In A. B. Watson, ed., Digital Images and Human Vision, Bradford Press: Cambridge Mass, pp.141-148, 1993.

IMAGE QUALITY: A MULTIDIMENSIONAL PROBLEM

A. Ahumada and C. H. Null
NASA Ames Research Center, Moffett Field CA

Abstract

Observers naturally vary in the degree to which they find different display artifacts objectionable, making image-based, quantitative prediction of display quality seem hopeless when artifacts are suprathreshold. Multidimensional scaling techniques, that are relatively easy to compute, can identify dimensions of image quality that are differentially weighted by observers. As compared with average observer ratings, these dimensions should relate more directly to physical properties of the stimuli and thus improve the predictability of image quality.

Introduction

The design of displays and image compression methods could proceed more efficiently if one could predict observer ratings of display quality from physical properties of displayed images. Zetzsche and Hauske (1989) report correlations between their image-quality model predictions and mean subjective ratings ranging from 0.95 down to 0.74, depending on the types of distortion in the images. Although they suggest that improvements in their models might allow adequate predictions, it is equally easy to suppose that models of image quality based on fixed visual system properties are fundamentally limited in their ability to predict subjective image-quality ratings. Even if the ratings reflect only the detectability of artifacts, there is significant variation in the contrast sensitivity functions and other critical visual paramenters (Ginsburg, Evans, Seculer, and Harp, 1982; Owsley, Seculer, and Siemsen, 1983).

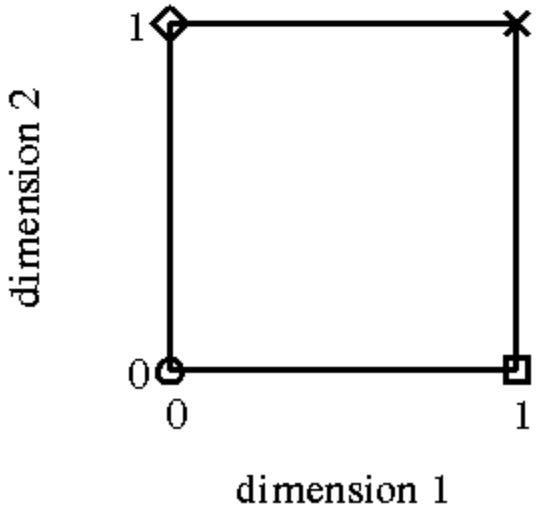
When the display artifacts are suprathreshold, the observers' different experiences with artifacts are bound to lead to differential weighting of the artifacts. Also, images are used for a range of purposes, so the objectionability of artifacts would also depend on the observers' presumptions as to the use of the display. The presumed use would also be expected to vary with the observers' past experiences. Because different groups of observers are bound to differ systematically, there is little likelihood that a measure calibrated for one group could perfectly predict results for all other groups.

The dimensions to which different observers give different weight should be more stable than the weightings over different groups of observers. It may be easier to generate predictions for these dimensions based on physical properties of the displays. The methods described here do not average out observer variability; they use inter-observer variability to extract dimensions of image quality. We describe one such method in detail, give an example of how it can find multiple dimensions of image quality, and describe some of the problems associated with the method. The method, a multidimensional scaling one, is a variant of the MDPREF method (Carroll, 1972). In many multidimensional scaling methods, observers provise estimates of the similarities or distances between stimuli in perceptual space (Kruskal and Wish, 1978; Shepard, Romney, and Nerlove, 1972). All that is needed for the methods described here are quality ratings.

Preference Factoring

An Imaginary Example

To see how dimensions of image quality can be recovered when observers weight them differently, imagine four displays that vary on two dimensions as illustrated in Fig. 1.



	○	◇	□	×
A	1	1	2	2
B	1	2	1	2
C	1	2	2	3

Fig. 1. The positions of the symbols indicate the values of 4 imaginary displays on two hypothetical quality dimensions. The three rows of the table represent ratings of these displays by 3 observers (A, B, and C) who differentially weight the two quality dimensions as described in the text.

An observer who only pays attention to the first dimension might rate the quality of the displays as in the first row of ratings in the figure. The second row of ratings shows the ratings of an observer who only attends to the second dimension, and the third row shows ratings of an observer who pays equal attention to both dimensions.

```
Mathematica 2.0 for SPARC
Copyright 1988-91 Wolfram Research, Inc.
In[1]:= Needs["Statistics`DescriptiveStatistics`"]

In[2]:= rawdata = {{1.,1.,2.,2.},
                   {1.,2.,1.,2.},
                   {1.,2.,2.,3.}};
In[3]:= data = rawdata - (Mean /@ rawdata)
Out[3]= {{ -0.5, -0.5,  0.5, 0.5 },
         { -0.5,  0.5, -0.5, 0.5 },
         { -1.0,  0.0,  0.0, 1.0 }}

In[4]:= SingularValues[data]

Out[4]= {{{-0.408, -0.408, -0.816},
          { 0.707, -0.707,  0.   }},

         {1.732, 1.},

         {{0.707,  0.   , 0.   , -0.707},
          {0.   , -0.707, 0.707,  0.   }}}
```

Fig. 2. A computer dialog in Mathematica. Keyboard inputs are in bold.

Fig. 2 shows a lightly edited transcript of a Mathematica dialog, with the input in bold type (Wolfram, 1991). The input is the data matrix from the previous figure. First the mean rating for each observer is subtracted from the ratings by that observer. Next the singular value decomposition (SVD) routine is called. The output is three matrices: a 2 x 3 matrix giving weights for our 3 observers on 2 new dimensions, a 1 x 2 list giving the weights of the two dimensions, and a 2 x 4 matrix giving the weights of the 4 displays on the 2 dimensions. Fig. 3 plots these results.

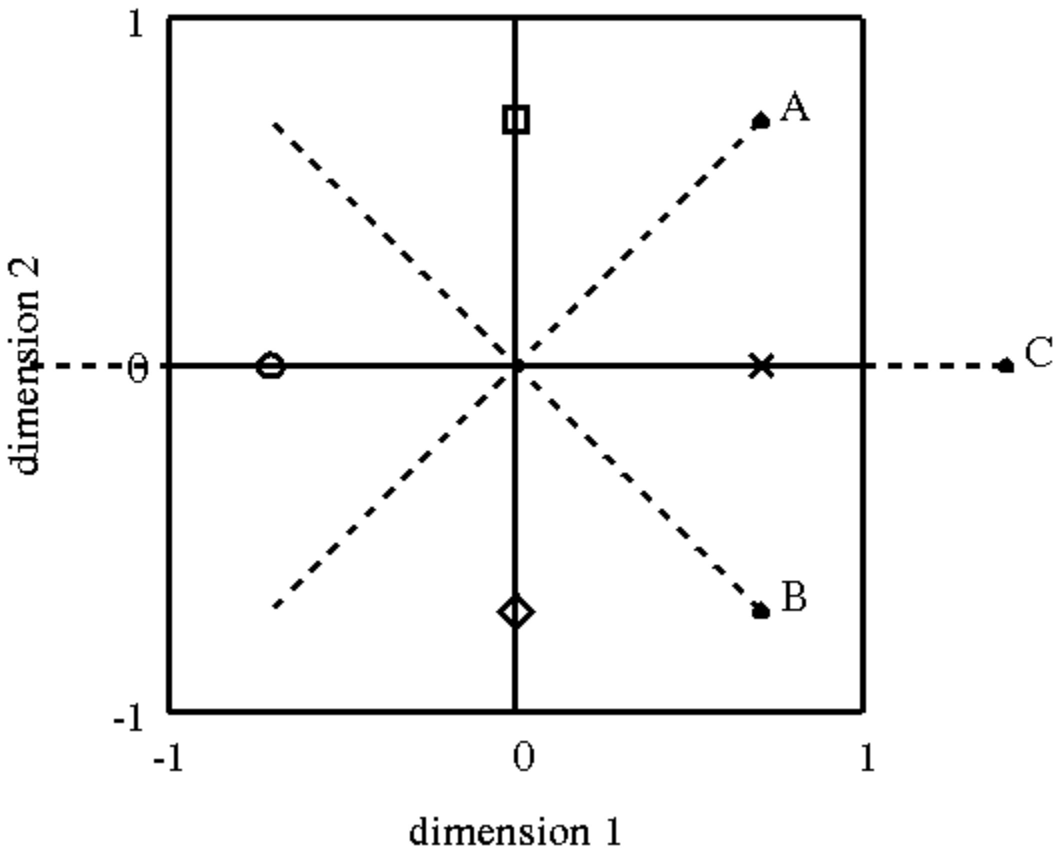


Fig. 3. Results of the SVD example of Fig. 2 plotted as in Fig 1. The positions of the same 4 symbols indicate the values of 4 imaginary displays on the two quality dimensions found by the SVD. The solid points are the subject weights of the 3 observers (A, B, and C) multiplied by the singular values. The dashed lines indicate the directions of the observers' preferences in the 2-dimensional space.

The symbols representing the 4 displays are in the same configuration as in Fig. 1, but this configuration has been rotated, reflected, and translated. Dimension 1 now represents the overall average quality, the single dimension which can best (in a least squares sense) represent the input data matrix. The new dimension 2 does not have such a nice name; it is represents the difference between the two original dimensions. The observer weights are also shown, scaled by the two dimension weights so that all of the SVD information is contained in the graph. The first subject is represented by the solid point in the lower right hand corner. The normalized ratings are the dot product of a subject's vector with each of the display vectors. These dot products are the projections onto his dotted line multiplied by his distance from the origin.

The SVD analysis has factored the data matrix into a display matrix and an observer matrix, representing the directions of the observers' preferences in display dimensions. The output dimensions are not the original dimensions, but they are related to them by a linear transformation.

Some Theory

Imagine now an experiment in which n_o observers are asked to give quality ratings to n_d displays, resulting in a matrix

$$(R)_{i,j} = r_{i,j},\ i=1,...,\ n_o,\ j=1,...,\ n_d. \quad (1)$$

One Dimension

If everyone agreed on the ratings perfectly, so that,

$$r_{i,j} = q_j,\ i=1,\ldots,\ n_o,\ j=1,\ldots,\ n_d,\quad (2)$$

all the rows of R are the same and the rank of R (the smaller of the number of linearly independent rows or columns) is one. In this case it would be possible for some single-valued function of the displays to predict the ratings, since they depend only on the display.

Suppose all the subjects have the same underlying quality ratings, but use different numerical scales, so that

$$r_{i,j} = a_i q_j + b_i,\ i=1,\ldots,\ n_o,\ j=1,\ldots,\ n_d.\quad (3)$$

R now has rank 2, since the rows are linear combinations of a column (q_j) and a column of 1's. An SVD analysis would find 2 dimensions even though there is really only one of interest. Subtracting the observer's mean rating from each rating leaves a data matrix of differences with elements

$$r'_{i,j} = r_{i,j} - \overline{r}_i$$

$$r'_{i,j} = a_i(q_j - \overline{q}\ ),\ i=1,\ldots,\ n_o,\ j=1,\ldots,\ n_d.\quad (4)$$

R' again has rank 1 and the SVD analysis returns observer weights and display weights that can be obtained simply by averaging.

Multiple Dimensions

The preference factoring model assumes that there are a number n(q) of different quality dimensions that observers, in general, weight differently, so that

$$r_{i,j} = \sum_{k=1}^{n_q} a_{i,k}\, q_{j,k},\ i=1,\ldots,\ n_o,\ j=1,\ldots,\ n_d,\quad (5)$$

where $q_{j,k}$ is the quality of display j on dimension k and $a_{i,k}$ is the weight that observer i gives to that dimension. The rank of this rating matrix is now, in general, the smaller of n_o, n_q, and n_d -1, assuming each observer's mean rating has been subtracted out. If the observers use the weightings and qualities of Eq. (5) and we find two matrices that multiply together as in Eq. (5) to give the matrix $R=(r_{i,j})$, we have probably not found the same dimensions used by the observers. There are many equivalent factorizations of a matrix R, since any invertible n_q x n_q matrix T can post-multiply $A = (a_{i,k})$ and then its transposed inverse post-multiply $Q = (q_{j,k})$ and the resulting matrices satisfy Eq. (5). In general, we can only find the dimensions up to an arbitrary invertible linear transformation T.

The Rating Matrix

If observers use numbers to report quality, it is not reasonable to assume that the numerical scales are the same for different observers. Subtracting the mean rating for an observer from the ratings for that observer can help in two ways. As we saw earlier, it makes the results blind to differences among observers in the absolute positioning of their scales, removing an additive factor dimension from the analysis. Second, the additive model has only to fit the smaller range of variations in the region of the mean, analogous to the improved functional fit of a Taylor expansion about the center of the desired region, rather than about zero.

The results can also be made independent of the scale factor used by the observer by dividing the observers' ratings by their standard deviations. This normalization (the default for most factor analysis programs) can cause problems if the scales were originally reasonably comparable, since then differences that one observer found inconsequential can be made as large as the important differences of another observer. If repeated judgments are available, their variability can be used to normalize the scale factors. Dividing an observer's ratings by the pooled repeated-judgment standard deviation allows each observer's responses to be weighted by an estimate of their precision. More complicated unidimensional scaling procedures such as Thurstone scaling allow the observers to use non-uniform rating functions, but these methods usually require more responses (Torgerson, 1958).

A rating matrix can also be generated from ranking or paired comparison experiments. If ranks are used, the only source of differential spacing is the different ordering, not accounted for by the multiple dimensions, so the results benefit from many closely spaced displays. Paired comparison data can be can be converted to ratings by scaling procedures, of which the simplest is computing the percentage of time each stimulus was chosen over the others (Torgerson, 1958).

Singular Value Decomposition

As was shown in the example above, the SVD can solve the problem of finding the rank n(r) of a rating matrix and then finding observer weights and display values that can be multiplied together to recreate the ratings as in Eq. (5). The SVD represents the rating matrix as the sum of the products of three numbers,

$$r_{i,j} = \sum_{k=1}^{n_r} u_{i,k}\ v_{j,k}\ w_k,\quad i=1,\ldots,\ n_o,\ j=1,\ldots,\ n_d, \quad (6)$$

where U = ($u_{i,k}$) is a matrix of normalized weights for each observer on each quality dimension, V = ($v_{j,k}$) is a matrix of normalized values for each display on each quality dimension, and W = (w_k) is an array of strengths for each dimension (the singular values). Eq. (6) can be put in the same form as Eq. (5) by arbitrarily associating the dimension strengths with the observer weights, that is setting $a_{i,k} = u_{i,k}w_k$ and $q_{j,k} = v_{j,k}$. A defining property for the SVD is that for any $n_q < n_r$ the first n_q rows of U and V and the first n_q w_k's are dimensions for a least squares representation of R by a matrix of rank n_q (Eckart and Young, 1936). In other words, if we constrain ourselves to a preference model with only n_q quality dimensions, then the best (least squares) version is given by the first n_q rows of U and V and the corresponding values of W. The sum of squares of the singular values w_k for k greater than n_q is the squared error of the representation. Another resulting property is that the rows of U are orthogonal to each other, as are the rows of V. The w_k values allow the vector lengths of all these rows to be set to one. The resulting representation is unique except for the signs of the rows of U and V, if the w_k are assumed positive.

Although the SVD is readily available in matrix operation subroutine collections (Dongarra et al., 1979; The Mathworks, Inc., 1991; Becker and Chambers, 1984; Wolfram, 1991), its results can also be obtained from a principal components factor analysis or eigenvector and eigenvalue analysis available in many statistical packages (Wilkinson, 1987; Dixon et al., 1977; Nie et al., 1975). Some programs will accept the data matrix as input and can subtract the observers' means (covariance about the mean option) and divide by the observers' standard deviations (correlation about the mean). Some will output the display weights multiplied by the singular values (test factor loadings). Some also output the observer weights (factor scores), and some also provide transformations (rotations) to possibly more interpretable dimensions. In the worst case you must subtract the means yourself and form the symmetric covariance matrix,

$$o_{i,i'} = \sum_{j=1}^{n_d} r_{i,j}\ r_{i',j},\quad i,i'=1,\ldots,\ n_o, \quad (7)$$

to get the program to output the subject weights; and then provide

n_o

$$d_{j,j'} = \mathop{S}_{i=1} r_{i,j}\ r_{i,j'},\ j,j'=1,...,\ n_d, \quad (8)$$

to get the display image configuration. The singular values are the square roots of the eigenvalues of either covariance matrix.

Some authors view each observer as selecting a preference direction in the stimulus space, represented by his unit length row vector in the U matrix. They leave the rows of U alone and multiply the rows of V by the w_k. Others prefer to leave the stimulus representation normalized and consider the w_k to represent the relative weights that the observers place on the dimensions. When the w_k span a large range, it is convenient to leave the model in the three separate parts so that plots of both subject weights and stimulus values can be scaled uniformly.

Error

Suppose that the ratings are all based on a single underlying quality value q_j, but that other unsystematic factors affect the ratings.

$$r_{i,j} = q_j + e_{i,j},\ i=1,...,\ n_o,\ j=1,...,\ n_d. \quad (9)$$

where $e_{i,j}$ is independent Gaussian noise,

$$e_{i,j} == N(0,s). \quad (10)$$

Now, with probability one the matrix $r_{i,j}$ has full rank, but if we use SVD to factor it, the dimension corresponding to the largest singular value will correspond most closely to the noise-free ratings, and the other dimensions will represent the noise. Although the true noise dimensions are equal in their singular values, the SVD will place the estimates of their singular values in descending order so that the actual number of dimensions is not obvious even in this simple case.

Let us return to the case of n_q dimensions of display quality (Eq. (5)), but with added noise as above.

$$r_{i,j} = \mathop{S}_{k=1}^{n_q} a_{i,k}\ q_{j,k} + e_{i,j},\ i=1,...,\ n_o,\ j=1,...,\ n_d, \quad (11)$$

If s is small enough, the SVD may result in a large drop in the singular values w_k after w_{nq}, but dimensions that we recover will not be exactly a linear transformation of the original dimensions. We are then left with the problem of recovering "true" dimensions that has plagued practitioners of factor analysis and multidimensional scaling. Solutions to these problems have been proposed and shown to do well in certain cases (Harmon, 1967; Torgerson, 1958). Statistical packages that include factor analysis usually also provide options to attempt to rotate the dimensions to make them more meaningful or interpretable.

A Real Example

Farrell, Trontelj, Rosenberg, and Wiseman (1991) report rankings of 12 displays by 18 observers. The displays were compressed versions of the classic 512 x 512 monochrome Lena image (the woman with the hat) from the USC data base. Six of the images were compressed using a discrete cosine transform (DCT) method (Wallace, 1991) and six images were compressed using the non-uniform sampling and interpolation (NSI) method of Rosenberg (1990). The bit-rate of the compressed images ranged between 1.7 to 0.3 bits/pixel. At low bit-rates, the DCT method created visible block-like patterns, while the NSI method produced vertical and horizontal smearing. Fig. 4 illustrates the nature of the artifacts in the highest compression (lowest bit rate) conditions.

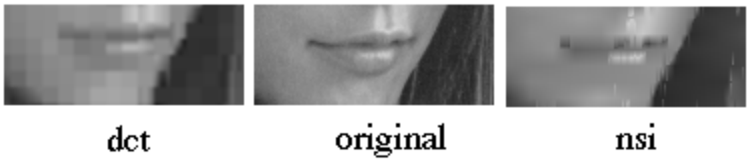


Fig. 4. Sections of the original image, a DCT compressed image, and an NSI compressed image from the study by Farrell et al., (1991), enlarged to illustrate the nature of the compression artifacts.

Observers viewed all 66 pairwise comparisons of the 12 images three times. Ratings were formed by computing the average proportion of times each each display image was chosen to be more like the original, which was displayed with the two compressed images.

Table 1. Singular values w_k from the data of Farrell et al., (1991).

k	1	2	3	4	5	6	7	8	9	10	11
w_k	4.28	0.61	0.37	0.34	0.31	0.22	0.22	0.13	0.13	0.08	0.05

Table 1 shows the 11 singular values from the SVD of the ratings. (Only 11 remain after first subtracting the mean response (0.5) from all ratings.) Although it is not clear from the pattern of sizes in Table 1, it is clear from the plot of the stimulus weights in Fig. 5, that the SVD recovers at least two meaningful dimensions differentially weighted by the observers.

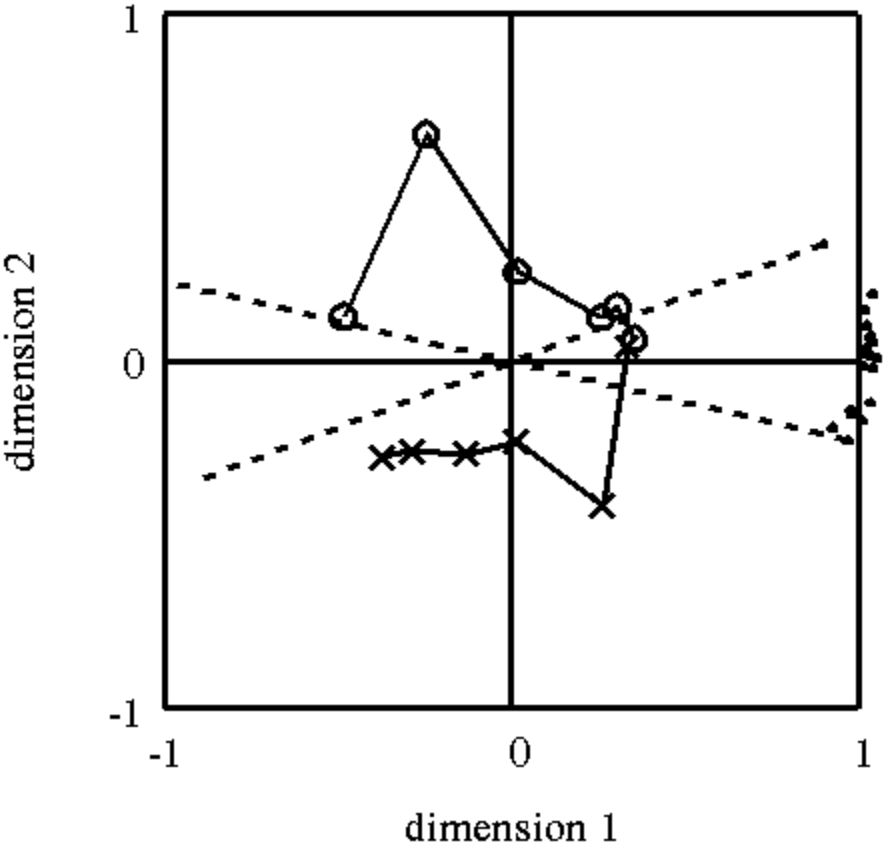


Fig. 5. The first two dimensions from the SVD of the data of Farrell et al. (1991).

Circles indicate the $v_{j,k}$ of the 6 DCT compressed images; x 's indicate the $v_{j,k}$ of the six NSI compressed images. Solid lines link neighboring compression levels, which decrease from left to right. Solid points are 18 subject weights $u_{i,k}$ multiplied by the singular values w_k. The orders of the rankings for two very different observers can be obtained by dropping perpendiculars to the dashed lines through their points.

The first dimension correlates with the amount of compression in bits/pixel and the second distinguishes between the two types of compression. The observer weights have been multiplied by the corresponding singular values and then plotted on the same figure. The order of the predictions for an observer can be seen by dropping perpendiculars from the compressed image points to the line through the origin and the observer point. These points of intersection can be converted to numerical predictions of proportion of times chosen by multiplying their distance from the origin by the distance of the observer's point (here nearly unity) and then adding 0.5.

Unfortunately, the data do not tell us the direction of the dimensions that the observers used. For example, the observers could have had two dimensions, "blockiness" and "blurriness", or the observers could have had one dimension of "quality" and another dimension of "nearsightedness", presuming (for the sake of argument) the NSI artifacts to be less visible to nearsighted observers. Whether the multidimensionality is important depends on the lability of the observer weights. If the observer weights are stable, representing something like the distribution of acuity in the population, the second dimension has little importance since most of the observers weights are close to zero on dimension 2. If the weights are easily modified by experience or task, the relative quality of the two compression methods could change greatly if the predicted ratings could change as much as the the difference between the two most extreme observers.

INDSCAL

Carroll and Chang's INDSCAL procedure (Carroll, 1972; Arabie, Carroll, and DeSarbo, 1987) is a variant of preference factoring that puts greater demands on the form of the data and makes stronger assumptions about the rating process, but provides, in return, dimensions with uniquely determined directions. The data must have three fully crossed factors: observers, displays, and conditions. That is, every observer must rate each display in each condition. The additional assumption is that the effect of the the conditions on either the observer weights or the display weights is purely multiplicative in each dimension, so that Eq. (5) becomes

$$r_{i,j,m} = \sum_{k=1}^{n_q} a_{i,k}\, q_{j,k}\, c_{m,k}, \quad i=1,\ldots,\, n_o,\; j=1,\ldots,\, n_d,\; m=1,\ldots,\, n_c, \quad (12)$$

where $c_{m,k}$ are weights for the n_c conditions on the n_q quality dimensions. Since the $c_{m,k}$ vary with both m and k, this equation does not in general reduce to Eq. (5). Also, in general, except for normalizations and reflections, the directions of the dimensions satisfying Eq. (12) are unique.

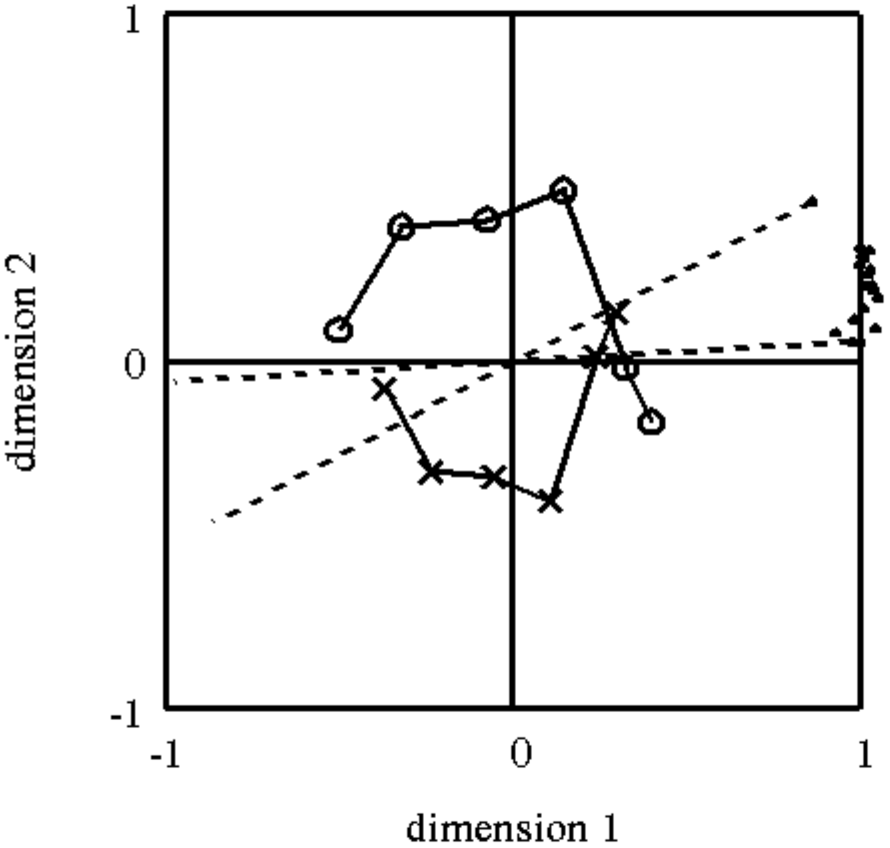


Fig. 6. The data of Fig. 5 analyzed by the INDSCAL procedure. Circles indicate the $q_{j,k}$ $c_{m,k}$ for the 6 DCT compressed images; x 's indicate these products for the 6 NSI compressed images. Solid lines link neighboring compression levels, which decrease from left to right. The solid points are 18 observer weights $a_{i,k}$. The dashed lines in Fig. 6 have the same interpretation as in Fig. 5 and are drawn for the same two observers.

An Example

The same data used to illustrate two-way preference factoring can be used to illustrate 3-way factoring by arbitrarily letting the two types of compression represent two condition levels for 6 display types, which then are the compression levels. The resulting INDSCAL analysis is illustrated in Fig. 6. In this example only two dimensions can result, since INDSCAL cannot extract more dimensions than the smallest of n_o, n_d, and n_c. The configuration of the 6 compressions for the two dimensions is shown for both the DCT and the NSI compressed images in Fig. 6. They are the same configuration scaled by the $c_{m,k}$ for the two different conditions. The configurations of Figs. 5 and 6 are extremely similar. Both show the same general pattern. When there is little compression, the methods are of course similar. Of more interest, the method differences get smaller again for the largest compressions. The shape similarity prevents Fig. 6 from showing the apparent difference for the two methods in the level of compression at which the observer differences are maximal.

The major possible advantage of the INDSCAL analysis is not well illustrated by this example. The directions of the dimensions in Fig. 6, while similar to those of Fig. 5 are determined by the functional interaction of Eq. (12), rather than being determined by which dimension can predict the most variance in the data. When there are only two dimensions, one can easily transform to another coordinate system by drawing in the new axes and projecting onto one axis using lines parallel to the other. For more dimensions, it is difficult to find "best" axes and the fixed solution of the INDSCAL method can be very helpful.

SUMMARY

Preference factoring can be easily performed using readily available computer programs. This multidimensional scaling procedure allows dimensions of image quality to emerge if observers vary in the relative weights they give to the dimensions. An example analysis of preference data for images compressed by two different methods shows that the method can find multiple dimensions even when there is strong agreement about the ratings from most of the observers. Dimensions found by these methods are likely to be better predicted by image properties. In addition to demonstrating the multidimensional nature of the quality ratings, the analyses showed that the differential effect of the two compression methods on different observers is greatest at moderate levels of compression.

ACKNOWLEDGMENTS

We are grateful to J. Farrell, H. Trontelj, and their colleagues at Hewlett Packard for the data and encouragement, and to H. Peterson for helpful comments. This work was supported by NASA RTOP 505-64-53 and 506-71-51.

REFERENCES

Arabie, P., Carroll, J. D., and DeSarbo, W. S. (1987) "Three-Way Scaling and Clustering." *Sage University Paper series on Quantitative Applications in the Social Sciences, 07-065,* Beverly Hills and London: Sage Publications.

Becker, R. A., and Chambers, J. M. (1984) *S: An Interactive Environment for Data Analysis and Graphics,* Belmont, Calif.: Wadsworth.

Carroll, J. D. (1972) "Individual Differences and Multidimensional Scaling." *In Multidimensional Scaling. Vol. I, Theory,* eds., R. N. Shepard, A. K. Romney, S. B. Nerlove, New York: Seminar Press.

Dixon, W. J., Brown, M. B., Engelman, L., Frame, J. W., and Jennrich, R. I. (1977) *BMDP-77: Biomedical Computer Programs, P Series,* Berkeley: Univ. of California Press.

Dongarra, J., Moler, C. B., Bunch, J. R., and Stewart, G. W. (1979) *LINPACK User's Guide,* Philadelphia, Pa.: SIAM.

Eckart, C., and Young, G. (1936) "The Approximation of One Matrix by Another of Lower Rank." *Psychometrika* **1**: 148-158.

Farrell, J. E., Trontelj, H., Rosenberg, C., and Wiseman, J. (1991) "Perceptual Metrics for Monochrome Image Compression." *Society for Information Display Digest* **22**: 631-634

Ginsburg, A. P., Evans, D. W., Sekuler, R., and Harp, S. A. (1982) "Contrast Sensitivity Predicts Pilots' Performance in Aircraft Simulators." *American Journal of Optometry and Physiological Optics* **59**: 105-109.

Harmon, H. (1967) *Modern Factor Analysis.* New York: McGraw-Hill.

Kruskal, J. B., and Wish, M. (1978) *Multidimensional Scaling.* Sage University Paper series on Quantitative Applications in the Social Sciences, 07-011, Beverly Hills and London: Sage Publications.

MATLAB User's Guide. (1991) Natick, Mass.: The MathWorks, Inc.

Nie, N. H., Hull, C. H., Jenkins, J. G., Steinbrenner, K., and Bent, D. N. (1975) *SPSS: Statistical Package for the Social Sciences, Second Edition,* New York: McGraw-Hill.

Owsley, C., Sekuler, R., and Siemsen, D. (1983) "Contrast Sensitivity throughout Adulthood". *Vision Research* **23**: 689-699.

Rosenberg, C. (1990) "A Lossy Image Compression Algorithm Based on Nonuniform Sampling and Interpolation of the Image Intensity Surface." *Society for Information Display Digest* **21**: 388-391.

Shepard, R. N., Romney, A. K., and Nerlove, S. B. (1972) *Multidimensional Scaling, Vol. I, Theory,* New York: Seminar Press.

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

Wallace, G. (1991) "The JPEG Still Picture Compression Standard." *Communications of the ACM* **34**: 31-44.

Wilkinson, L. (1987) *SYSTAT: The System for Statistics,* Evanston, I11: SYSTAT, Inc.

Wolfram, S. (1991) *Mathematica: A System for Doing Mathematics by Computer,* Redwood City, Calif.: Addison-Wesley.

Zetzsche, C., and Hauske, G. (1989) "Multiple Channel Model for the Prediction of Subjective Image Quality." In *Human Vision, Visual Processing, and Digital Display*, ed. B. Rogowitz, Proc. 1077, pp. 209-216, Bellingham, Wash. SPIE.