. Then, subjects' similarity judgments are compared to the distances derived from the hypothesized configuration. If there is a "good" region of the face space, where the subjective distances match the model, then the procedure has "found" the spatial configuration for that region. (If not, the hypothesized model may simply have been an inastute choice.) The advantage of this method is that considerably fewer distance observations are needed to confirm or refute an hypothesized model than to infer one. The disadvantage is, of course, that one must guess at a model rather than letting the data suggest one. Another disadvantage, as seen in the case presented here, is that even a large amount of data may not be sufficient to draw distinctions about regions within a many-dimensional space.

As in many conventional scaling experiments, subjects here

were not explicitly asked to rate the similarities between the stimuli, but such similarities were computed from their responses to a more familiar task. The data consist of subjects' responses to a categorization experiment. While they were originally obtained for a different purpose (Jacob, Egeth, and Bevan 1976), the responses are here used to investigate differences between regions within the face parameter space. Five widely-separated points *(prototypes)* were chosen from a 9-dimensional space. Then, a cluster of 10 more points *(deviants)* was generated randomly around each prototype, near it in Euclidean distance. Subjects were given the 5 prototypes and then instructed to assign each of the 50 deviants in turn to one of the prototypes ("the one they thought it belonged with"). These responses can be used to infer subjects' perceptual similarities between the 5 prototypes and the 50 deviants. This yields a total of 250 measurements-considerably less than the number of all possible pairs of stimuli in the set, which would be used for conventional scaling.

If many of the 24 subjects clustered a particular deviant with a particular prototype, then the perceptual similarity between those two points (combined over subjects) must be large-that is, greater than the similarity between that deviant and any of the other 4 prototypes. Since subjects did not differ markedly in their ratings, a combined measure for each of the 250 similarities was obtained by counting the number of times each deviant was judged to belong with each prototype. A linear regression was then performed to relate these similarities to the distances obtained by measuring the original points using some hypothesized distance metrics. Both the original Euclidean model

$$\sqrt{\sum_i (a_i - b_i)^2}$$

and a city block distance model

$$\sum_i |a_i - b_i|$$

were tested. The regressions predicted the subjects' similarity ratings from the hypothesized distances. Each yielded a correlation coefficient measuring how well the subjects' ratings corresponded to the model.

The simplest comparison that can be made is between the overall correlation coefficient for faces and that for the same experiment replicated using two other multivariate displays (polygons and digit matrices). Table I shows that faces fit either distance model better than polygons or digits.

One may also ask which model fits better-Euclidean or city block? It appears that the Euclidean model fits the responses *slightly* better for faces and polygons, but there is no difference for digits. In fact, for the particular 250 distance measurements in question, the two metrics give very similar relative distances, so it would be difficult to discriminate between them using this set of measurements.

TABLE I. Correlation Coefficients from Overall Linear Regressions

	Correlation coefficient for	
Display type	*Euclidean model*	*City block model*
Face	*0.556*	*0.511*
Polygon	*0.219*	*0.181*
Digits	*0.209*	*0.201*

Next, individual regions within the space may be considered. To do this, the 250 observed distances were divided into groups;

a regression was performed for each group; and the residuals for the groups were compared. Each of the original distance measurements was made between one of the 5 prototypes and one of the 50 deviants. One way to divide them is according to which of the 5 prototypes each distance is connected to. This gives 5 groups of 50 distances each. The average absolute residual for each of these, using a Euclidean distance model, is shown in Table II. To visualize the results, the actual prototypes are shown in Figure 1. For faces, ratings of distances from Prototype 2 appear least Euclidean, while those from Prototype 3 are most. One may hypothesize that Prototype 2 is the most distinctive face-that its location in the observers' perceptual configuration is further from the center than the location of the corresponding point in the 9-dimensional Euclidean data space. Prototype 2 is also the only face for which some of its deviants have mouths that extend beyond the face outline. Similar analyses may be made for the other display types. For polygons, observed distances from Prototype 1 fit the Euclidean model least well. Since it is the only regular polygon in the set, it is likely to be more perceptually distinctive than its location in the data space would indicate.

TABLE II. Residuals from Linear Regressions by Prototype[a]

Display type	*Prototype*				
	1	*2*	*3*	*4*	*5*
Face	*3.492*	*4.514*	*3.071*	*3.493*	*3.507*
Polygon	*4.099*	*3.515*	*3.143*	*3.392*	*3.491*
Digits	*2.928*	*3.669*	*2.644*	*2.908*	*2.801*

[a]*The similarity measurements from which these residuals are computed were integers between 0 and 24.*

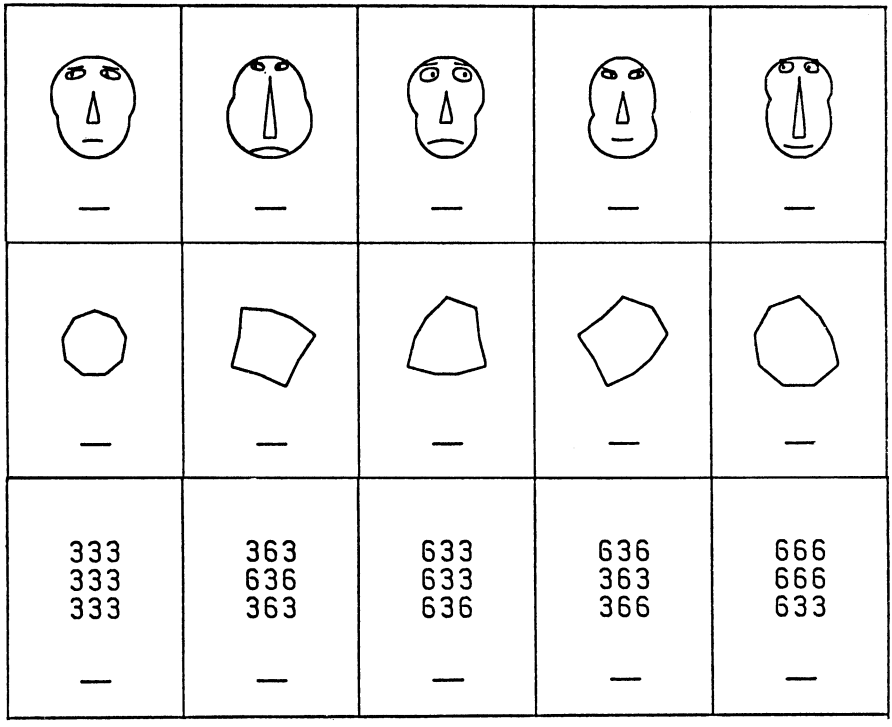

FIGURE 1. Five Prototypes Used for Distance Ratings

Aggregating all the distances *from* a particular prototype may be too gross a measure. Perhaps distances from some prototype to one region fit a Euclidean model well, and those from the same prototype to another region do not. The distances were therefore further divided so that all measurements from Prototype *i* to any of the 10 deviants around Prototype *j* were placed together in a single group. This yields 25 groups of 10 distances-one group for each pair that consists of a prototype and a cluster of 10 deviants. The residuals for the faces in Table III show some finer distinctions. The poor fit seen above for distances to Prototype 2 is attributable to the area immediately around Prototype 2; distances between Prototype 2 and the areas around the other prototypes are not particularly distorted. Prototype 5 exhibits a similar property. Thus, the regions of the face parameter space immediately surrounding these two points are relatively poor ones for representing Euclidean data.

TABLE III. Residuals from Linear Regressions by Prototype and Group

Display type	From deviant group	To prototype				
		1	2	3	4	5
Face	1	4.976	4.011	2.248	3.018	3.172
	2	4.532	7.862	3.403	3.056	2.914
	3	1.523	3.662	3.248	3.668	2.640
	4	2.909	4.349	3.290	4.510	2.437
	5	3.522	2.687	3.167	3.213	6.375
Polygon	1	4.937	2.656	2.906	2.821	2.611
	2	4.407	4.865	2.744	3.638	3.427
	3	5.226	2.243	4.969	3.145	2.148
	4	3.662	4.817	1.722	3.957	3.485
	5	2.265	2.995	3.374	3.401	5.784
Digits	1	3.021	3.766	1.987	3.633	3.206
	2	3.910	6.014	1.883	2.589	2.260
	3	3.338	2.720	3.286	2.228	3.454
	4	2.549	2.530	2.517	3.434	1.568
	5	1.825	3.315	3.547	2.659	3.514

Finally, a much finer measure may be considered. (This analysis was performed for the faces only.) Perhaps there are smaller good or bad regions of the space than the 5 or 25 considered above, and all distance measurements that *cross through* those regions of the space are distorted. To examine this, the 9-dimensional space was divided into uniform-size regions. Then, the average distortion for all distance measures that cross through a region was used as a measure of how well that region depicts Euclidean distances. The first step was to lay a grid over the 9-dimensional space to divide it into regions. The simplest grid, which divides each coordinate axis in half, was used; and this carves the space into 512 regions. (A finer grid would be difficult to support with only 250 observations. However, note that the two-level grid is not sufficient to test the hypothesis that regions near the center of the space are better than those near the edges, since all the regions are the same distance from the center.)

Each of the 250 subjective similarity observations was then compared to the similarity predicted by the hypothesized Euclidean metric. The absolute value of the resulting prediction error for each similarity observation provided a rating of the extent to which that observation fit the model. Next, a rating of the extent to which all the observations that cross through a particular region are Euclidean was needed. The rating for a region was defined as the mean of the ratings for all of the distance lines that traversed it, weighted by the portion of each distance line that lay in that region. This was approximated by dividing each distance line into 100 equal segments and assigning an equal share of the rating for that line to each of the regions that contained an endpoint of a segment. Of the 512 regions, only 189 are traversed by any of the distance lines. Hence this procedure yielded ratings of the "Euclidean-ness" of 189 out of the 512 regions of the 9-dimensional subspace of possible face parameter vectors that was

used to represent the data. It is not easy to apprehend the resulting mass of numerical data. One approach was to divide the regions into groups according to their relative ratings and then cluster the regions within each group with other nearby regions. One then searches for a large agglomeration of good or bad regions that could be characterized more concisely than the collection of smaller regions; but such was not found when these data were clustered.

The problem of obtaining an overall picture of these detailed results clearly calls for a good statistical graphics technique! Figure 2 plots the center point of each of the 10 "most Euclidean" regions as a Chernoff face, followed by the 10 regions around the median, and the 10 worst regions. The rating for each region is shown below the face. Here the data are easy to understand, but what one learns is that the good and bad regions appear to be scattered fairly uniformly throughout the space. No one large area or axis direction appears to be particularly good or bad with respect to Euclidean distances. It is possible that the changes made to the face plotting program removed the most obvious non-Euclidean regions from the space, and a finer analysis requires more observations.

II. AN EMPIRICAL METHOD FOR ASSIGNING COORDINATES TO FACIAL FEATURES

Another approach to approximating a multidimensional scaling procedure is to perform it for some specific type of data. Instead of inferring the perceptual configuration of a set of faces from a collection of distance judgments, here one hypothesizes a set of axes for the perceptual configuration and asks subjects to rate the faces on the given axes. This is appropriate when a particular type of data have a known or well-established set of axes. The size of the scaling experiment

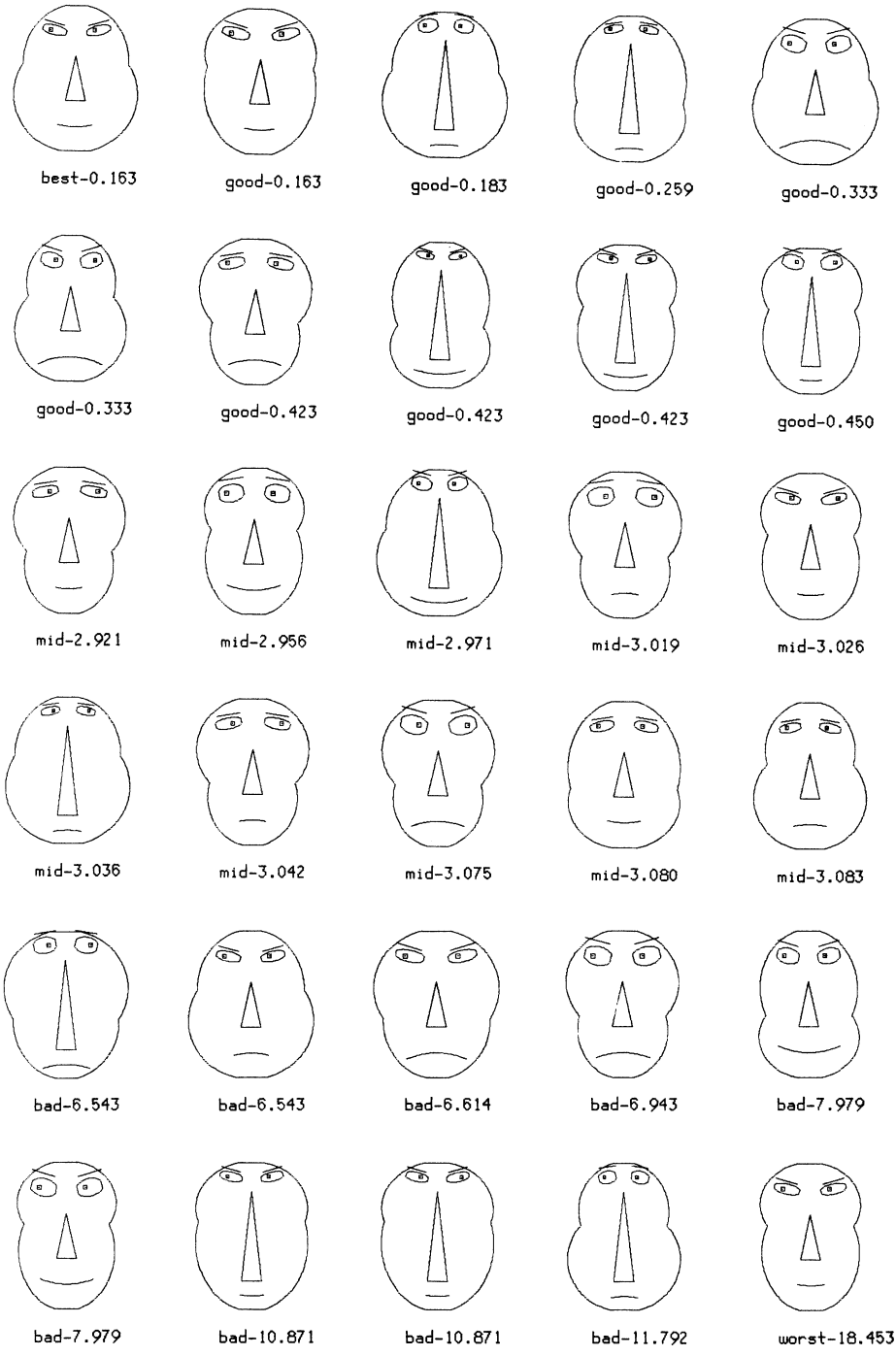

FIGURE 2. Facial Representation of the Centers of the Most, Median, and Least Euclidean Regions

that must now be performed is reduced by its square root, since only individual, rather than pairwise, ratings of the stimuli are needed. Each application of such a study yields a good mapping from one particular type of input data to faces. Obviously, it is most suitable where data of the same general type will be plotted again.

This technique was used to generate a data-to-face mapping for a type of psychiatric data. The resulting mapping was tested and found to be significantly more suggestive than other arbitrary mappings (Jacob 1978, 1976a). The data to be plotted consisted of the *Hypochondriasis, Depression, Paranoia, Schizophrenia,* and *Hypomania* scales of the Minnesota Multiphasic Personality Inventory (MMPI) psychiatric test. A naive approach would have been simply to assign the 5 scales arbitrarily to 5 facial features. Since the test was being given to many patients, however, it was worthwhile to conduct the scaling experiment to obtain a better mapping from the 5-dimensional space of MMPI scores to the 18-dimensional face parameter vector space. Once obtained, the same mapping could be used to plot the MMPI scores of new patients. It should provide a display of an MMPI score that intuitively suggests the meaning of the score, by exploiting observers' preconceptions or stereotypes. (The validity of these stereotypes is not at issue; it is only necessary that they be measurable and widely held.)

The first experiment attempted to find a linear transformation from the MMPI score space to the face parameter space. Unlike most applications of faces, in which each data coordinate is assigned to one face parameter, any linear transformation was permitted here. That is, each data coordinate could control any linear combination of all 18 face parameters. Subsequent analysis of the results for higher-order interactions showed the linear model to be an adequate approximation. A set of 200 faces was generated from uniform random face parameter vectors.

The subject then rated each face along the 5 MMPI axes. In effect, she was indicating what MMPI score she thought a person who looked like each of the faces might receive. A multiple linear regression of the face parameter vectors on the MMPI scores was computed, resulting in a regression equation that could then be used to produce a face for any given MMPI score.[1]

The resulting data-to-face transformation is described by an 18 by 5 matrix of regression coefficients. Again, the problem of interpreting it is solved with the faces themselves. To display the transformation, a series of face parameter vectors was computed, corresponding to equally-spaced points along the axes of the MMPI score space (i.e., the points represent patients who each have only one psychiatric disorder). Figure 3 shows the resulting plot; each row depicts a series of hypothetical patients with increasing amounts of a single disorder, from *0* (the first column), which represents an inverse amount of the disorder, through *1*, representing no disorder (the origin of the MMPI space), to *4*, representing a large, extrapolated amount of the disorder. These faces (particularly those in the column labeled *3*) appear to resemble common stereotypes of the personality traits they are purported to represent. In fact, the subject had not rated faces like these; most of the faces in the original stimuli had been reported to have more than one disorder. The scaling procedure generated an intuitively appealing linear transformation from MMPI scores to faces in an objective manner.

The resulting transformation can be used to plot new MMPI scores as faces. In a subsequent experiment, 30 subjects were each given 50 questions, each consisting of a text description of a patient with an hypothetical MMPI score vector and 5 faces, one of which was generated from this transformation. Subjects chose the "correct" face with significant ($p < 0.0005$) accuracy.

[1] *If the predicted face parameters are compared to the original parameters, the mean squared error over all components of all the vectors is 0.075, where vector components were between 0 and 1.*

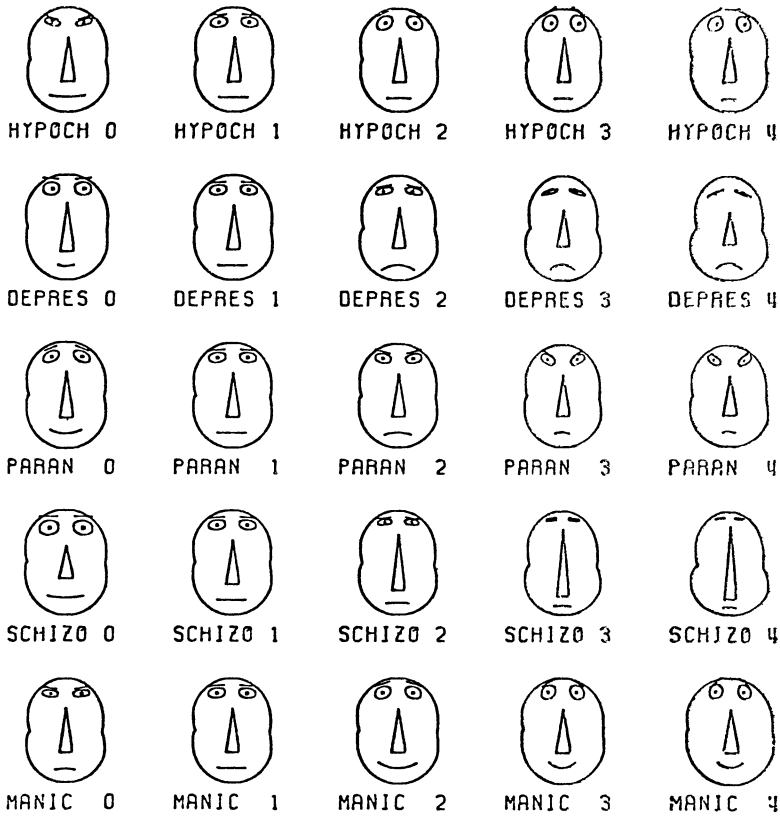

FIGURE 3. Facial Representation of MMPI Scores

These results can provide insight into the face parameter space in another way. The facial representations of all possible MMPI scores comprise a 5-dimensional subspace of the 18-dimensional space of all possible face parameter vectors. There remains an orthogonal 13-dimensional subspace of facial variation. Variation in this subspace should have little effect on the MMPI-related meaning of a facial expression. To see this subspace, Figure 4 shows faces corresponding to 5 arbitrarily-selected, mutually-orthogonal axes in this 13-dimensional subspace. (As before, each row of faces represents movement along one axis; and here the center of each row corresponds to the origin of the space.) Examination of faces in this subspace suggests that all facial variation could be partitioned into two distinct orthogonal subspaces. One (the 5-dimensional range of the transformation from MMPI scores to faces) depicts an *emotional* component of facial expressions, and the other (the remaining orthogonal subspace) depicts an *identification* component, representing variations that help distinguish the faces of individuals from one another but transmit little emotional content. In fact, most of the variation in the former space is in the facial features that a person can move (eyes, mouth) to indicate emotion, while that of the latter is in the immovable features (nose, outline) that distinguish individuals from one another.

To test the hypothesis that variation in the orthogonal subspace carries little psychological significance, 15 faces were generated. Five contained random variation in the MMPI space plus no variation in the orthogonal non-MMPI space, while the remaining 10 varied randomly only in the non-MMPI space. A group of 32 subjects rated the MMPI-varying faces significantly ($p < 0.0005$) more psychologically disturbed than the others. Thus the space of the faces appears to be divisible into two orthogonal subspaces, each of which carries a distinct type of facial information.

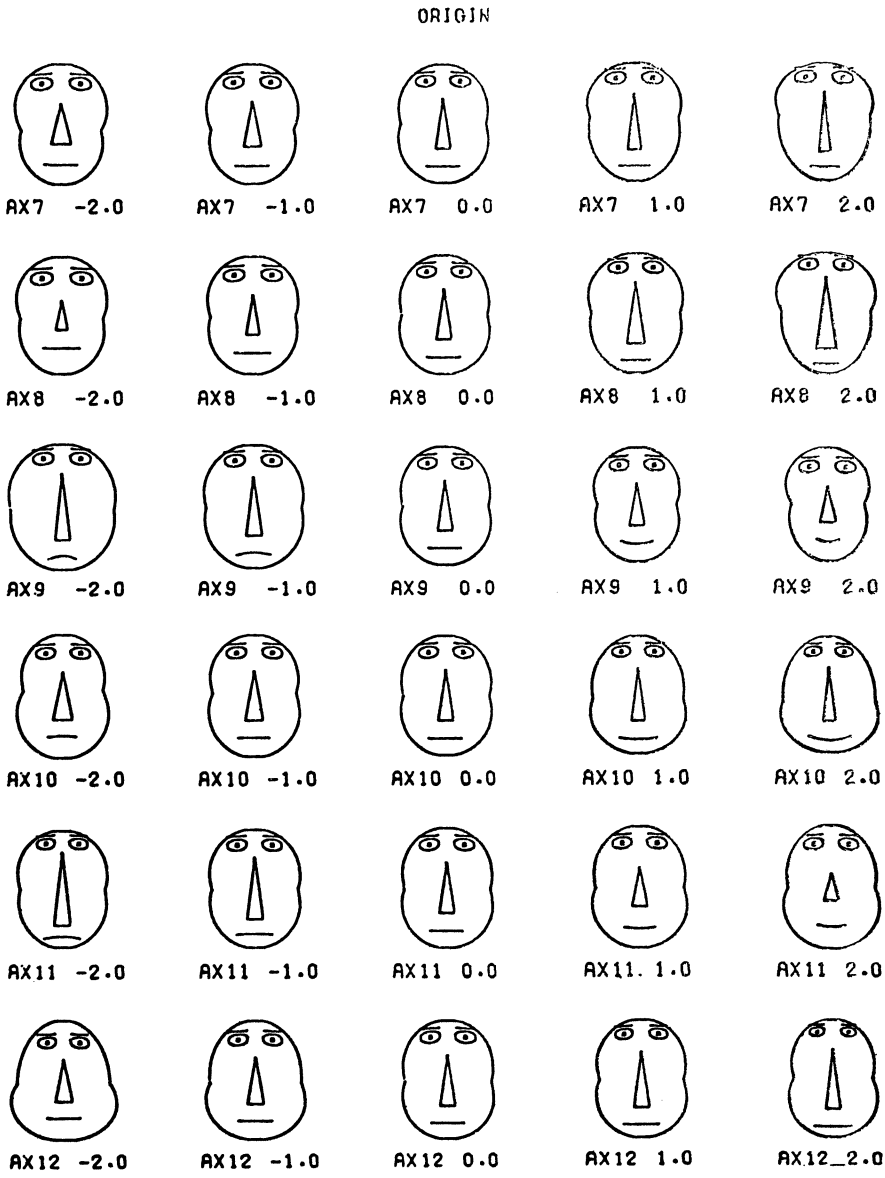

FIGURE 4. Facial Representation of the Axes of the Orthogonal Subspace

III. CONCLUSIONS

Using Chernoff faces to their best advantage to represent multivariate data requires that the perceptual configuration in which the faces are perceived resemble the configuration of the original data points. This means observers' judgments of similarities between faces should reflect the distances between the original data points. Multidimensional scaling solves this problem in principle, but it is not practical because of the dimensionality of the spaces. Three other approaches were considered:

- Intuitively obvious deficiencies in the way face parameters map onto faces can be corrected by trial and error, as was shown for the face outline parameters.
- Regions of the space of possible face parameter vectors can be evaluated according to the degree to which subjective distance judgments within each region match a particular distance model, as was attempted for the 9-dimensional data with the Euclidean distance model. If "bad" regions of the space can be identified from this procedure, one could then avoid them when producing face plots.
- A special case of the scaling paradigm can be used for any specific type of data, as was shown for the MMPI data. The result was a demonstrably mnemonic transformation from data to faces and also the identification of two basic components of facial variation.

ACKNOWLEDGMENTS

Professor William H. Huggins provided inspiration and insight for much of this work, while the author was a student at

the Johns Hopkins University. Professors Howard Egeth and William Bevan at Johns Hopkins also provided a great deal of valuable guidance.

Portions of this research were supported by a contract between the Johns Hopkins University and the Engineering Psychology Programs, Office of Naval Research; and by the U.S. Public Health Service Hospital in Baltimore.

And, of course, Herman Chernoff is father to all of this research. Happy Birthday!

REFERENCES

Bruckner, L.A. (1978). On Chernoff Faces, pp. 93-121, *Graphical Representation of Multivariate Data,* ed. P.C.C. Wang, Academic Press, New York.

Chernoff, H. (1971). The Use of Faces to Represent Points in n-Dimensional Space Graphically. Technical Report No. 71, Dept. of Statistics, Stanford University.

Chernoff, H. (1973). "The Use of Faces to Represent Points in k-Dimensional Space Graphically." *J. Amer. Statist. Ass. 68,* 361-368.

Chernoff, H. and Rizvi, M.H. (1975). "Effect on Classification Error of Random Permutations of Features in Representing Multivariate Data by Faces." *J. Amer. Statist. Ass. 70,* 548-554.

Fienberg, S.E. (1979). "Graphical Methods in Statistics." *Amer. Statist. 33,* 165-178.

Harmon, L.D. and Hunt, W.F. (1977). "Automatic Recognition of Human Face Profiles." *Computer Graphics and Image Processing 6,* 135-156.

Harmon, L.D., Kuo, S.C., Ramig, P.F., and Raudkivi, U. (1978). "Idenification of Human Face Profiles by Computer." *Pattern Recognition 10,* 301-312.

Harmon, L.D., Khan, M.K., Lasch, R., and Ramig, P.F. (1981). "Machine Identification of Human Faces." *Pattern Recognition 13,* 97-110.

Jacob, R.J.K., Egeth, H.E., and Bevan, W. (1976). "The Face as a Data Display." *Human Factors 18,* 189-199.

Jacob, R.J.K. (1976a). Computer-produced Faces as an Iconic Display for Complex Data. Doctoral dissertation, Johns Hopkins University. University Microfilms Inc. Order No. 76-22926.

Jacob, R.J.K. (1976b). PLFACE Program, Available from the author upon request.

Jacob, R.J.K. (1978). Facial Representation of Multivariate Data, pp. 143-168, *Graphical Representation of Multivariate Data,* ed. P.C.C. Wang, Academic Press, New York.

Jacob, R.J.K. (1981). "Comment on Representing Points in Many Dimensions by Trees and Castles." *J. Amer. Statist. Ass. 76,* 270-272.

Kleiner, B. and Hartigan, J.A. (1981). "Representing Points in Many Dimensions by Trees and Castles." *J. Amer. Statist. Ass. 76,* 260-269.

Mezzich, J.E. and Worthington, D.R.L. (1978). A Comparison of Graphical Representations of Multidimensional Psychiatric Diagnostic Data, pp. 123-141, *Graphical Representation of Multivariate Data,* ed. P.C.C. Wang, Academic Press, New York.

Tobler, W.R. (1976). The Chernoff Assignment Problem. Unpublished paper.

Torgerson, W.S. (1952). "Multidimensional Scaling: I. Theory and Method." *Psychometrika 17,* 401-419.

Torgerson, W.S. (1958). *Theory and Methods of Scaling,* Wiley, New York.