Multidimensional Modeling of Image Quality

JEAN-BERNARD MARTENS

Invited Paper

In this paper, multidimensional models of image quality are discussed. In such models, alternative images, for instance, obtained through different processing or coding of the same scene, are represented as points in a multidimensional space. The positioning is such that the correlation between geometrical properties of the points and the subjective impressions mediated by the corresponding images is optimized. More specifically, perceived dissimilarities between images are monotonically related to interpoint distances, while the strengths of image quality attributes (such as perceived noise and blur, or image quality) are, for instance, monotonically related to point coordinates along specified directions. The goal of multidimensional models is to capture subjective impressions into a single picture that is easy to interpret. We apply multidimensional models to two existing data sets to demonstrate that they indeed account very well for experimental data on image quality. The program XGms is introduced as a new interactive tool for constructing multidimensional models from experimental data. Although XGms is introduced here within the context of image-quality modeling, it is also potentially useful in other applications that rely on multidimensional models.

Keywords*—Image quality, image-quality models, multidimensional scaling.*

I. Introduction

People are very used to making judgements about many aspects of the things they encounter, be it objects, situations, or other people. Judgements made by different people often agree remarkably well when being compared explicitly. When judging the quality of images, for instance, people especially agree on the more perceptual aspects such as the brightness, contrast, and colorfulness of the images. This is partly due to the fact that their peripheral senses (for hearing, seeing, etc.) are very similar. At the same time, this agreement on perceptual aspects does not prohibit that people can have very different opinions about more cognitively related aspects of the same images. One aspect on which people often disagree is the aesthetical quality of images. There may be many different, often personal, reasons why some images (photographs or motion pictures) are considered to better convey the essence of the subject being depicted and appeal more to some people than to others. In this paper, we will demonstrate how multidimensional models can capture both common aspects and differences between individual judgements.

In engineering, we adopt a restricted perspective on image quality. We consider the input images as given and are interested in evaluating the effect of alternative image coding, processing, and/or display systems on the perceived quality of these images. Even within this limited scope, we can still distinguish many different aspects or attributes to image quality. The perceived sharpness, contrast, colorfulness, naturalness of colors, etc. may all be influenced in different ways by such technical systems. Alternative systems may also introduce different distortions, such as noise, sampling, and quantization artifacts that obviously do not belong to the scene, but are created by the imaging system. People often agree very closely in their judgements about the strengths of these image-quality attributes. The importance that they assign to these individual attributes in reaching their overall judgement on the technical image quality may however differ. This fair amount of agreement between people on underlying attributes will be referred to as the *principle of homogeneity of perception* [1] and forms an important motivation for the multidimensional scaling (MDS) approach [1]–[3] toward image-quality modeling that is presented in this paper. This principle states that a single multidimensional configuration representing the stimuli underlies the attribute and quality judgements by all subjects. Differences in attribute judgements of subjects can be accomodated in MDS by the fact that the mapping from the joint multidimensional stimulus configuration to the one-dimensional (1–D) attribute judgements may vary per attribute and per subject. Constructing the underlying stimulus configuration from experimental judgements by subjects or predicting it on the basis of instrumental (i.e., objective) measurements on the images, hence, becomes a major step in image-quality model building [4]. It should be noted that the proposed MDS approach, which will be

Manuscript received June 10, 1999; revised June 6, 2001.

The author is with the IPO Center for User-System Interaction, 5600 MB Eindhoven, The Netherlands (e-mail J.B.O.S.Martens@tue.nl).

Publisher Item Identifier S 0018-9219(02)00731-4.

illustrated by an example in Section II, is fundamentally different from most existing approaches [5] that attempt to model image quality as a 1-D entity without trying to provide insight into underlying image-quality attributes.

There are both theoretical and practical reasons for being interested in creating models for image quality. On the one hand, models guide the way we reason about a psychological concept such as image quality. The way we experiment with image quality and the data processing that is applied to experimental data is guided by such models [6]. Often, we collect experimental data in order to verify if they are in agreement with existing or emerging models. On the other hand, because of the large amount of imaging equipment being produced nowadays, there is a substantial economical interest in being able to predict the effect of variations in the technical parameters of such systems on the resulting quality. Especially in the case of alternative systems with similar functionality, perceived (image) quality is one of the major discriminating factors between products from the point of view of the user.

An instrumental quality measure is defined here as an algorithm for predicting the stimulus configuration that underlies human judgements. The usefulness of such an algorithm obviously depends on how well the resulting stimulus configuration correlates with human judgements. Instrumental models can not only make the design of new systems more efficient, but can also be used to monitor the performance of existing systems. An advantage of multidimensional models over 1-D quality models in such a monitoring application is that multidimensional models can also provide insight into which attributes of image quality are failing when the overall quality is insufficient.

The scientific field that is mostly concerned with measuring subjective sensations such as image quality is called psychophysics [7]. In a recent review paper [6], we have described in detail some of the major concepts and models underlying the psychophysical measurement of image quality and its attributes. In the current paper, we are mainly concerned with discussing how such psychophysical measurements on several different quality attributes can be combined into one overall model.

In Section III, we summarize some of the concepts and terminology that are needed when discussing the psychophysical measurement and modeling of image quality. We introduce the three major experimental paradigms used in image-quality evaluation: single-stimulus attribute scaling, double-stimulus difference (or preference) scaling, and double-stimulus dissimilarity scaling. We also discuss the important distinction between metric and nonmetric scales. We refer to the above-mentioned review paper [6] for a more in-depth discussion of these topics. Next, in Section IV, we discuss multidimensional modeling of image quality and give a brief historical overview of the application of MDS in image-quality measurement and modeling. In Section V, we discuss in detail how the parameters of a multidimensional model can be estimated from experimental data. We introduce a new interactive program, called XGms,[1] that can estimate a metric or nonmetric multidimensional model from data obtained for one or more subjects, using any combination of the three above-mentioned experimental paradigms. This program is a modified version of an existing program XGvis [8] that can only handle metric dissimilarity data from a single subject. Finally, in Section VI, we illustrate MDS modeling on some existing data sets.

II. Example of Multidimensional Modeling

In order to make the discussion on multidimensional modeling more concrete, we start with an example. More specifically, we describe the multidimensional modeling of experimental data concerning the quality of images degraded by noise and blur [9]. All combinations of four levels of blur with a binomial kernel and four levels of Gaussian additive white noise for three different scenes were used in an experiment. The blur was characterized by the standard deviation σ_b of the filter kernel (σ_b values of 0, $\sqrt{2}$, 2, and $2\sqrt{2}$). The noise was characterized by the standard deviation σ_n of the Gaussian probability density function (σ_n values of 0, 7, 10, and 14 for grey values in the range [0,255]). The noise was added after the images were blurred and the processed images were quantized to 8 bits/pixel. All pictures contained 512×512 pixels, but only a subregion of 240×470 pixels was viewed in the experiments. This restricted region was needed in order to allow simultaneous display of two images. The (cropped) images without noise or blur are shown in Fig. 1.

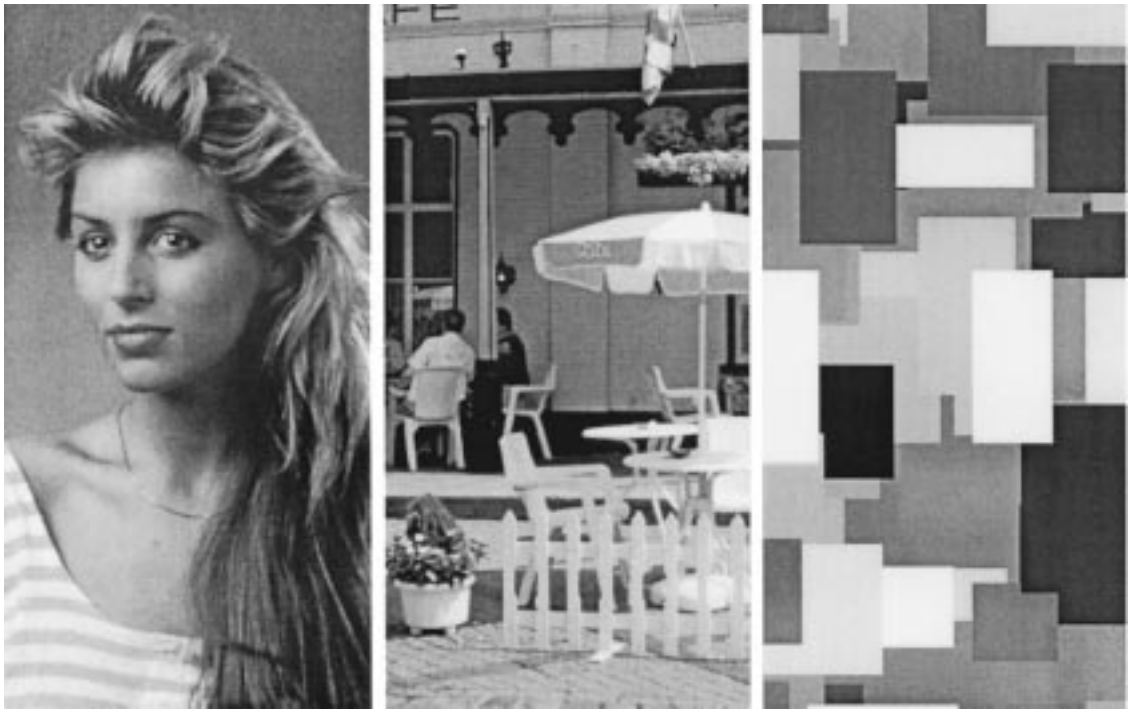
The images were converted to 50-Hz noninterlace video and displayed on a CCID-7351B high-resolution monitor. The grey-value-to-luminance characteristic of the monitor was measured and a lookup table was determined such that the relationship between the grey value g and the luminance L for the combined chain (lookup table, digital-to-analog video convertor, and monitor) became

$$L = \max\left[L_{\min}, L_{\max}\left(\frac{g}{g_{\max}}\right)^{\gamma}\right] \tag{1}$$

with $g_{\max} = 255$, $L_{\max} = 60$ cd/m^2, $L_{\min} = 0.2$ cd/m^2, and $\gamma = 2.5$. This calibrated characteristic was verified by a second luminance measurement.

Most viewing conditions satisfied the ITU-R BT.500 recommendation [10]. The viewing distance was 1.5 m, which was equivalent to six times the height of the monitor. Between two successive stimuli, a uniform adaptation field was displayed during the time it took subjects to enter their response by means of a keyboard. The minimum duration of this adaptation field was 2 s, while its luminance of 9 cd/m^2 was approximately equal to the average luminance of the images.

[1]XGms and XGobi Software: XGms is a modified version, developed by the author, of the existing program XGvis. The XGvis and XGobi programs are available via http://www.research.att.com/~dfs. The XGms program is also available for noncommercial use. E-mail requests can be made to J.B.O.S.Martens@tue.nl.



(a) (b) (c)

Fig. 1. Scenes used in experiments with blur and noise. (a) Wanda. (b) Terrace. (c) Mondrian.

Five subjects, aged between 25 and 39 years and with normal or corrected-to-normal visual acuity between 1.5 and 2 measured on a Landolt chart[2] at a distance of 5 m participated in a double-stimulus dissimilarity experiment. Each subject was presented all $15 \times 16/2 = 120$ different image pairs in random order and was asked to score the dissimilarity between the images in a pair using an integer number in the range from 0 to 10.

The same five subjects, plus two additional ones, also took part in three single-stimulus scaling experiments with the same images. In three separate experiments, subjects rated perceived noisiness, blur, and quality. The 16 stimuli were presented four times, in random order, to each of the subjects (resulting in $16 \times 4 \times 3 = 192$ attribute scores per subject). Again, integer scores between 0 and 10 were used by the subjects to express their judgements.

The XGms program that we will introduce in detail in Section V was used to derive multidimensional models from these data. The most concise picture representing the model output for one scene (i.e., Wanda) is shown in Fig. 2. The 16 processed images of this scene are represented by the crosses. The image in the upper left is the original, the image in the lower left contains no noise (only blur), the image in the upper right contains no blur (only noise), while the image in the lower right contains the maximum amount of blur and noise. The horizontal and vertical axes are not labeled since they do not have any physical meaning. This is a consequence of the fact that Fig. 2 can be arbitrarily translated, rotated, and scaled without influencing its interpretation.

[2]Visual acuities of 1 or 2 on a Landolt chart imply that a subject can detect the opening in a C-ring when this opening is 1 arcmin or 0.5 arcmin of visual angle wide, respectively.

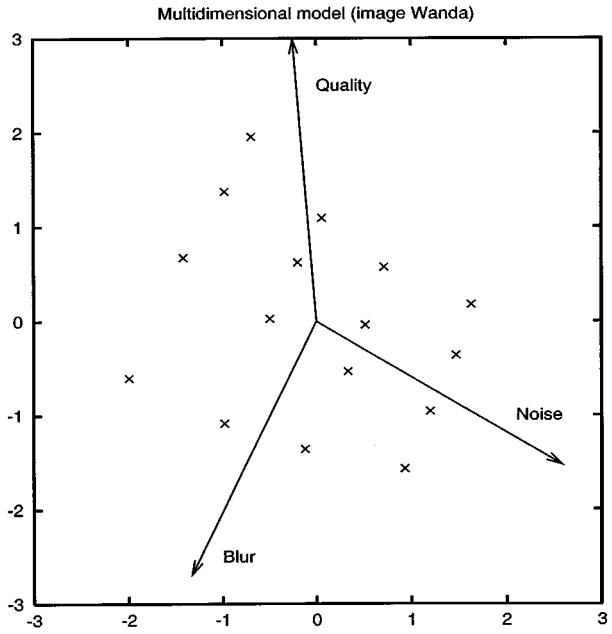


Fig. 2. Example of a multidimensional model for a scene degraded by noise and blur. Crosses indicate the 16 processed images, while the arrows indicate the directions in which perceived blur, noise, and quality increase.

The fact that the 16 images are not arranged in a rectangle, as might be expected *a priori* from the independent processing of blur and noise, indicates that there are some interactions between perceived noise and perceived blur. The directions that correspond to perceived blur and perceived noise are indicated by two of the arrows. The interpretation of the noise axis is, for instance, that the orthogonal projections of two image points onto this axis indicate (on average),

which, of the two images, is perceived to be more noisy. As expected, images with the same noise standard deviation are perceived to be (approximately) equally noisy. There is no such one-to-one relationship between the standard deviation of the filter kernel and perceived blur, however. More specifically, we can derive from the model in Fig. 2 that perceived blur increases with noise for an unblurred image, while it decreases with noise for the most heavily blurred image. The direction corresponding to overall quality is indicated by the third axis. Obviously, since both perceived noise and perceived blur contribute to overall quality, this direction of the quality vector is in between the directions of the attribute vectors for noise and blur. Blur is relatively more important than noise, since the angle between the blur and the quality direction is smaller than the angle between the noise and the quality direction. The model depicted in Fig. 2 assumes that the attribute directions are the same for all subjects. The output of an alternative model, which allows for a different attribute vector per subject, would look similar to Fig. 2, but would contain one vector per attribute and per subject.

A scatter diagram showing the relationship between the distances d_{ij} between stimulus points in Fig. 2 and scored dissimilarities D_{ij} for one subject is shown in Fig. 3. In Fig. 3(a), the dissimilarity data is assumed to be metric, i.e., d_{ij} is compared against D_{ij}. In Fig. 3(b), the dissimilarity data is nonmetric, i.e., d_{ij} is compared against transformed dissimilarities $T(D_{ij})$. In the example, the transformation is assumed to be of the form $T(D) \propto D^q$ and the power q is determined as part of the model optimization (in the specific example, $q = 0.68$). The linear regression coefficients between experimental dissimilarity data and interstimulus distances are $R = 0.982$ and $R = 0.987$ for the metric and nonmetric case, respectively.

III. Psychophysical Measurement of Image-Quality Attributes

In many past experiments, subjects have been asked to judge stimuli, such as images, on subjective attributes. The stimuli are for instance different versions of the same image, obtained by coding the original image at different bit rates with one or more codecs. They have to be evaluated with respect to overall image quality or one of its attributes (e.g., perceived noise or blur) [11]. When conducting such an experiment, one implicitly assumes that subjects are able to discriminate or rate stimuli on the given attribute in a reliable and consistent way. The existence of such a discriminating process [12], [13] is usually postulated and assumed to be an inherent part of the perceptual and/or cognitive abilities of the subject. Repeated intersubject and/or across-subject experiments with the same stimuli should give consistent results if this assumption is true.

Setting up a psychophysical experiment involves decisions on how the stimuli are to be presented to the observer and what the valid observer responses are. Different such experimental paradigms have been proposed, some of which have been standardized [10]. We give a brief overview of some of the most frequently used experimental procedures.

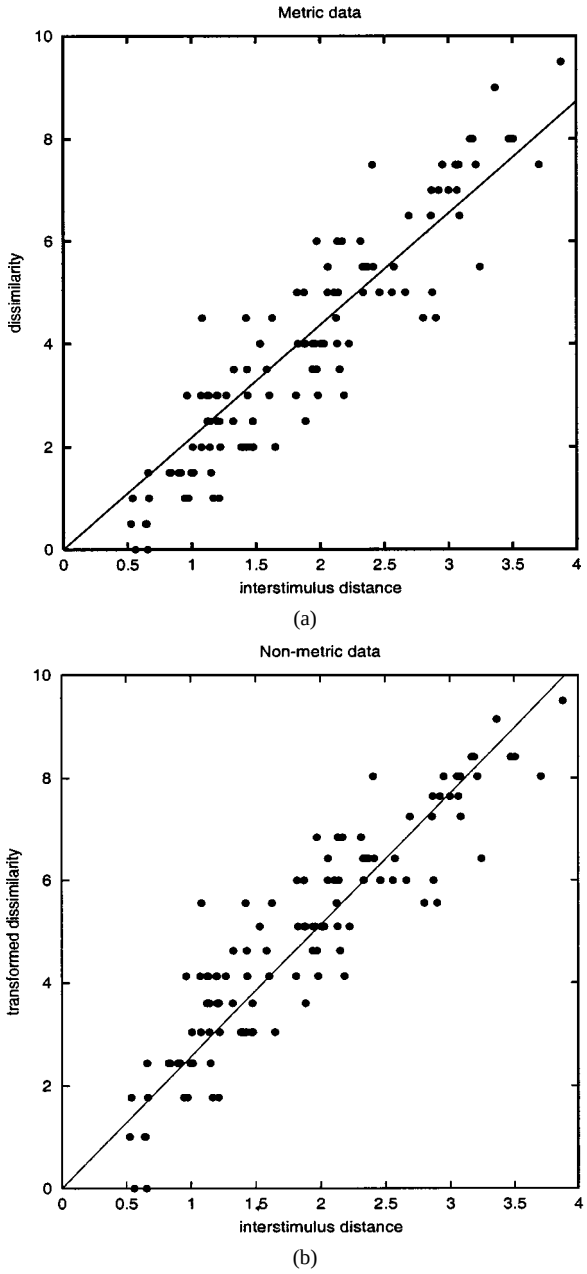


Fig. 3. (a) Original dissimilarity scores and (b) transformed dissimilarity scores as a function of interstimulus distances.

In *single-stimulus scaling*, the test images are presented one by one to the observer. A fixed reference image may be shown together with the test image, either simultaneously in space or sequentially in time. The task of the subject is to rate the test images. In the ITU-R BT.500 recommendation for the subjective assessment of quality or impairment [10], a graphical scale is proposed as a continuous rating device. The scale is divided into five equal-sized nonoverlapping intervals that are denoted by quality or impairment categories. The quality categories are "excellent," "good," "fair," "poor," and "bad," while the impairment categories are "very annoying," "annoying," "slightly annoying," "perceptible, but not annoying," and "imperceptible." The responses of the subjects are mapped into numbers (between zero and 100, for instance) that correspond to the coordinates

of the marks made on the continuous scale. In the experiments performed at our laboratory, we have almost exclusively used an alternative technique, called *numerical category scaling* [11], to rate stimuli. In this case, subjects use integer scores from a limited range, say, from zero to ten. The limited range is adopted because it has been shown that subjects cannot handle large number ranges in a linear way [14], [15]. Using adjectives to denote categories, as in the case of the ITU-R recommendation, potentially also introduces a bias in the use of these categories [16] and is avoided in numerical category scaling.

In *double-stimulus scaling*, pairwise combinations of test images are shown to the observer. The task of the subject is to scale either the dissimilarity between both stimuli in *dissimilarity scaling* or the difference in attribute strength for both stimuli in *difference scaling*. The latter method is also called *preference scaling* if the attribute to be scored is quality.

In dissimilarity scaling, the subjects' task is to indicate how dissimilar or different they perceive two images to be [17]. Any aspect that contributes to the dissimilarity can be taken into account. The order of presentation of the stimuli in the pair should be irrelevant. Continuous scaling or numerical category scaling can be used to express the response to a stimulus pair. Dissimilarity scores are by definition positive or zero.

In difference scaling, subjects can for instance respond to a stimulus combination by means of a numerical category from -5 to $+5$ or using a mark on a continuous interval that is symmetric around the origin. If the first (or leftmost) stimulus has the largest attribute strength, then a negative score is given, while a positive score corresponds to the second (or rightmost) stimulus having the largest attribute strength. Hence, the order of the stimuli does matter in this case. The absolute value of the response is monotonically related to the strength of the attribute difference. The zero category (or origin) can be used in case both stimuli are judged to have equal attribute strength.

In all the above cases, the subject responses can be expressed as real or integer numbers. These numbers correspond to sensations, such as image quality, that are not directly observable. Equal differences between the numbers can not be assumed *a priori* to correspond to equal differences in sensations. In this sense, psychophysical measurements differ fundamentally from many physical measurements [7]. Such numbered responses for which only the order and not the magnitude, is significant are said to belong to an *ordinal* or *nonmetric scale*.

It is usually assumed that there is an unknown monotonically increasing or decreasing nonlinear function that relates the internal sensation strengths to the subject responses [18]. More specifically, let us denote the attribute strengths of the images a used in the experiment by numbers $u(a)$. In difference scaling, for instance, the response $R(a,b)$ to two images a and b is assumed to be a function of the difference in the numbers $u(a)$ and $u(b)$, i.e.,

$$R(a,b) = F[u(a) - u(b)] \tag{2}$$

for some monotonic response function F [18]. This model has a geometric interpretation in one dimension. The images a are represented by their coordinates $u(a)$ in one dimension and the differences between stimulus coordinates are monotonically related to the subject responses. The responses only belong to a *metric scale* if the function F can be assumed to be linear.

Substituting $u^\star(a) = \alpha \cdot u(a) + \beta$ and $F^\star(s) = F(s/\alpha)$ in the above equation confirms that the scale $u(a)$ is only determined up to an arbitrary linear transformation, since

$$R(a,b) = F[u(a) - u(b)] = F^\star[u^\star(a) - u^\star(b)] \tag{3}$$

are two equivalent descriptions. The numbers $u(a)$ are, hence, said to belong to an *interval scale*. An interval scale that has a well defined origin, such as in the case of difference scaling, where zero corresponds to no perceived difference, is called a *ratio scale*.

Obviously, constructing a model from metric data is easier than from nonmetric data partly because many mathematical and statistical techniques [19], such as regression, principal component analysis, analysis of variance (ANOVA), etc., assume metric data. However, techniques for analysing nonmetric data are increasingly available [6], [20]–[22]. The program XGms, which is introduced in this paper, combines several of these techniques and makes them easily accessible through a graphical user interface (GUI).

IV. Multidimensional Modeling of Image Quality

When subjects are requested to evaluate image quality in a psychophysical experiment, they are often able to analyze and justify their judgements (especially if they are somewhat experienced in making image-quality judgements). They can for instance report different kinds of distortions in coded images [23] and are aware of the fact that their overall quality judgement is determined by the relative weight that they attribute to these individual impairments. They are also able to report the sensation strengths for individual impairments in a similar way as they can report their overal quality sensations.

In order to simultaneously model the results from different experiments with the same stimuli, a multidimensional geometrical model is proposed. In such a model, images are represented by points in a multidimensional space. All observations, obtained using one or several of the above-mentioned experimental paradigms, are related to geometrical properties of these points, such as distances between points and coordinates of point projections onto selected axes. This multidimensional model can be viewed as an extension to the 1-D geometrical model described above for modeling judgements on a single attribute.

Based on the above considerations, we divide the task of image-quality modeling into:

1) establishing the stimulus configuration (i.e., the discriminating process) that underlies all attribute judgements by different subjects;
2) determining how this stimulus configuration relates to the judgements for different attributes and/or individuals.

In case of an instrumental measure, the stimulus configuration is supplied by the measure and the remaining task (2) is to determine if this configuration does indeed agree with the subject responses. The XGms program that we will introduce below supports both options, i.e., derivation of a stimulus configuration from experimental data and comparison of a given stimulus configuration against experimental data.

A number of algorithms, usually referred to as MDS programs [2], [3], [24], have been developed within the field of mathematical psychology to derive stimulus configurations from experimental data. Most of these algorithms model dissimilarity data. The stimuli are positioned in space according to metric or nonmetric models. In nonmetric models, the distances between stimulus positions are only monotonically related to the judged dissimilarities, i.e., only the rank order of the experimental data is important [25]–[27]. Metric models, in contrast, maintain a linear relationship between the experimental dissimilarities and the distances between stimulus positions. The Euclidean distance is the metric most frequently used, although more general Minkowski metrics and weighted distances have also been used. In the latter case, the multidimensional psychological space in which the stimuli are positioned is not identical for all subjects (i.e., does not conform strictly to the principle of homogeneity of perception); the spaces for individual subjects are linked by linear (affine) transformations [3]. An example of a freely available MDS program for modeling dissimilarity data that implements most of the above metric and nonmetric options is ALSCAL.[3]

MDS has been used by Marmolin and Nyberg [28] and Goodman and Pearson [29] to study image quality. Dissimilarity judgements for pairs of impaired images were used in both studies. The dimensions of the multidimensional spaces, thus, established were labeled based on an examination of the positions of the impaired images. This labeling was however not verified by separate (independent) experiments. Escalante-Ramírez *et al.* [30] studied the perceptual quality of noise-reduced computer tomography images using MDS techniques. In addition to the space obtained using dissimilarity data, they also obtained a second space through a principal component analysis of the scaling data for the main attributes: noise, blur, and visibility of structures. They, hence, assumed the scaling data to be metric in this case. Both stimulus configurations found in this study were very similar and could be related by a linear transformation. The attribute data could also be used to identify the attribute directions in the multidimensional space that was obtained from dissimilarity data. A similar study was undertaken by Kayargadde and Martens [9] to model images degraded by noise and blur. A problem encountered in both studies is that no program is available for finding a single stimulus configuration based on all available experimental data. Although a program for the joint analysis of direct ratings, pairwise preferences and dissimilarities has been proposed [31], the available implementation[4] does not seem to function properly for the most general case.

Existing MDS programs are mostly offline programs. This implies that the user must enter his/her parameter choices, such as the dimension of the space or the selection between metric or nonmetric analysis and that the results are returned in a file to be examined after the program has finished. This makes it very difficult and cumbersome to appreciate the impact of alternative model choices. Also, since MDS programs are nonlinear optimization programs, the reported solution may correspond to a local optimum instead of a global one. The interactive MDS program XGvis [8] has been developed to help remedy such problems. The program XGvis allows to interactively control the main parameters in the MDS models and exchanges the calculated stimulus configuration with a second program, called XGobi [32]. This latter program is an interactive dynamic data visualization tool for the X Window environment. By combining the functionality of both programs, the user is not only able to dynamically alter the parameters of the MDS model within XGvis, but to also view and manipulate the stimulus configuration in XGobi. This interactivity for instance allows the user to assist the optimization algorithm in avoiding suboptimum solutions that correspond to local minima. The XGvis program, however, only implements a metric model for the dissimilarity data of a single subject and, hence, has too limited functionality to be very useful for modeling image quality. In order to overcome these limitations, XGvis has been extended to include the joint analysis of data from single-stimulus scaling, double-stimulus difference scaling, and dissimilarity scaling for multiple subjects. XGms supports both metric and nonmetric modeling of (parts of) the data.

V. XGMS

In this section, we describe the class of multidimensional models that is implemented in the interactive program XGms. We first describe the optimization criterium or stress function that is used in the program to estimate the model parameters from the experimental data. Many parameters of the multidimensional model (such as the dimension n of the space) can be controlled by the user at runtime, so that their effect on the stimulus configuration and on the relationship between experimental data and model predictions can be explored. Next, we show how optimized monotonic transformations can be used to replace the input data, which may be nonmetric, by transformed data that is approximately metric. Finally, in order to give a better impression of how a user can influence the construction of multidimensional models, the user interface to the program XGms is also described shortly.

A. Optimization Criterium

1) Dissimilarity Data: The experimental dissimilarity data are denoted by $D_{k,i,j}$ for subject $k = 1,\ldots,K_d$ and stimulus pair (i,j) with $i,j = 1,\ldots,N$. The goal is to

[3]http://forrest.psych.unc.edu/research/ALSCAL.html

[4]http://www.psych.mcgill.ca/faculty/ramsay/ramsay.html

construct a stimulus configuration $\mathbf{x}_1, \ldots, \mathbf{x}_N$ such that the experimentally observed dissimilarities are monotonically related to the interstimulus distances

$$d_{ij} = ||\mathbf{x}_i - \mathbf{x}_j||_l = \left[\sum_{m=1}^{n} |x_{im} - x_{jm}|^l \right]^{1/l} \quad (4)$$

where the distance is computed according to a Minkowski metric with power l. The default value $l = 2$ corresponds to a Euclidean distance metric.[5] More precisely, we pursue a linear relationship between transformed dissimilarity scores $TD_{k,i,j} = T_{dk}(D_{k,i,j})$ and interstimulus distances d_{ij}, i.e., $TD_{k,i,j} \approx d_k \cdot d_{ij}$, where d_k is a regression factor. In order to preserve the physical meaning of dissimilarity zero, we require that the applied transformations satisfy $T_{dk}(0) = 0$. An important implicition of the pursued relationship is that the transformed dissimilarities $TD_{k,i,j}$ are assumed to be metric (i.e., have ratio properties), since they are compared to the metric distances d_{ij}. We, hence, assume that there exists a monotonic transformation T_{dk} that maps the nonmetric observed dissimilarities $D_{k,i,j}$ into metric transformed dissimilarities $TD_{k,i,j}$. Note that this assumed transformation is the inverse of a response function that relates internal sensation strengths (on a metric scale) to external responses [as in (2)].

While the stimulus configuration is assumed to be shared by all subjects (according to the principle of homogeneity of perception), the transformations T_{dk} and the regression factors d_k may be subject dependent. For the time being, we assume that the transformations T_{dk} on the experimental data $D_{k,i,j}$ are known. If the input data can be assumed to be metric, no transformations need to be applied, i.e., $TD_{k,i,j} = D_{k,i,j}$, for all subjects $k = 1, \ldots, K_d$ and all stimulus combinations (i, j). This is the default choice of the program XGms at initialization. If the input data is nonmetric, then transformations T_{dk} that map the observed dissimilarities $D_{k,i,j}$ to transformed dissimilarities $TD_{k,i,j}$ that have ratio properties can be derived from the data, as will be discussed later.

The stress [27] for the dissimilarity data is a relative measure that expresses how much the model predictions $d_k \cdot ||\mathbf{x}_i - \mathbf{x}_j||_l$ differ from the (transformed) observations $TD_{k,i,j}$, i.e.,

$$S_d(\mathbf{x}_1, \ldots, \mathbf{x}_N) = \left\{ \frac{1}{K_d} \sum_{k=1}^{K_d} \frac{\sum_{(i,j)\in I_d(k)} |TDE_{k,i,j}|^r}{\sum_{(i,j)\in I_d(k)} |TD_{k,i,j}|^r} \right\}^{1/r} \quad (5)$$

with

$$TDE_{k,i,j} = TD_{k,i,j} - d_k \cdot ||\mathbf{x}_i - \mathbf{x}_j||_l. \quad (6)$$

Note that the sums can be taken over a subset

$$I_d(k) = \{(i,j) \mid i \neq j,\ D_{k,i,j} \neq NA,\ TD_{k,i,j} < U_d(k),\ \ldots\} \quad (7)$$

[5]The use of a Minkowski norm with $l \neq 2$ implies that coordinate differences in horizontal and vertical directions are treated differently from coordinate differences in oblique directions. Such Minkowski norms should, therefore, be handled with care. Amongst others, the optimum stimulus configuration will often depend critically on the initial configuration and convergence problems may occur, especially for $l < 2$ [20], [33].

of all possible stimulus combinations for subject k, so that missing data can be handled (a missing dissimilarity is indicated by NA). We can also select to exclude the largest dissimilarities (maybe because they are judged to be inaccurate) by setting $U_d(k)$ to a value smaller than the maximum transformed dissimilarity value $M_d(k)$ for subject k, where $k = 1, \ldots, K_d$. The number of observations $N_d(k)$ for subject k is, therefore, usually substantially smaller than the N^2 possible stimulus combinations.

Once the stimulus positions $\mathbf{x}_i$, for $i = 1, \ldots, N$, are known, minimization of individual terms in the above expression, i.e.,

$$\min_{d_k} \sum_{(i,j)\in I_d(k)} \left| TD_{k,i,j} - d_k \cdot ||\mathbf{x}_i - \mathbf{x}_j||_l \right|^r \quad (8)$$

can be used to determine the regression factors d_k, for $k = 1, \ldots, K_d$. These factors can subsequently be used to define the *normalized transformed dissimilarities*

$$TD^{\star}_{k,i,j} = \frac{1}{d_k} TD_{k,i,j}. \quad (9)$$

If all subjects behave similarly, then the normalized transformed dissimilarities for identical stimulus pairs (i, j) should have similar values across subjects, i.e., be approximately equal to the interstimulus distance d_{ij}. We will see in the next section how these normalized scores can be used to test the hypothesis that the MDS model can describe the average subject behavior when judging dissimilarity.

Since the transformed dissimilarities are assumed to belong to a ratio scale, i.e., are only determined up to an arbitrary linear scale factor, the stress function should be invariant under a linear scaling $TD_{k,i,j} \rightarrow \delta_k \cdot TD_{k,i,j}$, which can indeed be accomplished by the model parameter change $d_k \rightarrow \delta_k \cdot d_k$. Note that the normalized transformed dissimilarities are invariant under such a transformation.

The scaling of the regression parameter d_k by δ_k could also be replaced by a uniform dilation of the stimulus configuration, i.e., $\mathbf{x}_i \rightarrow \delta_k \cdot \mathbf{x}_i$. In order to avoid such undetermined model behavior, a condition is imposed on the stimulus configuration in order to uniquely determine its scale factor. Similarly, a translation of the stimulus configuration has no influence on the interstimulus distances either, so that an additional condition is imposed to uniquely determine this translation. In summary, the stress is invariant under linear translations and uniform dilation of the stimulus coordinates $\mathbf{x}_i = [x_{im}; m = 1, \ldots, n]$, i.e., mappings of the form

$$x_{im} \rightarrow d \cdot x_{im} + e_m \quad (10)$$

for $i = 1, \ldots, N$ and $m = 1, \ldots, n$. *A priori* conditions on the stimulus configuration are therefore required in order to guarantee a unique stimulus configuration. The translation vector $\mathbf{e} = [e_m; m = 1, \ldots, n]$ is determined by requiring that the configuration is centered at the origin, i.e., $\sum_{i=1}^{N} x_{im} = 0$, for $m = 1, \ldots, n$. The dilation factor d is derived from the condition that $\sum_{i=1}^{N} \sum_{m=1}^{n} x_{im}^2$ is constant (in our case, equal to $N \cdot n$). Stimulus configurations that satisfy these conditions are called normalized. Because of this

normalization of the stimulus configuration, the number of degrees of freedom (DOFs) in this configuration is equal to

$$F_{\mathbf{x}} = (N-1)\cdot n - 1. \tag{11}$$

If the distance metric is Euclidean, i.e., if $l = 2$, then an arbitrary orthogonal transformation (rotation or mirroring around the origin) of the stimulus configuration has no effect on the interstimulus distances either. This further reduces the DOF in the stimulus configuration to

$$\begin{aligned} F_{\mathbf{x}|l=2} =& (N-1)\cdot n - 1 - \frac{n(n-1)}{2} \\ =& \left(N - \frac{n+1}{2}\right)\cdot n - 1. \end{aligned} \tag{12}$$

In the default setting of the XGms program, no measures are taken to uniquely select the orientation of the stimulus configuration. This implies that the orientation of the output stimulus configuration will depend on the initial stimulus configuration. The default settings of the XGms program can, however, be modified such that an orthogonal transformation is selected that alligns the stimulus configuration along its principal axes [20].

2) Preference Data: The stress term for the double-stimulus preference (or, more generally, attribute difference) data can be defined in a similar way as

$$S_p(\mathbf{x}_1,\ldots,\mathbf{x}_N) = \left\{ \frac{1}{K_p} \sum_{k=1}^{K_p} \frac{\sum_{(i,j)\in I_p(k)} |TPE_{k,i,j}|^r}{\sum_{(i,j)\in I_p(k)} |TP_{k,i,j}|^r} \right\}^{1/r} \tag{13}$$

where $TP_{k,i,j} = T_{pk}(P_{k,i,j})$ denotes the transformed preference rating by subject k, with $k = 1,\ldots,K_p$, for stimulus pair (i,j), with $i,j = 1,\ldots,N$ and

$$TPE_{k,i,j} = TP_{k,i,j} - m_k \cdot ([\mathbf{p}_k,\mathbf{x}_i] - [\mathbf{p}_k,\mathbf{x}_j]) \tag{14}$$

is the corresponding prediction error. It is again assumed that the mapping T_{pk} is monotonically increasing and satisfies $T_{pk}(0) = 0$. The transformed preference $TP_{k,i,j}$, which is assumed to belong to a ratio scale, is compared against a prediction $[\mathbf{p}_k,\mathbf{x}_i]-[\mathbf{p}_k,\mathbf{x}_j]$, which is derived from the stimulus positions $\mathbf{x}_i$ and $\mathbf{x}_j$ and the *preference vector* $\mathbf{p}_k$ for subject k. Again, a subset

$$\begin{aligned} I_p(k) =& \{(i,j) | i \neq j, P_{k,i,j} \neq NA, \\ & |TP_{k,i,j}| < U_p(k), \ldots\} \end{aligned} \tag{15}$$

of all possible stimulus combinations can be selected. While dissimilarities are always positive numbers, preferences can be both negative or positive, depending on whether the ith stimulus is preferred over the jth stimulus or vice versa. Therefore, the upper limit $U_p(k)$ works on the absolute value of the (transformed) preferences. It is smaller than the maximum amplitude $M_p(k)$ of the transformed preferences of subject k, where $k = 1,\ldots,K_p$. The number of observed preferences for subject k is denoted by $N_p(k)$.

XGms allows to choose between two possible prediction models. The user can (interactively) switch between both models in order to decide which model best describes the data.

According to the *vector-product* or *inner-product* model, the prediction is equal to the vector product

$$\begin{aligned} [\mathbf{p}_k,\mathbf{x}_i] - [\mathbf{p}_k,\mathbf{x}_j] =& \langle \mathbf{p}_k, \mathbf{x}_i - \mathbf{x}_j \rangle \\ =& \sum_{m=1}^{n} p_{km} \cdot (x_{im} - x_{jm}) \end{aligned} \tag{16}$$

between the difference vector $\mathbf{x}_i - \mathbf{x}_j$, pointing from stimulus j to stimulus i, and the preference vector $\mathbf{p}_k$. The corresponding geometrical interpretation is as follows. The difference between the orthogonal projections (or coordinates) of the stimulus positions $\mathbf{x}_i$ and $\mathbf{x}_j$ onto a 1-D axis, with direction specified by the vector $\mathbf{p}_k$, determines the average preference between both stimuli for subject k. If the same stimuli are scored very differently by different subjects, then this should be reflected in preference vectors $\mathbf{p}_k$ with distinct orientations. Such different orientations are obviously only possible in case the dimension n of the space exceeds one [34].

A consequence of the ratio property of the transformed preferences is that only the directions of the preference vectors $\mathbf{p}_k$ are uniquely determined; their amplitudes $||\mathbf{p}_k||$ may be scaled arbitrarily. We can, hence, put $||\mathbf{p}_k|| = 1$ and conclude that the vector-product model contains n DOFs for each preference k, i.e., $n-1$ DOFs for the direction of the vector $\mathbf{p}_k$ and one DOF for the correlation factor m_k.

The alternative *ideal-point* or *unfolding model*

$$\begin{aligned} &[\mathbf{p}_k,\mathbf{x}_i] - [\mathbf{p}_k,\mathbf{x}_j] \\ &\qquad = -\log\left(||\mathbf{x}_i - \mathbf{p}_k||_l\right) + \log\left(||\mathbf{x}_j - \mathbf{p}_k||_l\right) \end{aligned} \tag{17}$$

is based on the ratio of the distances of the stimulus positions $\mathbf{x}_i$ and $\mathbf{x}_j$ from the "ideal point" $\mathbf{p}_k$ [3], [31]. This ideal point cannot coincide with a stimulus point since $[\mathbf{p}_k,\mathbf{x}_i]$ must remain finite.[6] If the same stimuli are scored very differently by different subjects, then this should be reflected by ideal points $\mathbf{p}_k$ at some distance apart. Unlike the vector-product model, the ideal-point model can also model different subject behaviors in case the space is 1-D. The model contains $n+1$ model parameters for each preference k, i.e., n DOFs for the position of the ideal point $\mathbf{p}_k$ and one DOF for the correlation factor m_k.

The predictions from both models can be related in case of a Euclidean metric with $l = 2$. More specifically, if $||\mathbf{p}_k||_2 \gg ||\mathbf{x}_i||_2, ||\mathbf{x}_j||_2$, then the first-order Taylor series expansion

$$-\log\left(||\mathbf{x}_i - \mathbf{p}_k||_2\right) \approx -\log\left(||\mathbf{p}_k||_2\right) + \frac{1}{||\mathbf{p}_k||_2^2} \langle \mathbf{p}_k, \mathbf{x}_i \rangle \tag{18}$$

[6]This can be avoided in XGms by using a slightly modified prediction formula with $\log(||\mathbf{x}_i - \mathbf{p}_k||_l)$ replaced by $\log(r_o + ||\mathbf{x}_i - \mathbf{p}_k||_l)$, where r_o is a small offset, such as $r_o = 10^{-3}$.

can be used to derive the following approximation:

$$-\log\left(\frac{||\mathbf{x}_i-\mathbf{p}_k||_2}{||\mathbf{x}_j-\mathbf{p}_k||_2}\right)\approx\frac{1}{||\mathbf{p}_k||_2^2}\langle\mathbf{p}_k,\mathbf{x}_i-\mathbf{x}_j\rangle. \quad (19)$$

Hence, if the ideal-point $\mathbf{p}_k$ is far removed from the stimulus positions, then only the direction of the vector $\mathbf{p}_k$ is relevant for the predictions and the ideal-point model becomes equivalent to a vector-product model. In practice, this will imply that estimation of the amplitude $||\mathbf{p}_k||_2$ becomes ill conditioned in case of the ideal-point model and that a vector-product model, with one less DOF, should be preferred.

Very often, the available preference data may be divided into groups. For example, the indexes $k=1,\ldots,K_p/2$ may refer to $K_p/2$ different subjects rating preference (differences in image quality), while indexes $k=K_p/2+1,\ldots,K_p$ may refer to the same subjects rating another attribute difference (such as the difference in amount of perceived blur). In such a case, it often makes sense to look for a common prediction model for all subjects within a group. This corresponds to estimating a single prediction vector $\mathbf{p}_k=\mathbf{p}$ for all indexes k in a group. In case of a single group with K_p subjects, this reduces the number of parameters from K_pn to K_p+n-1 for the vector-product model and from $K_p(n+1)$ to K_p+n for the ideal-point model. XGms allows to define such groups and can also export the *normalized transformed preferences*

$$TP^{\star}_{k,i,j}=\frac{1}{m_k}TP_{k,i,j} \quad (20)$$

that allow for an easy comparison of experimental data across subjects within a group. Indeed, for stimulus pair (i,j), the repeated judgements $TP^{\star}_{k,i,j}$ across subjects k in a group should aggregate around the "group" prediction $[\mathbf{p},\mathbf{x}_i]-[\mathbf{p},\mathbf{x}_j]$.

3) Attribute Data: The stress term for the single-stimulus attribute scaling data is defined as

$$S_a(\mathbf{x}_1,\ldots,\mathbf{x}_N)=\left\{\frac{1}{K_a}\sum_{k=1}^{K_a}\frac{\sum_{(i,j)\in I_a(k)}|TAE_{k,i,j}|^r}{\sum_{(i,j)\in I_a(k)}|TA_{k,i,j}-TA_{k,\cdot\cdot}|^r}\right\}^{1/r} \quad (21)$$

where $TA_{k,\cdot\cdot}$ denotes the average transformed attribute score for subject k, $A_{k,i,j}$ is the attribute rating given by subject k for stimulus i on the jth repetition ($k=1,\ldots,K_a$, $i=1,\ldots,N$, and $j=1,\ldots,L_a$), and

$$TAE_{k,i,j}=TA_{k,i,j}-(c_k+s_k\cdot[\mathbf{a}_k,\mathbf{x}_i]) \quad (22)$$

is the corresponding prediction error. The transformed attribute scores $TA_{k,i,j}=T_{ak}(A_{k,i,j})$ are compared against their predictions $[\mathbf{a}_k,\mathbf{x}_i]$. These predictions for subject k are derived from the stimulus positions $\mathbf{x}_i$ and the *attribute vector* $\mathbf{a}_k$. The same two prediction models as in the case of preference data are available for modeling attribute data in XGms.

The vector-product model compares the transformed attribute scores $TA_{k,i,j}$ with the linear prediction

$$\begin{aligned}c_k+f_k\cdot[\mathbf{a}_k,\mathbf{x}_i]&=c_k+f_k\cdot\langle\mathbf{a}_k,\mathbf{x}_i\rangle\\&=c_k+f_k\cdot\sum_{m=1}^{n}a_{km}\cdot x_{im}.\end{aligned} \quad (23)$$

The average strength of attribute k for stimulus i, hence, increases linearly with the coordinate of the stimulus projection on a 1-D axis with direction indicated by the attribute vector $\mathbf{a}_k$. The offset value c_k, the scale factor f_k, and the direction of the attribute vector $\mathbf{a}_k$ comprise a total of $n+1$ parameters. We again adopt the convention that $||\mathbf{a}_k||=1$ in this case.

The alternative to this vector-product model is the ideal-point (or unfolding) modelideal-point modelunfolding model in which the transformed attribute scores for stimulus i are related to the distance $||\mathbf{x}_i-\mathbf{a}_k||_l$ of the stimulus at position $\mathbf{x}_i$ from an "ideal" image at position $\mathbf{a}_k$, i.e.,

$$c_k+f_k\cdot[\mathbf{a}_k,\mathbf{x}_i]=c_k+f_k\cdot\{\alpha_k-\log(||\mathbf{x}_i-\mathbf{a}_k||_l)\} \quad (24)$$

where the offset

$$\alpha_k=\frac{1}{N}\sum_{i=1}^{N}\log(||\mathbf{x}_i-\mathbf{a}_k||_l) \quad (25)$$

is included to obtain that the attribute predictions are centered around the origin. The offset value c_k, the scale factor f_k, and the position $\mathbf{a}_k$ are the $n+2$ parameters of the ideal-point model.

Again, a subset

$$I_a(k)=\{(i,j)\mid A_{k,i,j}\neq NA,\ |TA_{k,i,j}-B_a(k)|<U_a(k),\ \ldots\} \quad (26)$$

off all possible attribute scores can be used in the stress function. The transformed attribute values for subject k belong to the interval $[m_a(k),M_a(k)]$, for $k=1,\ldots,K_a$. The upper limit $U_a(k)$ and the base value $B_a(k)$ can be used to indicate that only the transformed attribute scores in the range $[B_a(k)-U_a(k),B_a(k)+U_a(k)]$ contribute to the stress. The number of observed attribute scores for subject k is denoted by $N_a(k)$.

The zero point for dissimilarity data and preference data has a physical meaning (i.e., no dissimilarity or preference) and is, hence, uniquely determined. We could, therefore, assume that the transformed dissimilarities $TD_{k,i,j}$ and preferences $TP_{k,i,j}$ belonged to ratio scales, provided of course that the applied transformations satisfied $T_{dk}(0)=T_{pk}(0)=0$. Attribute data usually do not have such a natural origin, so that the transformed attribute scores $TA_{k,i,j}$ are assumed to belong to an interval scale, i.e., they are only determined up to an arbitrary linear transformation. The stress is invariant under such a linear transformation $TA_{k,i,j}\rightarrow\delta_k\cdot TA_{k,i,j}+\epsilon_k$ of the transformed attribute data, since such a transformation can be absorbed in the

regression parameters by mapping $c_k \rightarrow \delta_k \cdot c_k + \epsilon_k$ and $f_k \rightarrow \delta_k \cdot f_k$.

As in the case of preference data, the attribute data may be subdivided into groups. This corresponds to estimating a single prediction vector $\mathbf{a}_k = \mathbf{a}$ for all indexes k in a group. In case of a single group with K_a subjects, this reduces the number of parameters from $K_a(n+1)$ to $2K_a+n-1$ for the vector-product model and from $K_a(n+2)$ to $2K_a+n$ for the ideal-point model. The *normalized transformed attributes*

$$TA^{\star}_{k,i,j} = \frac{1}{f_k}\left(TA_{k,i,j} - c_k\right) \tag{27}$$

across subjects k and repetitions j in a group are expected to aggregate around the "group" prediction $[\mathbf{a}, \mathbf{x}_i]$ for stimulus i.

B. Interactively Controlled Power-Like Transformations

The monotonically increasing transformations from $D_{k,i,j}$, $P_{k,i,j}$, and $A_{k,i,j}$ to $TD_{k,i,j}$, $TP_{k,i,j}$, and $TA_{k,i,j}$ are fixed in the above discussion. These transformations can, however, be modified interactively in XGms. In order to limit the possible variations, the available nonlinear transformations are based on a power-like function

$$P\left(x; q, x_t\right) = \operatorname{sign}(x) \cdot \frac{\left(|x| + x_t\right)^q - x_t^q}{q} \tag{28}$$

that can be characterized by two parameters q and x_t. This function varies from zero for values $|x| \ll x_t$ well below the threshold to a power function with exponent q for values $|x| \gg x_t$ well above the threshold. Note that the often-used power-law relationship [20]

$$P(x; q, 0) = \operatorname{sign}(x) \cdot \frac{|x|^q}{q} \tag{29}$$

is included as a special case. The above definition needs to be modified to a logarithmic relationship

$$P\left(x; 0, x_t\right) = \operatorname{sign}(x) \cdot \left[\log\left(|x| + x_t\right) - \log x_t\right] \tag{30}$$

in case $q = 0$. The continuity in this transition from a power law to a logarithmic relationship is most easily observed by verifying that the derivative to the funcion is

$$P'\left(x; q, x_t\right) = \left(|x| + x_t\right)^{q-1} \tag{31}$$

in all cases. Note that this derivative is strictly positive, so that $P(x; q, x_t)$ is indeed monotonically increasing.

The nonlinear transformations that are used in the XGms program are designed such that they preserve the range of the input data. It is easily verified that, if the input data x is within the range $[m, M]$, that

$$T\left(x; q, x_t, b\right) = \frac{(M-m) \cdot P(x-b) - M \cdot P(m-b) + m \cdot P(M-b)}{P(M-b) - P(m-b)} \tag{32}$$

where $b \in [m, M]$ is the origin of the power function and $P(x) = P(x; q, x_t)$, is also limited to this same range. The linear relationship

$$T\left(x; 1, x_t, b\right) = x \tag{33}$$

is used as the default transformation by the program XGms. It corresponds to assuming that the input data is metric.

In case of dissimilarity and preference data, we choose $m = -M$, where M is the maximum (absolute) score and $b = 0$, since the zero value is the natural origin for the transformation. The resulting simplified transformation

$$T\left(x; q, x_t\right) = \frac{M \cdot P\left(x; q, x_t\right)}{P\left(M; q, x_t\right)} \tag{34}$$

satisfies $T(x; q, x_t) = 0$ and $T(M; q, x_t) = M$ and contains two parameters, i.e., q and x_t. In XGms, the threshold value $x_t = t \cdot M$ is specified by t as a fraction of M.

In the case of attribute data, there is no natural origin for the transformation, so that the bias b is also a parameter. The transformed bias value is denoted by $B = T(b; q, x_t, b)$. In this case, the threshold $x_t = t \cdot (M - m)$ is specified by t as a fraction of the overall range.

C. Nonmetric MDS Through Optimized Transformations

The multidimensional model described above is essentially metric, since the nonlinear transformations on the experimental data are assumed to be known *a priori*. Rather than fitting the model to the experimentally observed dissimilarities $D_{k,i,j}$, preferences $P_{k,i,j}$ and attribute scores $A_{k,i,j}$, the fit is made to the transformed values $TD_{k,i,j}$, $TP_{k,i,j}$, and $TA_{k,i,j}$ that are assumed to have ratio or interval properties. In nonmetric modeling, the monotonic transformations on the data are not specified *a priori*, but are determined as part of the optimization process. The XGms program allows for such nonmetric optimizations under user control. The user can select to replace some or all of the experimental data by optimally transformed data in the above minimization of the stress.

The stress functions are composed of expressions of the form

$$S^r = \frac{\sum_{i=1}^{I} \left|T(o_i) - \hat{o}_i\right|^r}{\sum_{i=1}^{I} \left|T(o_i)\right|^r} \tag{35}$$

where $T(o_i)$ denotes the transformed observation o_i for case i and $\hat{o}_i$ is the corresponding prediction according to the stimulus configuration and regression parameters in the current model. These predictions $\hat{o}_i$ are, for instance, linearly related to the distances between the stimulus positions in case of dissimilarity and to the coordinates or coordinate differences of the stimulus positions along known attribute or preference directions.

In nonmetric MDS [3], [26], the monotonic transformation T that minimizes S for known predictions $\hat{o}_i$ is selected as the transformation for the experimental data. In XGms, the monotonic transformations are such that they preserve the extreme values. The extreme values are equal to zero and the

maximum (absolute) value in case of dissimilarity and preference data and equal to the minimum and maximum value in case of attribute data. In the case of preference data, the optimum transformation is also restricted to be asymmetric with respect to the origin, i.e., $T(-o) = -T(o)$. Amongst others, this implies that $T(0) = 0$.

1) Optimum Power-Like Transformation: The first possibility is to assume that the optimum transformation can be closely approximated by (32). A nonlinear optimization over the parameters q and t (b in case of attribute scores) can then be performed in XGms. The parameters are restricted to the range $q \in [-6, 6]$ and $t \in [0.02, 4]$ in order to avoid runaway arguments in the optimization.[7] In case of attribute scores, the origin b for the transformation is limited to the range $[m, M]$ of scores that occur. The number of DOFs in an optimized power-like transformation is, hence, $F_t = 2$ or $F_t = 3$, depending on whether the transformed data is ratio (dissimilarity and preference data) or interval (attribute data), respectively.

Although this optimization over a restricted set of power-like transformations often gives reasonable results, it may be interesting to compare the obtained optimum power-like transformation with the optimum transformation out of all possible monotonic transformations. For example, in case of attributes, knowledge of the optimum monotonic transformation often helps to select an appropriate bias parameter b for the power-like transformations (it is usually advantageous to put this bias b at the attribute score for which this optimum transformation is most asymmetric). Therefore, XGms also allows for an optimization across all possible monotonic transformations.

2) Optimum Monotonic (Kruskal) Transformation: An algorithm for determining the optimum monotonic transformation in nonmetric MDS was originally developed by Kruskal [3], [26]. We use an alternative and more flexible method based on spline interpolation introduced by Ramsay [35].

Suppose that the observations $\{o_i; i = 1, \ldots, I\}$ contain $J \leq I$ distinct values. Without loss of generality, we can assume that the data are sorted such that the first J observations $\{o_j; j = 1, \ldots, J\}$ are all distinct and in increasing order. Such sorting has no effect on the value of the stress. A monotonically increasing transformation function

$$T(o) = T(o_1) + \int_{o_1}^{o} T'(x)dx \tag{36}$$

can be obtained by integrating a positive-valued derivative function $T'(o)$. We take a derivative function $T'(o)$ that linearly interpolates between J positive values $T'(o_j)$, for $j = 1, \ldots, J$. The corresponding integrated function will be piecewise quadratic with a continuous derivative, as shown in the example of Fig. 4. A nonlinear optimization of the stress in (35) as a function of $T'(o_j)$, for $j = 1, \ldots, J$, can, hence, be to used to find the optimum monotonically increasing transformation on the data. The numbers $T'(o_j)$ have to be rescaled at each iteration step to guarantee that the transformation T preserves the data range $[o_1, o_J]$, i.e., the normalization conditions

$$T(o_J) - T(o_1) = \int_{o_1}^{o_J} T'(x)dx = o_J - o_1 \tag{37}$$

and $T(o_1) = o_1$ are imposed. Since only the $J - 2$ intermediate values are modified from o_j to $T(o_j)$, for $j = 2, \ldots, J-1$, we obtain that the number of DOFs in the above *Kruskal optimization* is $F_t = J - 2$.

3) Optimum Spline Transformation: The Kruskal approach can be easily generalized to spline transformations

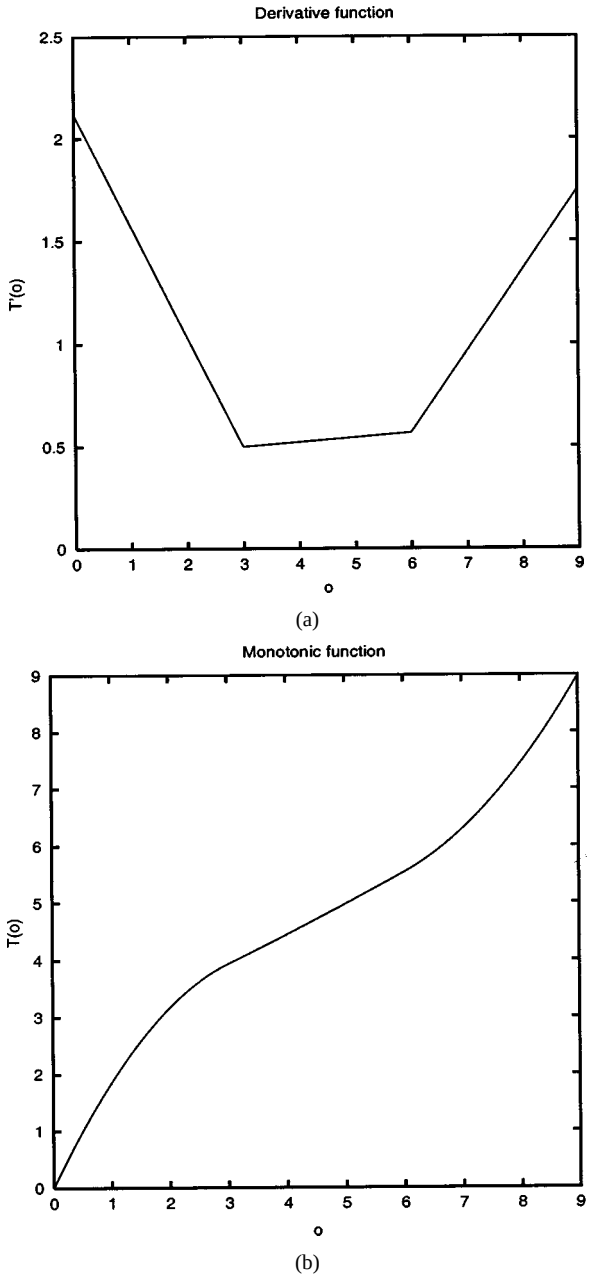


Fig. 4. (a) Piecewise-linear derivative function $T'(o)$ that is strictly positive. (b) Corresponding monotonically increasing function $T(o)$ that is piecewise quadratic with a continuous derivative.

[7]Nonlinear optimizations for bounded parameters are performed using the DMNFB routine from the Netlib library at http://www.netlib.org.

in which case the derivatives $T'(\tilde{o}_j)$ are only specified for a limited number of knot points $\tilde{o}_j$, for $j = 1, \ldots, s+2$ (with $s+2 < J$). The first and last knot point are chosen equal to the minimum and maximum value, i.e., $\tilde{o}_1 = o_1$ and $\tilde{o}_{s+2} = o_J$, respectively, and the number of internal knot points is s. All required derivatives $T'(o)$ are obtained by linear interpolation of the values at the knot points, while the monotonic transformation $T(o)$ is again obtained by integrating the normalized derivative function. XGms allows for *spline optimization* with $0 \leq s < J-2$ equally spaced internal knot points. Especially when J is large, such a spline optimization of a reduced order is a practical alternative to the general Kruskal optimization. The number of DOFs in the optimized monotonic transformation is $F_t = s+1$ in case of a spline optimization with s internal knot points. One DOF is added by the integration, but two DOFs are consumed by the normalization conditions that guarantee that the range of the data is preserved.

D. Optimization of the Stress Function

The XGms program minimizes an overall stress function of the form

$$\mathbf{Stress}(\mathbf{x}_1, \ldots, \mathbf{x}_N) = \left\{ \frac{|w_d S_d|^r + |w_p S_p|^r + |w_a S_a|^r}{|w_d|^r + |w_p|^r + |w_a|^r} \right\}^{1/r} \tag{38}$$

as a function of the stimulus positions $\mathbf{x}_1, \ldots, \mathbf{x}_N$, the regression parameters d_k (for dissimilarity), m_k and $\mathbf{p}_k$ (for preference), and c_k, f_k, and $\mathbf{a}_k$ (for attribute scaling). In case of nonmetric MDS, the monotonic transformations T_{dk}, T_{pk}, and T_{ak} must also be optimized per subject. The weights are initialized to $w_d = w_p = w_a$, but can be used to vary the relative contribution of the dissimilarity, preference, and attribute data to the overall stress.

All required optimizations are performed iteratively.[8] If the stimulus configuration needs to be optimized (as in MDS), then one iteration step involves three stages. If the stimulus configuration is fixed (as in regression analysis), then one iteration step involves only two stages.

In the first stage, which is only performed in case of MDS, the stimulus positions are optimized, assuming fixed values for the regression parameters and the monotonic transformations. The stimulus configuration is normalized after each optimization step, i.e., its translation and dilation factor are determined by *a priori* conditions (see above).

In the second stage, the regression parameters are optimized for a fixed stimulus configuration and known monotonic transformations on the data. This latter optimization involves only individual terms in the stress function and is, therefore, fairly simple and efficient. In case $r = 2$, it reduces to solving a set of (simultaneous) linear equations. If $r \neq 2$, then a (slower) nonlinear optimization is required.

[8]Nonlinear optimizations in XGms are performed using the iterative routine DMNF from the Netlib library.

Nonmetric models add a third stage to each iteration step in the minimization of the stress criterium. In this third stage, the monotonic transformations that minimize the stress for the current stimulus configuration and regression parameters are updated.

E. Graphical User Interface to XGms

The GUI to XGms is depicted in Fig. 5 for an experimental data set containing both dissimilarity data and attribute data. This interface was derived from the GUI of the XGvis program [8].

The panel at the top contains the major action buttons. The user can either specify an initial stimulus configuration at startup or can load a new stimulus configuration (using "Load CONF") at any time during the XGms session. He/she can also switch to a random initial stimulus configuration using the "Scramble CONF" button and can return at any time to the last specified initial configuration using "Init CONF." The current stimulus configuration can be kept fixed and used for regression analysis against the experimental data. A single iteration step in the regression can be triggered with the "Step REG" button, while multiple iterations can be started and stopped with the "Run REG" button. Alternatively, the current stimulus configuration can be used as the initial configuration in an MDS analysis. Starting and stopping of the iterations in such an MDS analysis is controlled by the "Step MDS" and "Run MDS" buttons.

Either the original data (Di and Ai), power-transformed data (Dt and At), or spline-transformed data (Ds and As) can be used in the regression or MDS analysis. Dissimilarities and/or attribute scores from individual subjects can be selected for transformation or the scores from all subjects can be transformed simultaneously (Di-current, Dt-current, or Ds-current versus Di-all, Dt-all, or Ds-all). In the example, the original data are used for the attributes, while the dissimilarity data are transformed using a power-like function. In order to allow processing by other data analysis or visualization programs, most intermediate data in the XGms program can also be output to ASCII files using the options in the "file" menu. Amongst others, the normalized transformed scores $TD^{\star}_{k,i,j}$, $TP^{\star}_{k,i,j}$, and/or $TA^{\star}_{k,i,j}$ and their model predictions $||\mathbf{x}_i - \mathbf{x}_j||_l$, $[\mathbf{p}_k, \mathbf{x}_i] - [\mathbf{p}_k, \mathbf{x}_j]$, and $[\mathbf{a}_k, \mathbf{x}_i]$ can all be exported in this way.

The bottom left panel allows to control the parameters of the multidimensional model, such as:

1) the dimension n of the stimulus space;
2) the power r used in the stress function;
3) the Minkowski power l of the distance metric;
4) the number of iteration steps in one call to the optimization routines;
5) the weights w_d, w_p and/or w_a on the different stress terms;
6) the prediction model used (either vector product or ideal point);
7) the parameters q and t (and b in case of attributes) that control the nonlinear power-like transformations on the individual subsets of the data;

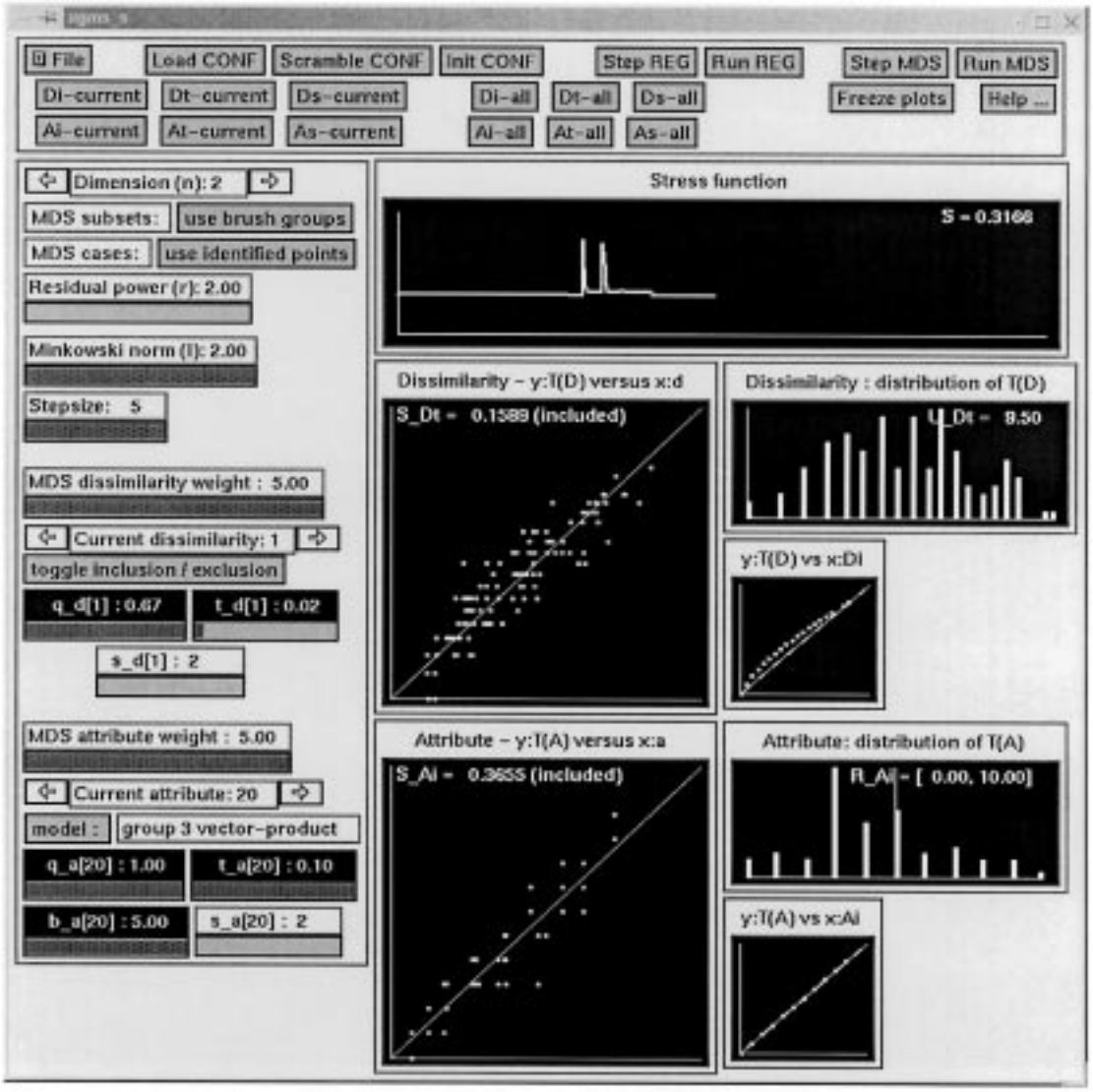


Fig. 5. XGms GUI.

8) the number s of internal knot points in a spline transformation.[9]

If data for several subjects and/or attributes are available, then the "current" selection for the dissimilarity or attribute allows to view the settings and model fits for the current subset of the data. The user can, for instance, select which parameters of the power-like transformation are fixed to a user-specified value and which can be optimized by XGms. In the case of preference and attribute data, a choice can be made between a vector-product model and an unfolding model. Attributes and preferences can also be grouped so that they share a common prediction vector. It is, moreover, possible to (temporarily) exclude the subset of the data currently being viewed from the stress criterium. Another way of selecting a subset of the available data is through the "use brush groups" and "use identified points" options [8]. The XGobi program that is used to visualize the stimulus configuration (see below) allows to partition the stimuli into nonoverlapping subsets by means of brushing.[10] In case "use brush groups" is selected, then only those dissimilarities and preferences for which both stimuli belong to the same subgroup are used in the regression and/or MDS analysis. Similarly, XGobi also allows to select a subset of the stimulus points. In case "use identified stimuli" is selected, then only those data that involve the identified stimuli are used in the analysis. Both options can also be active at the same time.

The graph at the top in the bottom right panel of Fig. 5 shows how the stress has varied as a function of time during the XGms session. In addition, scatterplots of transformed experimental data versus predicted model data and barplots representing the histograms of the transformed experimental data are shown for the "current" dissimilarity and attribute. The small panels depict the overall transformations from the input data to the transformed data. The limits $U_d(k)$, $U_p(k)$, or $[B_a(k) - U_a(k), B_a(k) + U_a(k)]$ used to define a subset

[9]Kruskal transformation is considered as a special case of spline transformation, indicated by s_d or s_a equal to max.

[10]Brushing means that the attributes (such as color, shape, etc.) of the stimulus points can be changed. Stimulus points with the same attributes are assumed to belong to the same brush group.

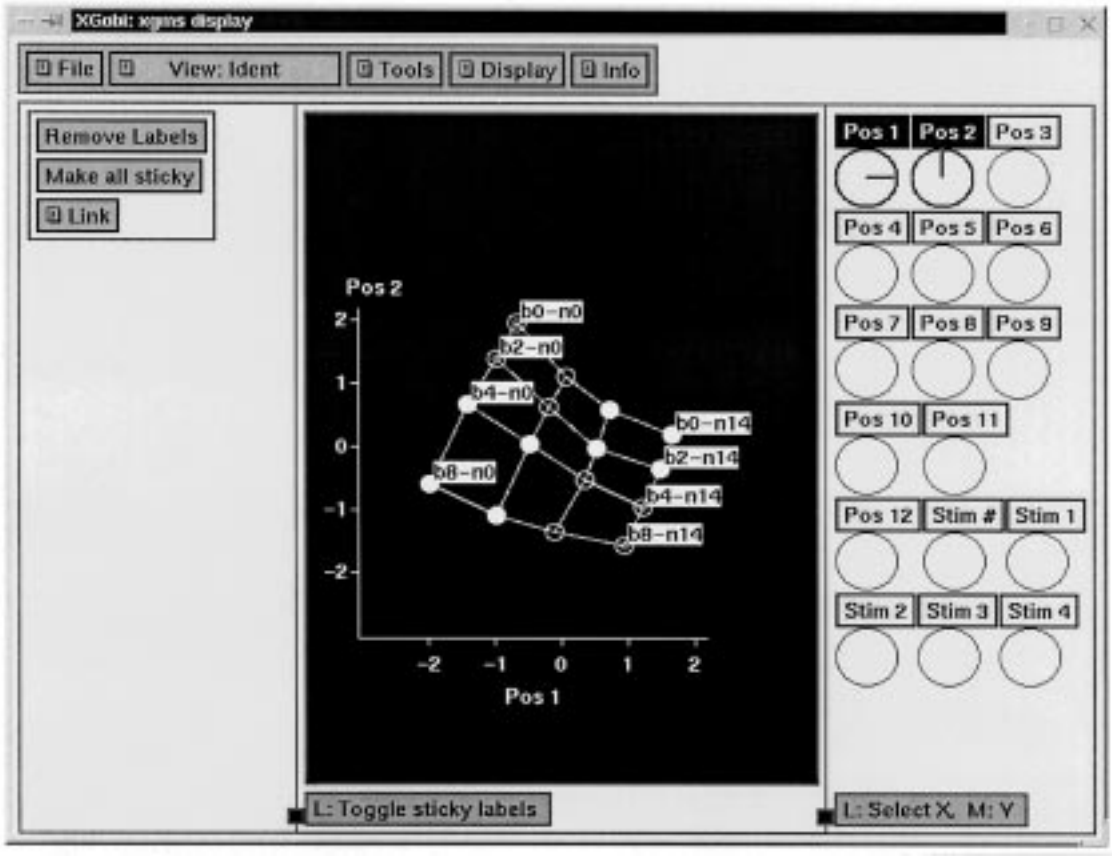


Fig. 6. XGobi GUI.

of the "current" data can be changed by clicking at the desired position in the histogram of the transformed data. The data in the scatterplots can also be output to the visualization program XGobi in order to examine them more closely.

As mentioned before, the stimulus configuration is exchanged between the MDS program XGms and the visualization program XGobi. The GUI to the XGobi program is shown in Fig. 6. The two-dimensional (2-D) stimulus configuration in the example is displayed as a point plot and the user can view and/or alter this configuration. In Fig. 6, two brush groups with eight stimuli each have been created using open and filled circles, respectively, while eight of the 16 available stimuli have been identified. The preference and attribute vectors (or ideal points) can also be visualized in XGobi. Higher dimensional stimulus configurations can be viewed dynamically by moving a 2-D projection plane through the stimulus space [32].

VI. Image Quality Data

In this section, we apply multidimensional modeling to two experimental data sets.

A. Images With Blur and Noise

The first data set that we consider concerns images degraded by noise and blur [9]. The experimental setup used for collecting these data was described in Section II. The measurements performed per scene with the 16 different images were as follows: 1) double-stimulus dissimilarity scaling by five subjects and 2) single-stimulus attribute scaling of perceived noisiness, blur, and quality by seven subjects. The 16 images of a scene corresponded to all possible combinations of four levels of blur and four levels of Gaussian additive white noise. We used XGms with an inner-product prediction model for the attribute scores to find separate 2-D stimulus configurations for each of the scenes. The subjects were assumed to form a homogeneous group, so that a single attribute vector was used to describe the mapping from stimulus configuration to attribute strengths for all subjects. A nonmetric analysis showed that the dissimilarity and attribute scores were approximately metric, so that no monotonic transformation was performed on the data in the following analysis (i.e., $TD_{k,i,j} = D_{k,i,j}$ and $TA_{k,i,j} = A_{k,i,j}$). The resulting 2-D model for the image Wanda was already illustrated in Fig. 2. The corresponding 2-D models for the images Terrace and Mondrian are shown in Fig. 7. We now discuss in somewhat more detail how we can determine if these models do indeed provide an adequate description of the experimentally obtained attribute and dissimilarity data.

Since a Minkowski power of $r = 2$ was used in the minimization of the stress function, ANOVA can be used to analyze the goodness of fit between the multidimensional model predictions and the experimental data. Briefly stated, ANOVA attributes the variance in observed data to several (potential) causes and determines the statistical evidence

for these causes. In our case, the data is split into model predictions and prediction errors. The hypothesis to be tested is that the prediction errors can be attributed solely to noise, so that the model is adequate and need not be improved further. The results of ANOVA analyses in our example are summarized in Table 1. We briefly discuss how these results were obtained and refer to the excellent book by Draper and Smith [19] for a more in-depth discussion on linear regression and ANOVA.

The observed attribute scores $A_{k,i,j}$ for different subjects k cannot be compared directly. XGms can, however, export normalized transformed attribute scores $TA^{\star}_{k,i,j}$ that can be used to compare responses across subjects that belong to a homogeneous group (i.e., that are assumed to share a common attribute vector). The most frequently used statistic for testing the goodness of fit of a model to (repeated) measurements is the *linear correlation coefficient* R, which is defined[11] as

$$\begin{aligned} R^2 &= 1 - \frac{\mathrm{SS(residue)}}{\mathrm{SS(total)}} \\ &= 1 - \frac{\sum_{k=1}^{k_a} \sum_{(i,j)\in I_a(k)} \left| TA^{\star}_{k,i,j} - \hat{\beta}_i \right|^2}{\sum_{k=1}^{k_a} \sum_{(i,j)\in I_a(k)} \left| TA^{\star}_{k,i,j} \right|^2}. \end{aligned} \quad (39)$$

The stimulus predictions $\hat{\beta}_i = [\mathbf{a}, \mathbf{x}_i]$ are derived from the stimulus configuration $\mathbf{x}_1, \ldots, \mathbf{x}_N$ and attribute vector $\mathbf{a}$ of the multidimensional model. This attribute vector is assumed to be shared by all k_a subjects in a group. The number of DOFs in the sum of squares SS (residue) is

$$\mathrm{DOF(residue)} = \sum_{k=1}^{k_a} N_a(k) - 2 \cdot k_a - n_a \quad (40)$$

since two DOFs are used per subject for obtaining the regression coefficients c_k and f_k that are involved in defining the normalized attribute scores and n_a DOF are needed to specify the attribute vector $\mathbf{a}$. In the current example, an inner-product model was used in two dimensions, so that $n_a = n - 1 = 1$ (in case of an ideal-point model $n_a = n$). If transformations are applied to the original attribute scores (i.e., if $TA_{k,i,j} \neq A_{k,i,j}$), then the number of DOFs in these transformations should also be subtracted from the above DOF (residue).

Note that no DOF are counted for the stimulus configuration itself, which is formally only allowed if the attribute data being tested are not used in determining this stimulus configuration. This is not strictly true in the current example in which the configuration is derived from the experimental data using MDS. However, the regression analysis addresses only a subset of the available experimental data and the obtained stimulus configurations with and without the data in this subset included are usually very similar. The statistics in Table 1 were, therefore, derived for a fixed configuration,

[11]This definition of the linear correlation coefficient R relies on the fact that the normalized scores are centered on the origin.

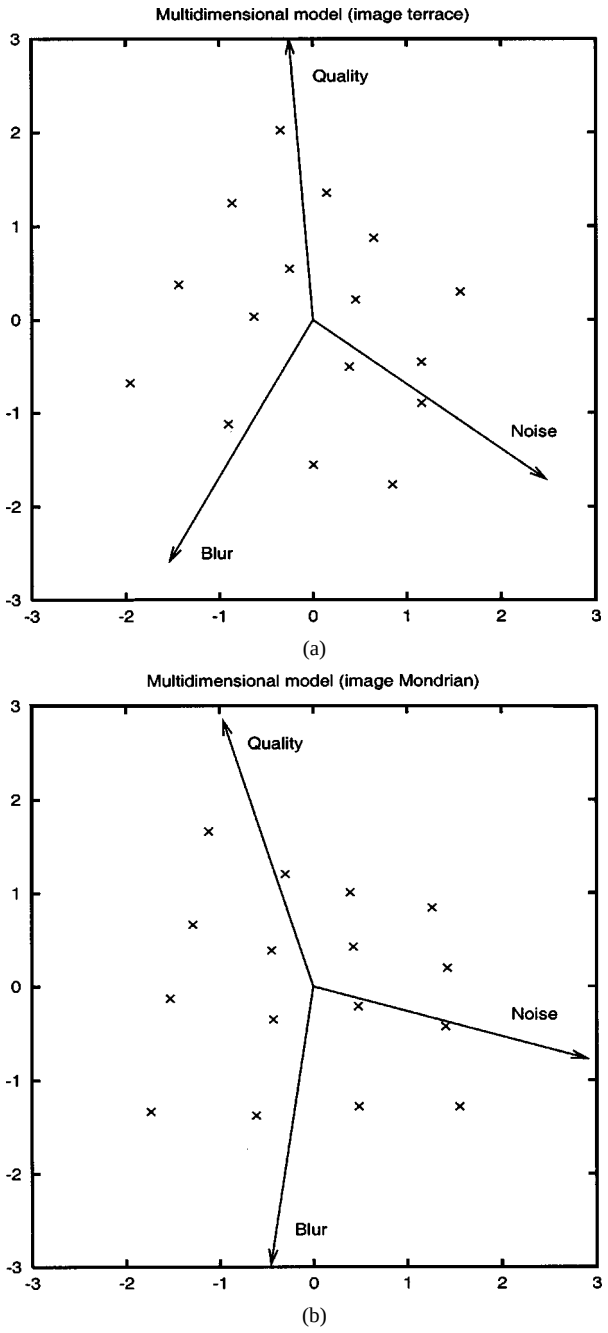


Fig. 7. 2-D stimulus configurations from the blur/noise experiment for scenes (a) "Terrace" and (b) "Mondrian." Directions for perceived quality, blur, and noise are indicated by the vectors.

based on all available data. We expect that this approximation has only limited consequences for the conclusions drawn from the following analyzes. If necessary, DOF (residue) can be further reduced to reflect the number of DOF in the stimulus configuration.

Linear correlation coefficients indicate the percentage of variance in the data that can be described by a linear prediction model, but do not allow to judge whether such a model can potentially be improved. This requires that the variance in the prediction error is compared against the inherent variance that is in the experimental data. We expect a smaller correlation between the actual data and the model predictions in case the responses to identical stimuli on repeated

Table 1
MDS Model Statistics for Dissimilarity (D) and Attributes Blur (B), Noise (N), and Overall Quality (Q) in Case of Images With Blur and Noise

Scene	*Attribute*	R_p	R	*Lack of fit measure*
Wanda	D	0.983	0.979	$F(119,476) = 1.163 < 1.258$
	B	0.878	0.874	$F(14,419) = 0.902 < 1.715$
	N	0.942	0.941	$F(14,419) = 0.495 < 1.715$
	Q	0.920	0.911	$F(14,419) = 3.210 > 1.715$ (*)
Terrace	D	0.980	0.976	$F(119,476) = 0.937 < 1.258$
	B	0.893	0.890	$F(14,419) = 0.878 < 1.715$
	N	0.957	0.955	$F(14,419) = 1.958 > 1.715$ (*)
	Q	0.939	0.932	$F(14,419) = 3.022 > 1.715$ (*)
Mondrian	D	0.979	0.972	$F(119,476) = 1.263 < 1.258$(*)
	B	0.886	0.881	$F(14,419) = 1.128 < 1.715$
	N	0.953	0.951	$F(14,419) = 1.260 < 1.715$
	Q	0.908	0.902	$F(14,419) = 1.700 < 1.715$

Cases where the statistics indicate that the model fit is not completely adequate are marked by (*).

trials vary more. This accuracy of the experimental data can be estimated if the stimuli have been scored more than once (either within or across subjects). In such case, the best possible predictor[12] for stimulus i is the average score

$$\beta_i = \frac{1}{M_i} \sum_k \sum_j TA^{\star}_{k,i,j} \tag{41}$$

where M_i is the number of repetitions for stimulus i. Using these average scores, SS (residue) can be split into

$$\text{SS(residue)} = \text{SS(pure error)} + \text{SS(lack of fit)} \tag{42}$$

where

$$\text{SS(pure error)} = \sum_{k=1}^{k_a} \sum_{(i,j)\in I_a(k)} \left(TA^{\star}_{k,i,j} - \beta_i\right)^2 \tag{43}$$

accumulates the variances on successive trials around the average score β_i and

$$\text{SS(lack of fit)} = \sum_{k=1}^{k_a} \sum_{(i,j)\in I_a(k)} \left(\hat{\beta}_i - \beta_i\right)^2 \tag{44}$$

is the SS of the differences between the average scores β_i and the predictions $\hat{\beta}_i$ according to the multidimensional model. The maximum value R_p that can be obtained for the linear correlation coefficient is

$$\begin{aligned} R_p^2 &= 1 - \frac{\text{SS(pure error)}}{\text{SS(total)}} \\ &= 1 - \frac{\sum_{k=1}^{k_a} \sum_{(i,j)\in I_a(k)} \left| TA^{\star}_{k,i,j} - \beta_i \right|^2}{\sum_{k=1}^{k_a} \sum_{(i,j)\in I_a(k)} \left| TA^{\star}_{k,i,j} \right|^2} \end{aligned} \tag{45}$$

since SS (residue) cannot become smaller than SS (pure error).

[12]Obviously, an even better fit to the data can be obtained by using predictors of the form $\beta_{k,i}$ that depend in a more general way on the subject k and stimulus i. Such a predictor is, however, not in agreement with the assumption that the subjects form a homogeneous group.

If the normalized scores $TA^{\star}_{k,i,j}$ are derived from minimizing a stress function with exponent $r = 2$, as is the case in the current example, then the following relationships

$$\sum_{i=1}^{N} \beta_i = \sum_{i=1}^{N} \hat{\beta}_i \text{ and } \sum_{i=1}^{N} \beta_i \cdot \hat{\beta}_i = \sum_{i=1}^{N} \hat{\beta}_i^2 \tag{46}$$

hold, so that there are only $N-2$ independent values β_i.

If the multidimensional model "fits" the data, then the standard deviation of the regression errors in case of the model predictors $\hat{\beta}_i$ should not exceed the standard deviation of the regression errors in case of the best predictors β_i. Unbiased estimates for these standard deviations are the mean squares (MS), i.e.,

$$\begin{aligned} \text{MS(lack of fit)} &= \frac{\text{SS(lack of fit)}}{\text{DOF(lack of fit)}} \\ \text{MS(pure error)} &= \frac{\text{SS(pure error)}}{\text{DOF(pure error)}} \end{aligned} \tag{47}$$

where DOF denotes the number of degrees of freedom

$$\begin{aligned} n_f &= \text{DOF(lack of fit)} = N - 2 \\ n_e &= \text{DOF(pure error)} \\ &= \sum_{k=1}^{k_a} N_a(k) - [2 \cdot k_a + n_a + (N-2)] \end{aligned} \tag{48}$$

in the respective SS. The ratio

$$F = \frac{\text{MS(lack of fit)}}{\text{MS(pure error)}} \tag{49}$$

is, hence, a good statistic for testing the goodness of fit of the multidimensional model. In case of equal standard deviations, we expect $F \approx 1$. A large value of F, on the other hand, indicates that the prediction error for the model is significantly larger than the prediction error for the best predictor and, hence, that R is substantially smaller than R_p. A better model than the multidimensional model under test can be pursued in such cases. A lack of fit ,however, does not prohibit that a large part of the variance in the attribute data may be explained by the available model, i.e., that both R and R_p are well above 0.9, for instance, so that the model may still be a very useful (but not perfect) description of the data.

It has been shown that the ratio F satisfies an $F(n_f, n_e)$ distribution under the hypothesis that *the standard deviations of the underlying (Gaussian) distributions are equal* [19]. If this is the case, then the probability that a value greater than F_α occurs is equal to

$$P(F > F_\alpha | n_f, n_e) = \frac{I_x\left(\frac{n_e}{2}, \frac{n_f}{2}\right)}{B\left(\frac{n_e}{2}, \frac{n_f}{2}\right)} = \alpha \tag{50}$$

with $x = n_e/(n_e + n_f \cdot F_\alpha)$, where we have introduced the notation

$$I_x(a,b) = \int_0^x z^{a-1}(1-z)^{b-1} dz \tag{51}$$

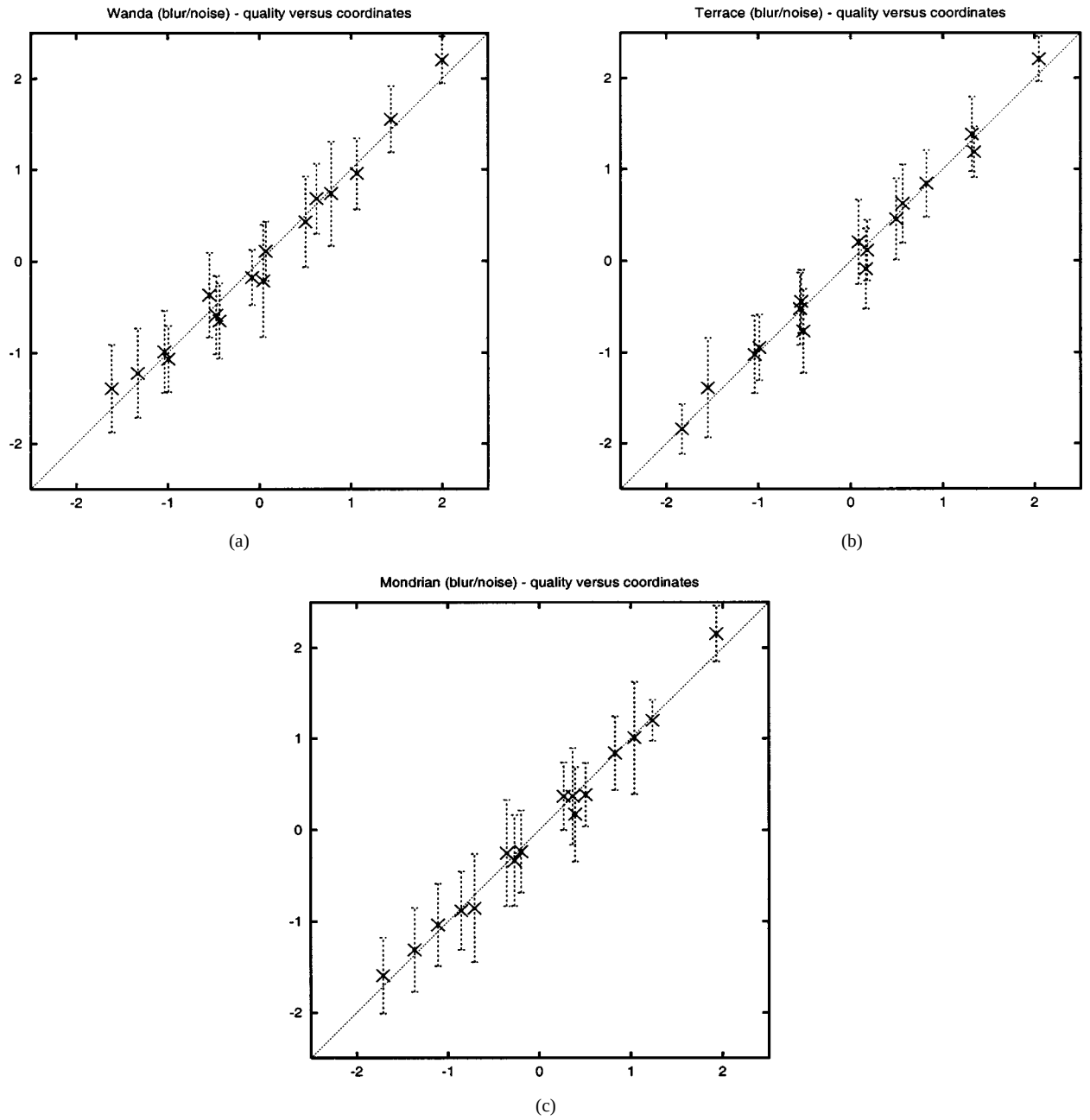


Fig. 8. Experimental quality scores versus MDS model predictions in the experiment with blur and noise for scenes (a) "Wanda," (b) "Terrace," and (c) "Mondrian." Estimated standard deviations of the experimental scores, obtained by regarding judgements within and across subjects as repeated measurements, are also indicated.

for the incomplete beta function and $B(a,b) = I_1(a,b)$ for the complete beta function. This result can be used to set up a quantitative F *test* for the goodness of fit. If the observed ratio F is larger than F_α, where most often $\alpha = 0.05$, then it is considered very unlikely that both standard deviations are equal and hence that all the variance in the data is accounted for by the model.

In case of our example data set, the number of DOFs in the prediction errors is $\mathrm{DOF(residue)} = 448 - 15 = 433$, since 16 stimuli where judged four times by seven subjects ($448 = 16 \cdot 4 \cdot 7$). The attribute scores were linearly correlated with the stimulus coordinates $\mathbf{x}_i$ in a 2-D space as independent variables. A single attribute vector with one DOF was used to derive attribute strengths from stimulus coordinates for all subjects, so that the total number of regression parameters is $\mathrm{DOF(regression)} = 1 + 7 \cdot 2 = 15$. The optimal predictor has $\mathrm{DOF(lack\ of\ fit)} = 16 - 2 = 14$ parameters, so that $\mathrm{DOF(pure\ error)} = 433 - 14 = 419$. The goodness of fit can, hence, be tested based on the distribution $F(14, 419)$. More specifically, with $\alpha = 0.05$, an observed ratio $F = \mathrm{MS(lack\ of\ fit)/MS(pure\ error)} > F_\alpha = 1.715$ is interpreted as a lack of fit.

The statistics in Table 1 indicate that the MDS models describe most of the variance in the experimental data

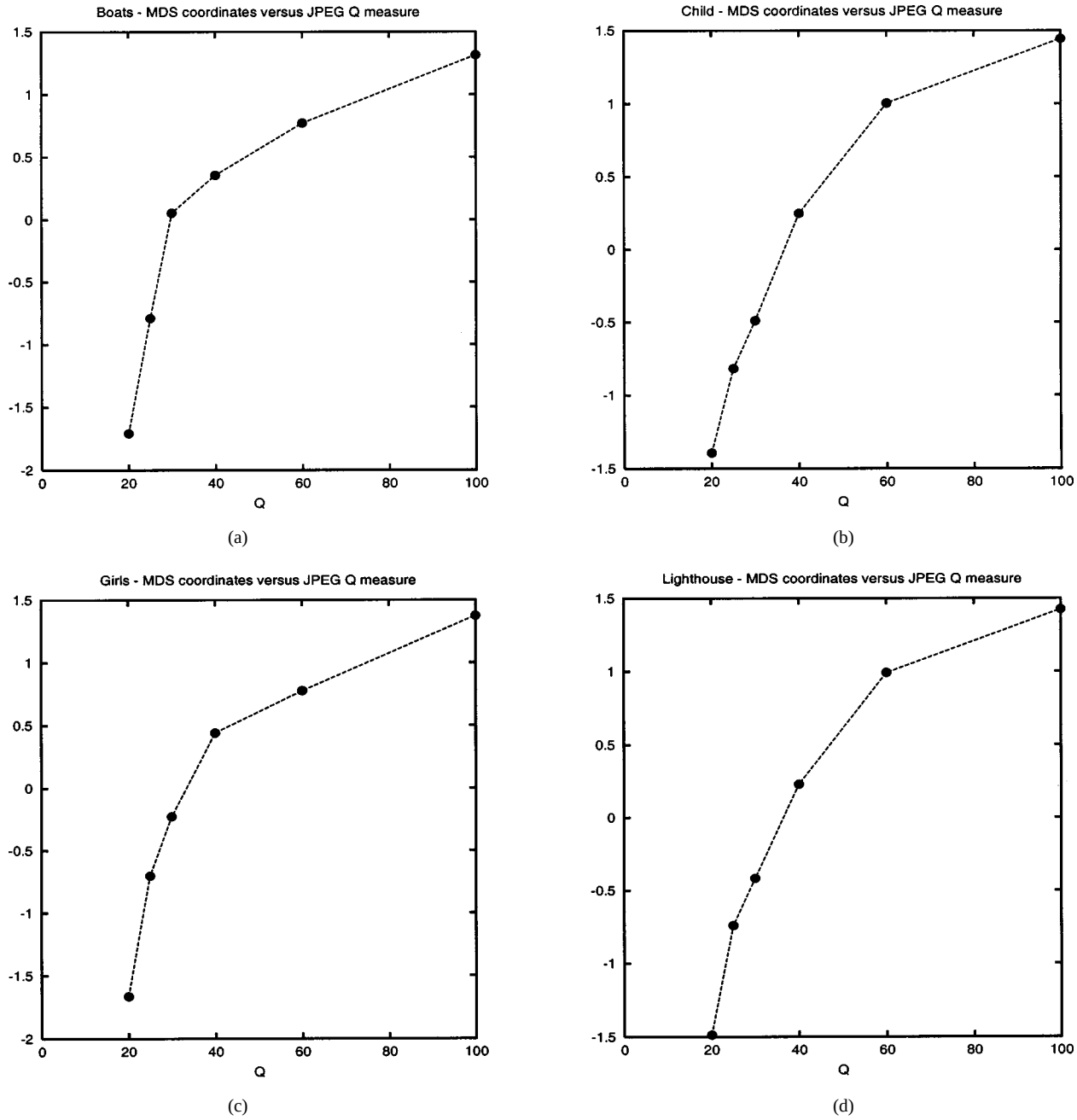


Fig. 9. 1-D stimulus configurations from JPEG experiment are shown along the ordinates for scenes (a) "Boats," (b) "Child," (c) "Girls," and (d) "Lighthouse." Abscissas show the JPEG quality (Q) parameter that is derived from the quantizer step size; the original image is plotted at value $Q = 100$.

($R > 0.87$ in all cases). The linear correlation coefficient R is largest for perceived noise and smallest for perceived blur. Since this trend is also observed in the pure-error correlations R_p, this only reflects the fact that blur is harder to judge consistently than quality, while noise is most easy to judge. The largest deviations from linear regression are observed for overall quality. This can for instance be verified in Fig. 8, where the normalized quality scores for all stimuli are plotted against the stimulus coordinates along the quality direction. The lack of fit seems to be mainly caused by the fact that the last point, which corresponds to the original image, falls above the regression line.

The regression analysis that is presented above for single-stimulus attribute data must be modified slightly in case of double-stimulus data, such as dissimilarity data. The correlation coefficient R for dissimilarity data is defined by

$$\begin{aligned} R^2 &= 1 - \frac{\mathrm{SS(residue)}}{\mathrm{SS(total)}} \\ &= 1 - \frac{\sum_{k=1}^{k_d} \sum_{(i,j)\in I_d(k)} \left| TD^{\star}_{k,i,j} - d_{ij} \right|^2}{\sum_{k=1}^{k_d} \sum_{(i,j)\in I_d(k)} \left| TD^{\star}_{k,i,j} \right|^2} \end{aligned} \quad (52)$$

where $d_{ij} = ||\mathbf{x}_i - \mathbf{x}_j||_l$ is the distance between the points representing stimulus i and j. The number of DOF in the residue error is

$$\mathrm{DOF(total)} = \sum_{k=1}^{k_d} N_d(k) - k_d \tag{53}$$

in case k_d subjects are assumed to form a homogeneous group (and no transformations are applied on the dissimilarity data, i.e., $TD_{k,i,j} = D_{k,i,j}$). The maximum correlation coefficient R_p satisfies

$$\begin{aligned} R_p^2 &= 1 - \frac{\mathrm{SS(pure\ error)}}{\mathrm{SS(total)}} \\ &= 1 - \frac{\sum_{k=1}^{k_d} \sum_{(i,j)\in I_d(k)} \left| TD^{\star}_{k,i,j} - \delta_{ij} \right|^2}{\sum_{k=1}^{k_d} \sum_{(i,j)\in I_d(k)} \left| TD^{\star}_{k,i,j} \right|^2} \end{aligned} \tag{54}$$

and corresponds to the best possible dissimilarity predictors

$$\delta_{ij} = \delta_{ji} = \frac{1}{M_{ij} + M_{ji}} \sum_{k=1}^{k_d} \left(TD^{\star}_{k,i,j} + TD^{\star}_{k,j,i} \right) \tag{55}$$

for $i, j = 1, \ldots, N$, where M_{ij} and M_{ji} denote the respective number of times that stimulus pairs (i, j) and (j, i) are repeated (across subjects). Note that the optimal predictor is assumed to satisfy $\delta_{ij} = \delta_{ji}$ and $\delta_{ii} = 0$, so that it is specified by $N \cdot (N-1)/2$ values. The lack-of-fit measure

$$\mathrm{SS(lack\ of\ fit)} = \sum_{k=1}^{k_d} \sum_{(i,j)\in I_d(k)} (d_{ij} - \delta_{ij})^2 \tag{56}$$

sums the squared differences between the optimal dissimilarity predictions and the interstimulus distances according to the multidimensional model. In case of a stress function with exponent $r = 2$, the best predictors satisfy the condition

$$\sum_{(i,j)\in I_d} d_{ij} \cdot \delta_{ij} = \sum_{(i,j)\in I_d} d_{ij}^2 \tag{57}$$

so that $\mathrm{DOF(lack\ of\ fit)} = N \cdot (N-1)/2 - 1$. For our example data, we obtain that $\mathrm{DOF(total)} = 5 \cdot 120 - 5 = 595$, $\mathrm{DOF(lack\ of\ fit)} = 119$ and $\mathrm{DOF(pure\ error)} = 476$, so that an $F(119, 476)$ test with 5% confidence value $F_{0.05} = 1.258$ is used to test the goodness of fit of the multidimensional model to the experimental dissimilarity data.

B. JPEG-Coded Images

The second data set that we analyze concerns JPEG-coded images [17]. The images were obtained by applying six different quality levels in the JPEG-baseline encoding [36] of four different scenes. The measurements performed per scene were as follows: double-stimulus dissimilarity scaling by ten subjects and double-stimulus difference scaling of perceived blockiness and quality by the same ten subjects. We used XGms with an inner-product model for preference to derive the 1-D stimulus configurations in Fig. 9. An inner-product model in one dimension automaticallly implies that the subjects are considered to form a homogeneous group. Nonmetric analysis again confirmed only a marginal improvement over metric analysis, so that the reported analysis is again based on metric data (i.e., $TD_{k,i,j} = D_{k,i,j}$ and $TP_{k,i,j} = P_{k,i,j}$).

Table 2
MDS Model Statistics for Dissimilarity (D) and Attributes Blockiness (B) and Overall Quality (Q) in Case of JPEG-Coded Images

Scene	*Attribute*	R_p	R	*Lack of fit measure*
Boats	D	0.948	0.941	$F(14,126) = 1.014 < 1.771$
	B	0.893	0.886	$F(14,126) = 0.522 < 1.771$
	Q	0.893	0.889	$F(14,126) = 0.318 < 1.771$
Child	D	0.970	0.956	$F(14,126) = 4.059 > 1.771$ (*)
	B	0.961	0.949	$F(14,126) = 2.839 > 1.771$ (*)
	Q	0.966	0.956	$F(14,126) = 2.749 > 1.771$ (*)
Girls	D	0.970	0.956	$F(14,126) = 4.054 > 1.771$ (*)
	B	0.939	0.931	$F(14,126) = 1.149 < 1.771$
	Q	0.959	0.941	$F(14,126) = 3.968 > 1.771$ (*)
Lighthouse	D	0.945	0.929	$F(14,126) = 2.538 > 1.771$ (*)
	B	0.946	0.940	$F(14,126) = 0.981 < 1.771$
	Q	0.956	0.947	$F(14,126) = 1.689 < 1.771$

Cases where the statistics indicate that the model fit is not completely adequate are marked by (*).

Table 2 should be interpreted in a similar way as Table 1. The normalized scores of different subjects for the same stimulus combination (i, j) and attribute (dissimilarity, blockiness or quality) were treated as repetitions, so that the statistics in Table 2 indicate how the derived 1-D stimulus configurations can describe the average subject responses.

The results of the ANOVA analyses for dissimilarity (D), blockiness (B) and quality (Q) indicate that large correlation coefficients are found (i.e., $R > 0.88$), so that the majority of the variance in the date can be described by a 1-D model. However, significant deviations from linear regression are especially observed for the scenes "Child" and "Girls." This can for instance be verified in Fig. 10, where the normalized quality scores for all stimulus pairs are plotted against the differences in the stimulus coordinates. Fig. 10 illustrates that the deviation from the regression line is larger than the standard deviation of the experimental error for some of the data points. A more complex (for instance, higher dimensional) model will be needed to describe these remaining deviations.

VII. Summary

In this paper, we have shown how experimental data on image quality and its attributes can be integrated using multidimensional models. The interactive program XGms for estimating multidimensional models from data has been introduced as an interesting extension to available programs. It has been shown that the models implemented in XGms can indeed provide adequate descriptions for two available experimental data sets.

The experimental data are treated as continuous variables in the data analyses presented in this paper. In case subjects

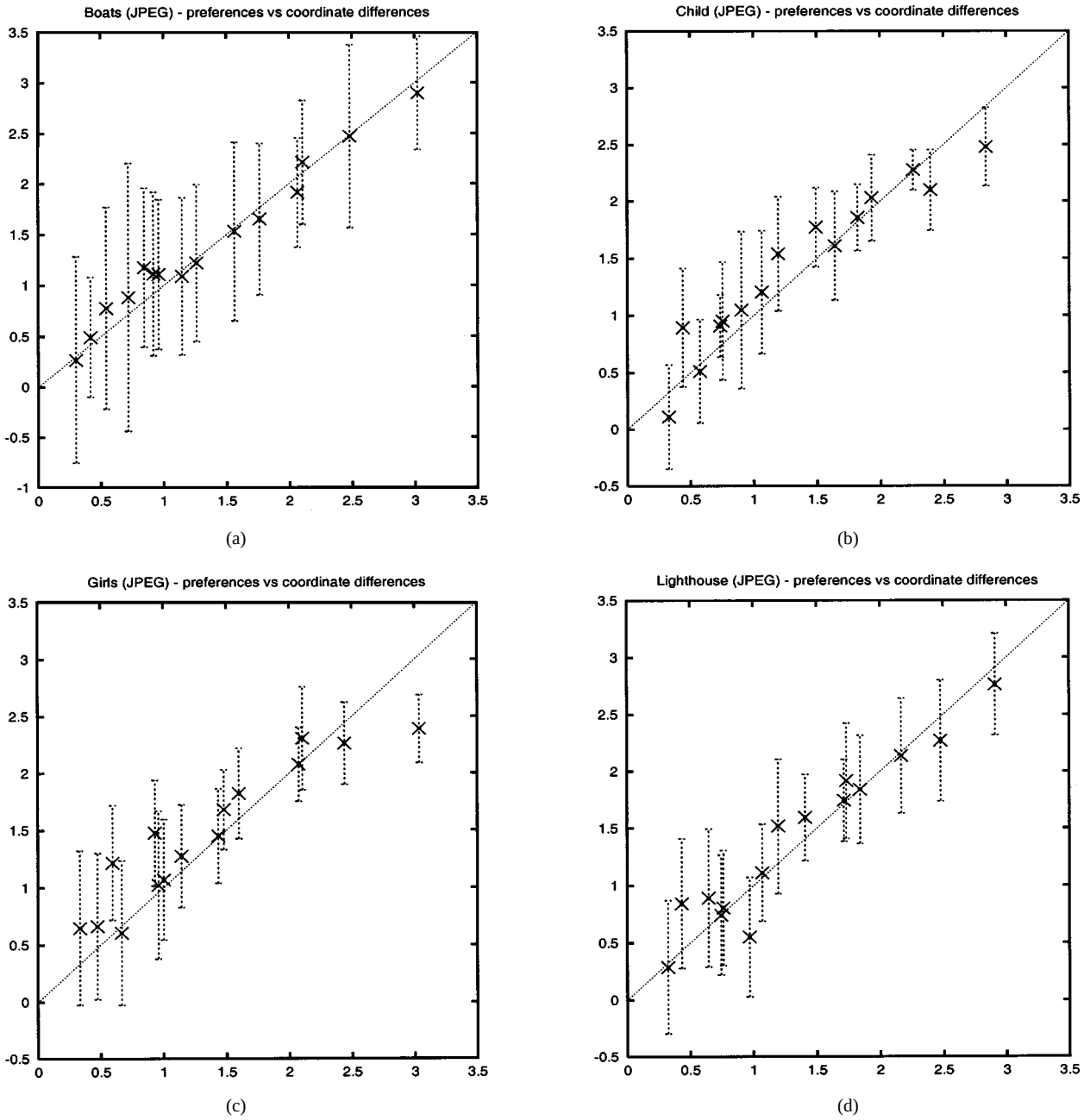


Fig. 10. Experimental quality scores versus MDS model predictions in the JPEG experiment for scenes (a) "Boats," (b) "Child," (c) "Girls," and (d) "Lighthouse." The estimated standard deviations of the experimental scores, obtained by regarding judgements across subjects as repeated measurements, are also indicated.

use numerical category scaling with only a limited number of discrete categories for expressing their sensations, this is an obvious approximation. If the quantization noise introduced by the categorical scaling can be assumed to be substantially smaller than the internal noise underlying the judgements, this approximation is sufficiently accurate. However, in order to remedy this limitation, we are currently developing a version of XGms in which the input data can also be interpreted as discrete (categorical) data. The main implication of this change is that the stress optimization will have to be replaced by a maximum-likelihood optimization [6], [20].

ACKNOWLEDGMENT

The author would like to thank the many people that have contributed to the public software in XGobi, XGvis, and Netlib. Without this freely available software, realization of the XGms program would not have been feasible.

REFERENCES

[1] P. E. Green, F. J. Carmone Jr., and S. M. Smith, *Multidimensional Scaling, Concepts and Applications*. Boston, MA: Allyn & Bacon, 1989.

[2] F. W. Young and R. M. Hamer, *Multidimensional Scaling: History, Theory, and Applications*. New York: Erlbaum, 1987.
[3] T. F. Cox and M. M. A. Cox, *Multidimensional Scaling*. London, U.K.: Chapman & Hall, 1994.
[4] J. B. Martens and L. Meesters, "Single-ended intrumental measurement of image quality," in *Vision Models*, C. J. van den Branden Lambrecht, Ed. Norwell, MA: Kluwer, 2001.
[5] M. P. Eckert and A. P. Bradley, "Perceptual quality metrics applied to still image compression," *Signal Processing*, vol. 70, no. 3, pp. 177–200, Nov. 1998.
[6] J. B. Martens and M. Boschman, "The psychophysical measurement of image quality," in *Vision Models*, C. J. van den Branden Lambrecht, Ed. Norwell, MA: Kluwer, 2001.
[7] R. D. Luce and C. L. Krumhansl, "Measurement, scaling and psychophysics," in *Stevens' Handbook of Experimental Psychology—Perception and Motivation*, R. C. Atkinson, R. J. Herrnstein, G. Lindzey, and R. D. Luce, Eds. New York: Wiley, 1988, pp. 3–74.
[8] A. Buja, D. F. Swayne, M. L. Littman, and N. Dean, "Xgvis: Interactive data visualization with multidimensional scaling," J. Comput. Graphical Statistics, 2001, to be published.
[9] V. Kayargadde and J. B. Martens, "Perceptual characterization of images degraded by blur and noise: Experiments," *J. Opt. Soc. Amer. A*, vol. 13, no. 6, pp. 1166–1177, June 1996.
[10] "Methodology for the Subjective Assessment of the Quality of Television Pictures," ITU-R, Recommendation ITU-R BT.500–8, 1998.
[11] H. de Ridder and G. M. M. Majoor, "Numerical category scaling: An efficient method for assessing digital image coding impairments," in *Human Vision and Electronic Imaging: Models, Methods and Applications*, B. E. Rogowitz and J. P. Allebach, Eds., 1990, vol. 1249, Proc. SPIE, pp. 65–77.
[12] L. L. Thurstone, "A law of comparative judgement," *Psychol. Rev.*, vol. 34, pp. 273–286, 1927.
[13] W. S. Torgerson, *Theory and Methods of Scaling*. New York: Wiley, 1958.
[14] W. A. Wagenaar, "Stevens versus Fechner: A plea for dismissal of the case," *Acta Psychologica*, vol. 39, pp. 225–235, 1975.
[15] J. A. J. Roufs, F. F. J. Blommaert, and H. de Ridder, "Brightness-luminance relations: Future developments in the light of the past," in *Proc. CIE*, Melbourne, Australia, 1991.
[16] K. Teunissen, "The validity of CCIR quality indicators along a graphical scale," *SMPTE J.*, vol. 105, pp. 144–149, Mar. 1996.
[17] J. B. Martens and L. Meesters, "Image dissimilarity," *Signal Processing*, vol. 70, no. 3, pp. 155–176, Nov. 1998.
[18] J. C. Falmagne, "Psychophysical measurement and theory," in *Handbook of Perception and Human Performance—Sensory Processes and Perception*, K. R. Boff, L. Kaufman, and J. P. Thomas, Eds. New York: Wiley, 1986, pp. 1–66.
[19] N. R. Draper and H. Smith, *Applied Regression Analysis*. New York: Wiley, 1998.
[20] J. O. Ramsay, "Maximum likelihood estimation in multidimensional scaling," *Psychometrika*, vol. 42, no. 2, pp. 241–266, June 1977.
[21] F. W. Young, "Scaling," *Annu. Rev. Psychol.*, vol. 35, pp. 55–81, 1984.
[22] A. Agresti, *Categorical Data Analysis*. New York: Wiley, 1990.
[23] M. Yuen and H. R. Wu, "A survey of hybrid mc/dpcm/dct video coding distortions," *Signal Processing*, vol. 70, no. 3, pp. 247–278, Nov. 1998.
[24] J. B. Kruskal and M. Wish, "Multidimensional scaling," in *Sage University Paper Series 07–011 on Quantitative Applications in the Social Sciences*. Beverley Hills, CA: Sage Publications, 1978.
[25] R. N. Shepard, "The analysis of proximities: Multidimensional scaling with an unknown distance function," *Psychometrika*, vol. 27, no. 3, pp. 219–246, Sept. 1962.
[26] J. B. Kruskal, "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," *Psychometrika*, vol. 29, no. 1, pp. 1–27, Mar. 1964.
[27] ——, "Nonmetric multidimensional scaling: A numerical method," *Psychometrika*, vol. 29, no. 2, pp. 115–129, June 1964.
[28] H. Marmolin and S. Nyberg, "Multidimensional Scaling of Subjective Image Quality," Swedish Nat. Defense Res. Inst., Stockholm, Sweden, FOA Rep. C 30039-H9, 1975.
[29] J. S. Goodman and D. E. Pearson, "Multidimensional scaling of multiply-impaired television pictures," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-9, pp. 353–356, June 1979.
[30] B. Escalante-Ramírez, J. B. Martens, and H. de Ridder, "Multidimensional characterization of the perceptual quality of noise-reduced computed tomography images," *J. Vis. Commun. Image Represent.*, vol. 6, no. 4, pp. 317–334, Dec. 1995.
[31] J. O. Ramsay, "The joint analysis of direct rating, pairwise preferences and dissimilarities," *Psychometrika*, vol. 45, no. 2, pp. 149–165, June 1980.
[32] D. F. Swayne, D. Cook, and A. Buja, "Xgobi: Interactive dynamic data visualization in the x window system," *J. Comput. Graphical Statistics*, vol. 7, no. 1, pp. 113–130, Mar. 1998.
[33] R. N. Shepard, "Representation of structure in similarity data: Problems and prospects," *Psychometrika*, vol. 39, no. 4, pp. 373–424, Dec. 1974.
[34] A. J. Ahumada and C. H. Null, "Image quality: A multidimensional problem," in *Digital Images and Human Vision*, A. Watson, Ed. Cambridge, MA: MIT Press, 1993, pp. 141–148.
[35] J. O. Ramsay, "Monotonic weighted power transformations to additivity," *Psychometrika*, vol. 42, no. 1, pp. 83–109, Mar. 1977.
[36] W. B. Pennebaker and J. L. Mitchell, *JPEG Still Image Compression Standard*. New York: Van Nostrand Reinhold, 1993.

Jean-Bernard Martens was born in Eeklo, Belgium, in 1956. He received the M.S. and Ph.D. degrees in electrical engineering from the University of Ghent, Ghent, Belgium, in 1979 and 1983, respectively.

Since October 1984, he has been with the Eindhoven University of Technology, Eindhoven, The Netherlands. His current research interests lie in perceptually related image coding and processing, measurement and modeling of perceived image quality, mathematical modeling of the human visual system, data visualization, computer vision, the use of visual pattern recognition in multimodal interfaces, and the design and application of such new multimodal interfaces. His previous research interests include number theory, with applications in efficient algorithms for digital signal processing.