Journal of Electronic Imaging 10(1), 47–55 (January 2001).

Cognitive issues in image quality measurement

Huib de Ridder
Delft University of Technology
Department of Industrial Design
Jaffalaan 9, 2628 BX Delft, The Netherlands
E-mail: h.deridder@io.tudelft.nl

Abstract. *Designers of imaging systems, image processing algorithms, etc., usually take for granted that methods for assessing perceived image quality produce unbiased estimates of the viewers' quality impression. Quality judgments, however, are affected by the judgment strategies induced by the experimental procedures. In this paper the results of two experiments are presented illustrating the influence judgment strategies can have on quality judgments. The first experiment concerns contextual effects due to the composition of the stimulus set. Subjects assessed the sharpness of two differently composed sets of blurred versions of one static image. The sharpness judgments for the blurred images present in both stimulus sets were found to be dependent on the composition of the set as well as the scaling technique employed. In the second experiment subjects assessed either the overall quality or the overall impairment of manipulated and standard JPEG-coded images containing two main artifacts. The results indicate a systematic difference between the quality and impairment judgments that could be interpreted as instruction-based different weighting of the two artifacts. Again, some influence of scaling technique was observed. The results of both experiments underscore the important role judgment strategies play in the psychophysical evaluation of image quality. Ignoring this influence on quality judgments may lead to invalid conclusions about the viewers' impression of image quality.* © 2001 SPIE and IS&T. [DOI: 10.1117/1.1335529]

1 Introduction

Present-day technology begins to make it feasible to communicate complex information in a natural, dynamic way. In the coming years, for example, electronic imaging technology is expected to contribute substantially to the development of communication media that must lead human observers to believe that they are actually present in the environment displayed (e.g., virtual space teleconferencing,[1] immersive television[2]). At the same time, persisting limitations in transmission bandwidth and data storage will keep on forcing system designers to employ high levels of data compression introducing content-dependent temporal fluctuations in the quality of information presentation.[3] When annoying to the end user, such quality fluctuations may seriously threaten the acceptability of new media.

The impact of experienced quality on the acceptability of media is one of the main reasons why in the field of electronic imaging there has always been much interest in methods and tools for assessing and predicting perceptual image quality.[4–7] In the last two decades, the prospect of being able to predict the viewer's quality impression directly from the physical image has elicited a large variety of so-called objective quality metrics, i.e., measures generated by algorithms which aim to correlate well with quality judgments of human observers. The complexity of these algorithms has gradually increased in time by incorporating increasingly more properties of the early human visual system and, more recently, higher level cognitive processes (e.g., attention, memory).[3,7] The main reasons for this trend are the relatively low correlation between most of the objective measures and the viewers' judgments and the frequently observed scene dependency of the objective measures. Looking at this trend, it is surprising to realize that, in general, the problems encountered with objective measures are attributed to the limited capacities of the algorithms and not to the quality judgments. Apparently, there is an implicit assumption that quality judgments are a faithful representation of quality impressions. But is that always the case? And does it hold for generally accepted standardized evaluation techniques like the ones recommended by the International Telecommunication Union (ITU)?

Recommendation 500 of the International Telecommunication Union/Radio Communication (ITU/R)[8] is probably the most frequently cited document in the field of image quality evaluation. This document describes, among others, scaling methods and viewing conditions for assessing the perceived quality of television pictures in a standardized way. The objective of these methods is to generate opinion scores reflecting the viewers' quality impression. To date, subjective evaluation is regarded as the most effective and reliable way of assessing image quality, especially because widely used objective measures like root-mean-squared error and peak signal-to-noise ratio are, in general, not able to provide a good indication of perceived quality.[9]

As already suggested above, designers of imaging systems, image processing algorithms, etc., usually take for granted that such quality ratings are unbiased estimates of the viewers' quality impression.[9] In general, this will not be the case. In his book *Sensation and Judgment: Complementary Theory of Psychophysics*, Baird[10] points out that in psychophysical experiments subjects' responses are deter-

This paper is a slightly modified version of an invited paper presented on the tenth anniversary of the IS&T/SPIE conference on Human Vision and Electronic Imaging, 26–29 January 1998, San Jose, CA.

Paper HVEI-03 received Dec. 23, 1999; revised manuscript received July 11, 2000; accepted for publication Aug. 24, 2000.
1017-9909/2001/$15.00 © 2001 SPIE and IS&T.

mined not only by the percept itself but also by the judgment strategies induced by the experimental procedures. As a consequence, the same stimulus may elicit different responses under varying conditions. Quality judgments are no exception to this rule as will be shown in this paper.

What can be done about this apparent malleability of opinion scores? One possibility is to identify the sources of this context-dependent flexibility and to control them by meticulously specifying the conditions under which experiments have to be carried out.[10–12] According to Gescheider,[12] this agrees with ''...the approach of the sensory scientist whose goal is to obtain unbiased scales of sensory magnitude to study sensory processes such as summation, inhibition, adaptation, and sensory channels.'' (p. 183). A possible disadvantage of this approach is its inability to generalize to other conditions. An alternative approach is to accept the malleability of opinion scores[10,12] and to establish rules for deriving quantitative measures of perceived quality from context-dependent quality judgments. This ''...represents the approach of the cognitive scientist whose goal is to understand the process of judgment. To these investigators, biased responses influenced by context are interesting—even welcome—and no doubt represent the way most people make judgments outside the laboratory'' (Gescheider,[12] p. 183). It is in line with this approach to assume that judgment strategies have such an impact on quality ratings that ignoring their influence may lead to invalid conclusions about the viewers' impression of image quality.

The objective of this paper is to demonstrate the influence judgment strategies can have on quality judgments. To this end, the results of two experiments are presented. These experiments were specially designed for this purpose and concern two well-known issues in the field of judgmental processes,[10] namely, the influence of context, in this case the composition of the stimulus set (experiment 1), and the influence of instructions, in this case quality versus impairment judgments (experiment 2). The image material consisted of blurred versions of one static image in experiment 1 and both manipulated and standard JPEG-coded images in experiment 2. In the General Discussion (Sec. 4) it is argued that the influence of judgment strategies can be demonstrated not only in specially designed experiments but also in ''normal'' experiments. As an illustration, the results of a third experiment are briefly mentioned. The goal of that experiment was to link instantaneously perceived quality to overall quality judgments of MPEG-2-coded video sequences.[13] This led to a model incorporating a recency effect and a nonlinear averaging procedure stressing the importance of strong impairments.

2 Experiment 1: Contextual Effects in Sharpness Judgments

2.1 *Introduction*

Scaling is one of the most efficient methods for assessing perceived image quality and its underlying dimensions (sharpness, brightness, colorfulness, etc.).[6] Experiments with simple stimuli such as squares, circles and dot patterns have shown, however, that the outcome of a scaling experiment is susceptible to contextual effects.[14,15] That is, the response to a stimulus depends not only on the stimulus itself but also on the other stimuli to be judged in a session. Contextual effects due to stimulus spacing or frequency of occurrence of stimuli have been found to have a substantial influence on the results of a single stimulus (or ''direct'') scaling experiment.[15] The question becomes whether and, if so, to what extent contextual effects are present when complex stimuli, e.g., digitally coded images of natural scenes, are evaluated.

In practice, the composition of the stimulus set to be evaluated on, for example, image quality is often fixed and cannot be manipulated. Accordingly, contextual effects due to stimulus spacing may seriously threaten the reliability and, particularly, the validity of the outcome of such a quality assessment. It was, therefore, decided to limit the investigation to the possible influence of stimulus spacing. In the experiment, subjects were instructed to assess the perceived sharpness of low-pass filtered versions of one static image. The advantage of this image material is that the mapping from Gaussian spatial filtering to perceived sharpness can be predicted quantitatively.[16–18] The influence of stimulus spacing on the evaluation of perceived sharpness was measured for the following three scaling techniques: single stimulus scaling, double stimulus scaling and a scaling procedure based on difference judgments (''comparison scaling''). Note that these techniques represent the three kinds of evaluation methods recommended by the International Telecommunication Union (ITU).[8]

2.2 *Method*

2.2.1 *Image material*

The sharpness of the static image of a terrace scene was manipulated using a Gould deAnza Image Processing System IPS8400. The video signal obtained by scanning the slide of this scene was digitized with 8 bits/pixel on a grid of 512×512 pixels. During the experiment, however, only the central part of the scene was displayed (476×471 pixels). Low-pass filtering the original with the aid of a two-dimensional (2D) separable binomial filter generated blurred versions of the original image. The resulting perceptual strength of blur is related to the standard deviation of the corresponding Gaussian kernel (σ, expressed in pixels) by the following equation[17,18]:

$$S_b = 1 - ((\sigma/\sigma_0)^2 + 1)^{-0.25}, \tag{1}$$

where S_b denotes the perceptual strength of blur and σ_0 can be interpreted as the standard deviation of the eye's internal blurring kernel. In the present study, σ_0 was fixed at a value of 0.73 (Fig. 1, left-hand panel). There were two sets of blurred images, which had seven images in common. The perceptual strength of blur S_b of these seven images ranged from 0 (the original image; $\sigma=0$) to 0.58 ($\sigma=4.07$) in regular steps of about 0.1. Accordingly, these images were evenly distributed with respect to their perceived (un)sharpness (Fig. 1, filled symbols). A stimulus set with a negatively skewed distribution was created by adding 15 comparatively unsharp images (Fig. 1, right-hand panel, open symbols in upper row). Similarly, a stimulus set with a positively skewed distribution was created by adding 15 comparatively sharp images (Fig. 1, right-hand panel, open symbols in lower row).

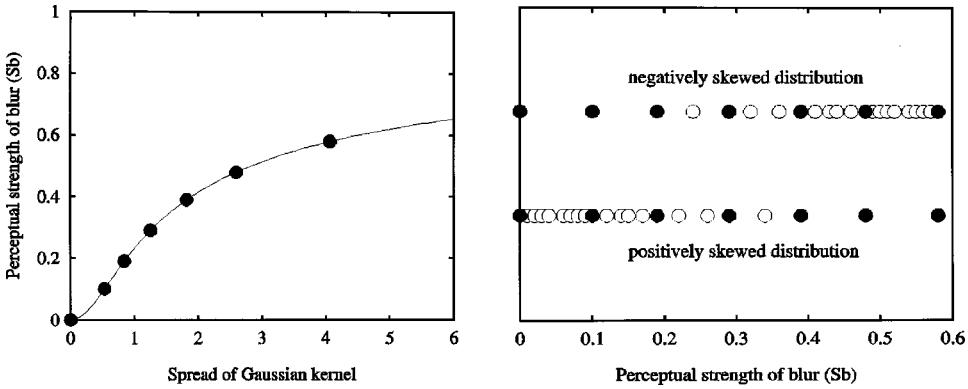

Fig. 1 Left-hand panel: Perceptual strength of blur S_b as a function of spread parameter σ. Filled symbols denote the seven images that belong to the negatively as well as the positively skewed stimulus set. Right-hand panel: Schematic representation of the negatively and positively skewed stimulus sets.

2.2.2 *Procedure*

The black-and-white images were displayed on a 70 Hz interlaced Barco CCID7351B CRT monitor placed in a dark room in front of a dimly lit ''white'' background. The monitor was corrected such that the screen luminance was linearly related to the optical density of the original slide. The images were presented for 5 s after which a 9 cd/m^2 adaptation field appeared on the screen. This luminance level was the average luminance of the test images. Viewing conditions were in accordance with ITU Recommendation 500.[8] The subjects viewed the monitor at a distance of about 1.5 m. At this distance, the pixel size is about 1 min of arc. This implies that under these experimental conditions the pixel structure was just not visible. During a session, the subjects saw the images of either the positively or the negatively skewed stimulus set. The sharpness of these images was assessed in three ways: single stimulus scaling on a ten-point numerical category scale ranging from 1 (lowest sharpness) to 10 (highest sharpness), double stimulus scaling using the same ten-point numerical category scale and comparison scaling. For the double stimulus as well as the comparison scaling experiment reference images had to be introduced. In these experiments each trial consisted of a test and a reference image that were displayed sequentially with an interval of 2 s between the two 5 s presentations. After each trial the subjects had to rate the sharpness of the two images on two separate ten-point numerical scales in the case of double stimulus scaling and the difference between the perceived sharpnesses on a single scale ranging from -10 (the first image is much sharper than the second one) to 10 (the second image is much sharper than the first one) in the case of comparison scaling. Before the results obtained by the double stimulus method were analyzed, the ratings for the test and the reference image (the original with $S_b=0$) were always subtracted. Three images with $S_b=0$, $S_b=0.29$ and $S_b=0.58$ were used as references in the comparison scaling experiment. Per session the subjects assessed the difference in sharpness between the 22 images of either the positively or negatively skewed stimulus set and each of these three reference images.

2.2.3 *Subjects*

Eight inexperienced subjects in the age from 20 to 28 years participated in the experiment. They had normal or corrected-to-normal vision. Their visual acuity measured with the aid of a Landolt chart at a distance of 5 m varied between 1.5 and 2.5. Four subjects took part in both the single stimulus scaling experiment and the comparison scaling experiment. The other subjects carried out the double stimulus scaling experiment.

2.2.4 *Data analysis*

The possible influence of the stimulus spacings on sharpness judgments was analyzed with the aid of Parducci's range-frequency model.[19,20] This model states that subjects tend to cover the perceptual range under investigation by the whole response scale and at the same time try to use each category an equal number of times. This implies that category judgments are a compromise between two principles, namely, a range principle postulating that each stimulus is judged in relation to the extreme stimuli that form the stimulus range and a frequency principle postulating that the same number of stimuli is assigned to each category. Judgment $J_{i,c}$ of stimulus i in context c is assumed to be the weighted sum of these two principles, or

$$J_{i,c}=w*R_{i,c}+(1-w)*F_{i,c}, \tag{2}$$

where $J_{i,c}$ is the category judgment linearly transformed to a scale running from 0 to 1. Range value $R_{i,c}$ is related to perceptual strength S_i by means of the following equation:

$$R_{i,c}=(S_i-S_{\min,c})/(S_{\max,c}-S_{\min,c}), \tag{3}$$

in which $S_{\min,c}$ and $S_{\max,c}$ are the perceptual strengths of the extreme stimuli. The frequency value $F_{i,c}$ of stimulus i in context c is related to the rank $r_{i,c}$ of this stimulus, or

$$F_{i,c}=(r_{i,c}-1)/(N_c-1), \tag{4}$$

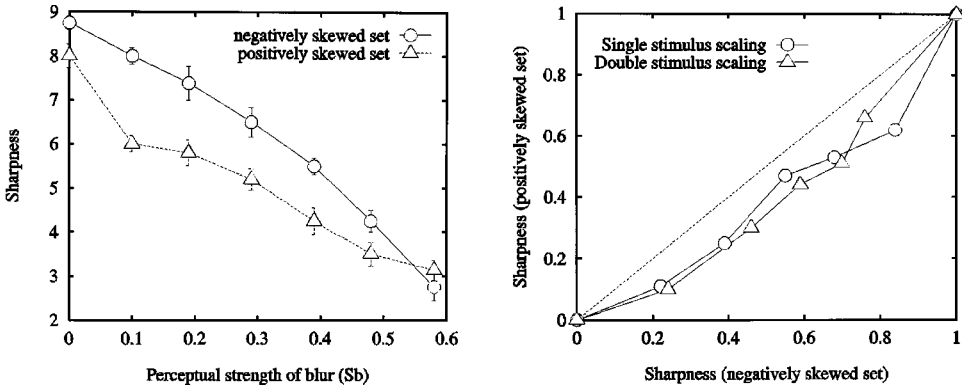

Fig. 2 Sharpness judgments for the images that were present in both the negatively and positively skewed stimulus set. Left-hand panel: Data for one subject (CD). Vertical bars denote twice the standard error of the mean. Data are based on eight repetitions. Right-hand panel: Sharpness judgments averaged across subjects. The judgments have been linearly transformed to a scale running from 0 to 1. The average standard error of the mean is 0.02 and 0.03 for the positively skewed and negatively skewed set, respectively. Error bars have not been added because the standard errors of the mean are smaller than the size of the symbols in the figure. Dotted line indicates predicted results in the absence of contextual effects.

N_c being the total number of stimuli in context c. In the present study, frequency value $F_{i,c}$ increased with the perceptual strength of blur (S_i). At the same time, sharpness judgment $J_{i,c}$ decreased with the perceptual strength of blur. The original image, for example, had the lowest rank number but the highest sharpness judgment. To settle this discrepancy, judgment $J_{i,c}$ in Eq. (2) was replaced by ''$1-J_{i,c}$.''

In the present study the extreme stimuli were always the same (Fig. 1). Hence, the range values are independent of stimulus spacing. Furthermore, the frequency values are based on the perceptual strength of blur, whereas the subjects evaluated the sharpness of the images. Taking this into account, the weighting factor w, reflecting the influence of context effects, can easily be derived from Eqs. (2), (3) and (4). This results in the following expression:

$$w=1-(J_{i,\mathrm{neg}}-J_{i,\mathrm{pos}})/(F_{i,\mathrm{pos}}-F_{i,\mathrm{neg}}), \tag{5}$$

where pos and neg represent the positively and negatively skewed stimulus sets, respectively. If w equals one, then the judgments are independent of stimulus spacing. If w is less than one, a clear context effect is present. For simple stimuli, the value of w varies around 0.5.[15] In the present study, the value of w was estimated by fitting the following expression to the experimental data:

$$(J_{i,\mathrm{neg}}-J_{i,\mathrm{pos}})=(1-w)(F_{i,\mathrm{pos}}-F_{i,\mathrm{neg}}). \tag{6}$$

2.3 *Results*

The left-hand panel of Fig. 2 shows the unprocessed data of one subject for the single stimulus condition. It denotes the sharpness judgments for the seven images that were present in both the negatively and positively skewed stimulus set. Comparable results were obtained for the other subjects in both the single stimulus and double stimulus condition. The general trend in the data is that the images in the middle of the blur range are consistently judged sharper when they appear in the negatively skewed stimulus set than when they are part of the positively skewed stimulus set (Fig. 2, right-hand panel). This trend is in accordance with Parducci's model. To quantify this phenomenon, the value of w was determined by fitting Eq. (6) to the averaged data. The resulting values were: 0.74 ± 0.09 ($r^2=0.80$; 95% confidence interval: $0.57-0.92$) for the single stimulus method and 0.71 ± 0.06 ($r^2=0.91$; 95% confidence interval: $0.57-0.83$) for the double stimulus method. These values deviate significantly from one, implying that there is a context effect. At the same time they are significantly higher than 0.5, suggesting that the contextual effects observed with natural images will not be as strong as those observed with simple stimuli.[15]

The left-hand panel of Fig. 3 shows the differences in sharpness, as judged by one subject, for the seven images belonging to both stimulus sets and the three references. Again, similar results were found for the other subjects. The data in Fig. 3 indicate that sharpness judgments obtained via comparison scaling are hardly influenced by the composition of the stimulus set. This implies that for this scaling procedure the influence of stimulus spacing is negligible. This is confirmed by the finding that fitting Eq. (6) to the data did not yield a significant relation between ''$J_{i,\mathrm{neg}}-J_{i,\mathrm{pos}}$'' and ''$F_{i,\mathrm{pos}}-F_{i,\mathrm{neg}}$'' ($r^2=0.02$). The resulting value of weighting factor w is almost one, namely 0.98 ± 0.03 (95% confidence interval: $0.92-1.05$). A similar result has been reported for simple stimuli provided a single perceptual dimension is involved in the judgments.[14,21]

It can be argued that the nonexistence of a context effect for comparison scaling is a trivial finding because the distribution of the differences in perceptual strength of blur is identical for the two stimulus sets. Fortunately, this can be checked as the differences in perceptual strength of blur S_i-S_j can be calculated for every combination of test image i and reference image j by means of Eq. (1). Suppose that difference judgment J_{ij} is related to S_i and S_j by the following equation:

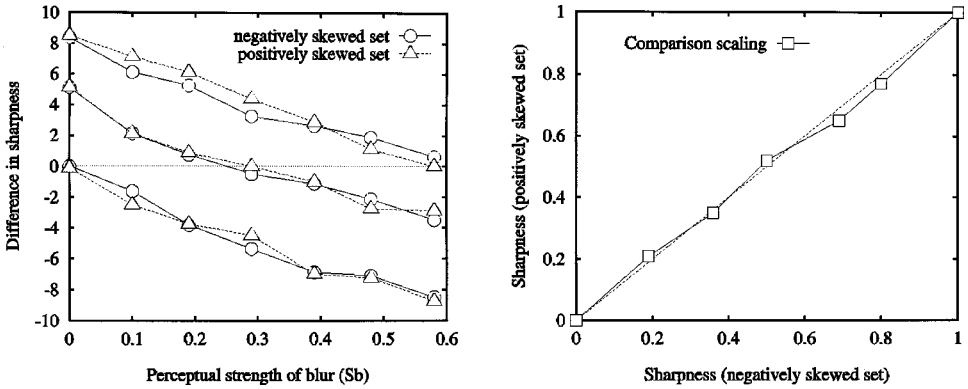

Fig. 3 Sharpness judgments obtained by comparison scaling for the images that were present in both the negatively and positively skewed stimulus set. Left-hand panel: Difference judgments for one subject (SY). Data are based on eight repetitions. The references are images with $S_b=0$ (lower curve), $S_b=0.29$ (middle curve) and $S_b=0.58$ (upper curve). The average standard error of the mean is 0.31 and 0.32 for the positively and negatively skewed set, respectively. Error bars have not been added because the standard errors of the mean are smaller than the size of the symbols in the figure. Right-hand panel: Sharpness judgments averaged across the curves for the three references and subjects. The judgments have been linearly transformed to a scale running from 0 to 1. Averaging across the curves for the three references results in estimates of sharpness that according to the additive functional measurement model of Anderson (see Ref. 35) represent perceived strength of sharpness but for a linear transformation. Dotted line indicates predicted results in the absence of contextual effects.

$$|J_{ij}| = w*(|S_i - S_j|)/0.58 + (1-w)*F_{ij}, \quad (7)$$

where $|J_{ij}|$ is linearly transformed to a scale running from 0 to 1 and F_{ij} is the frequency value based on the whole set of 66 test-reference pairs. Then weighting coefficient w can be estimated because the other parameters are known. Figure 4 presents the difference judgments as a function of $S_i - S_j$ for the test images belonging to both stimulus sets. For both conditions the judgments are linearly related to the calculated differences in blur implying no context effect due to stimulus spacing. This is confirmed by the estimated value of w: 0.95 ± 0.14 ($r^2 = 0.988$).

2.4 *Discussion*

The present study shows that (1) sharpness judgments obtained via comparison scaling are hardly influenced by the composition of the stimulus set, suggesting that for this procedure the influence of stimulus spacing is negligible, (2) there are no differences between the single and double stimulus method with respect to their sensitivity to contextual effects due to stimulus spacing, and (3) contextual effects observed with natural images are not as strong as those observed with simple stimuli. Recent experiments have extended these conclusions to other scenes and to colored images.

In 1997, a comparable experiment was carried out by a consortium of four laboratories (CCETT, France, SPTT, Switzerland, CRC, Canada, and IRT, Germany).[22] Their test material consisted of MPEG-2-coded video sequences at varying bit rates. Three ITU-recommended methods were tested, viz., double-stimulus continuous quality scale method (DSCQS), double-stimulus impairment scale method (DSIS), and comparison scaling.[8] The results agree with those of the present study in that a weak, almost neg-

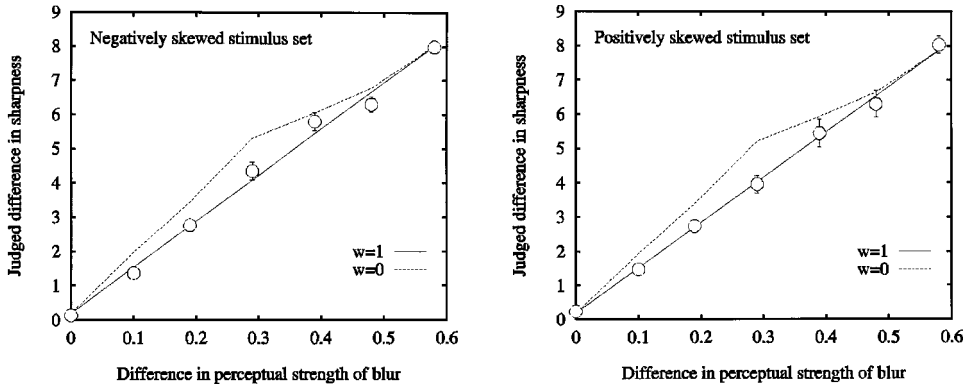

Fig. 4 Judged difference in sharpness as a function of calculated difference in perceptual strength of blur ($S_i - S_j$). Judgments have been averaged across subjects. Vertical bars denote twice the standard error of the mean.

ligible context effect ($w=0.89$) was found for the comparison scaling method and a much stronger effect ($w=0.73$) for the DSIS method. In contrast with the double-stimulus data in Fig. 2, no context effect was found for the DSCQS method.

A possible explanation for the last-mentioned result is the kind of scale employed in the DSCQS method: a 10 cm graphical scale divided into five equal intervals with descriptors (''bad,''..., ''excellent'') as labels of the categories. However, sharpness judgments obtained by single stimulus scaling using different rating scales (ten-point numerical category scale, five-point numerical category scale, five-point category scale with the Dutch equivalents of the ITU recommended descriptors as labels) suggest that the magnitude of the context effect due to stimulus spacing does not depend on the kind of response scale employed.[23] This conclusion can be generalized to graphical scaling; Schifferstein and Frijters[24] found no differences between a seven-point numerical scale and a line scale during the assessment of the sweetness of various solutions of sucrose. These results do not rule out the possibility that a combination of a graphical scale and certain descriptors, e.g., the ones employed in DSCQS but not the ones employed in DSIS, eliminates contextual effects due to stimulus spacing.

The finding that the number of categories has no effect on the influence of stimulus spacing agrees with experimental data gathered with simple stimuli.[14,15] These studies also demonstrated, however, that this insensitivity to the number of response categories relies on the kind of context effect involved. For example, the contextual effects due to the varying degree of occurrence of the stimuli were strongly influenced by the number of categories.[15] It is not known whether this also holds when complex stimuli like natural images are evaluated.

3 Experiment 2: Quality Versus Impairment Judgments

3.1 *Introduction*

Experiment 1 has shown that judgment strategies can influence the evaluation of the perceived quality of images comprising one varying attribute. It is reasonable to expect that this will also occur for images varying along multiple dimensions. For example, Boschman and Roufs[25] have found that individual differences in quality judgments of visual display units can be understood as a subject-dependent weighting of underlying perceptual attributes like sharpness and brightness contrast. Experiment 2 elaborated on this finding by asking subjects to evaluate both manipulated and standard JPEG-coded images. The JPEG-coding algorithm produces several artifacts at high levels of data compression, the most prominent ones being ''blockiness'' and ''ringing.'' [26,27] In the present study, the algorithm was manipulated such that the amount of blockiness could be varied independently of the amount of ringing.[28] The subjects were instructed to assess either the overall quality or the overall impairment of these ''manipulated'' images as well as standard JPEG-coded images. The observed differences between quality and impairment judgments could be interpreted as an instruction-based different weighting of the two artifacts.

3.2 *Method*

3.2.1 *Image material*

The image material consisted of JPEG-coded black-and-white versions of five pictures of natural scenes taken from a Kodak Photo CD test disk. These pictures were sampled with 8 bits/pixel on a grid of 512×480 pixels. During the experiment, only part of the pictures was used (240×480 pixels) in order to allow for the simultaneous display of two images on one screen. The goal of manipulating the JPEG coding was to vary independently the perceptual strength of two artifacts: blockiness (appearance of artificial rectangular blocks, especially in uniform regions) and ringing (appearance of light and dark artificial lines in the vicinity of edges). In standard JPEG coding[25] the perceptual strength of both artifacts decreases monotonically with the value of quality factor Q. This factor represents the degree of quantization of the 64 discrete cosine transform (DCT) coefficients $F(u,v)$ and varies between 0 (highest compression, lowest quality) and 100 (lowest compression, highest quality). Willemsen[28] observed that blockiness is caused mainly by quantizing the DCT coefficients at the lowest spatial frequencies (dc-component $F(0,0)$ plus low-frequency ac-components $F(1,0)$, $F(0,1)$ and $F(1,1)$, i.e., frequency bands 0 and 1 in Fig. 5). This observation led to the idea of creating images in which blockiness could be varied without introducing ringing by applying low Q factors to DCT coefficients in the three lowest frequency bands and high Q factors to the other DCT coefficients. In this study, the Q factor was set equal to 20, 30 or 40 for the three lower frequency bands and 80 for the other ones. These images will be referred to as the manipulated JPEG-coded images. The rest of the test images consisted of standard JPEG-coded images with Q factors equal to 20, 30 and 40. At these Q factors blockiness is the most prominent artifact. The reference was a JPEG-coded image with Q factor equal to 80. This image is almost free of distortions.

3.2.2 *Procedure*

The viewing conditions are comparable with those in experiment 1, except that the processed images were displayed on a calibrated 50 Hz noninterlaced BARCO CRT monitor and that the subjects were seated at a distance of about 0.9 m from the screen. At this distance, the pixel size is about 2 min of arc. This distance was chosen because it is the ''natural'' distance for watching JPEG-coded images on a monitor. The luminance of the adaptation field was 13 cd/m^2 being the average luminance of the test images. Half of the subjects evaluated overall quality and the other half overall impairment. In both conditions half of the subjects first performed a comparison scaling experiment and then a single stimulus scaling experiment, whereas the reverse holds for the other subjects. In both experiments quality/impairment was assessed on an 11-point numerical category scale ranging from 0 (no difference; lowest quality/impairment) to 10 (largest difference; highest quality/impairment). Per Q factor three stimulus pairs were created to be used in the comparison scaling experiment: (1) one pair of reference image (no distortions) and standard JPEG-coded image (blockiness plus ringing), (2) one pair of reference image (no distortions) and manipulated JPEG-coded

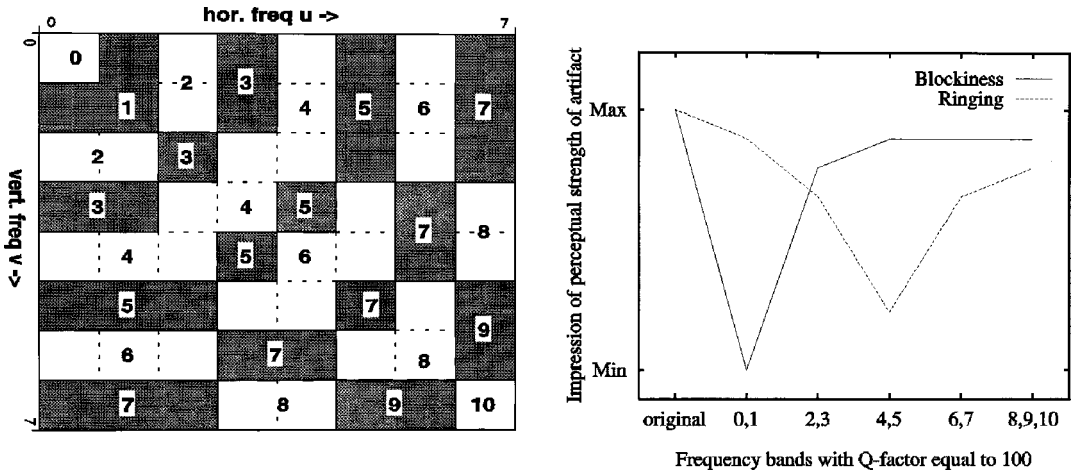

Fig. 5 Left-hand panel: DCT frequency bands. Right-hand panel: Impression of the perceptual strength of blockiness and ringing for images in which the *Q* factor was equal to 100 (i.e., size of the quantization step equal to one) for at least two frequency bands. For the remaining frequency bands the *Q* factor was equal to 20. Adapted from Willemsen (see Ref. 28).

image (blockiness only), and (3) one pair of standard JPEG-coded image (blockiness plus ringing) and manipulated JPEG-coded image (blockiness only).

3.2.3 *Subjects*

Twenty inexperienced subjects in the age from 18 to 28 years participated in the experiment. They had normal or corrected-to-normal vision. Their visual acuity measured with the aid of a Landolt chart at a distance of 5 m varied between 1.0 and 2.5.

3.3 *Results and Discussion*

Based on research in the field of human decision making, Willemsen[28] hypothesized that quality judgments are determined by the most prominent attribute (blockiness) and that impairment judgments are based on all attributes (blockiness and ringing). A sensitive test of this hypothesis is to compare the difference judgments of pairs consisting of the reference image and standard JPEG-coded images with those of pairs consisting of the reference image and manipulated JPEG-coded images. Ideally, the difference of these differences should be equal to zero for quality (both differences are based on blockiness) and positive but diminishing with increasing *Q*-factor for impairment. In contrast with these predictions, Fig. 6 demonstrates that the difference of differences is always positive for impairment and quality. This implies that both quality and impairment judgments of JPEG-coded images are based on blockiness and ringing. However, the influence of ringing is larger for impairment than for quality when subjects started with comparison scaling (Fig. 6, left-hand panel). This can be interpreted as an instruction-based different weighting of the two artifacts. The observed difference in judgment strategy disappears when subjects begin with single stimu-

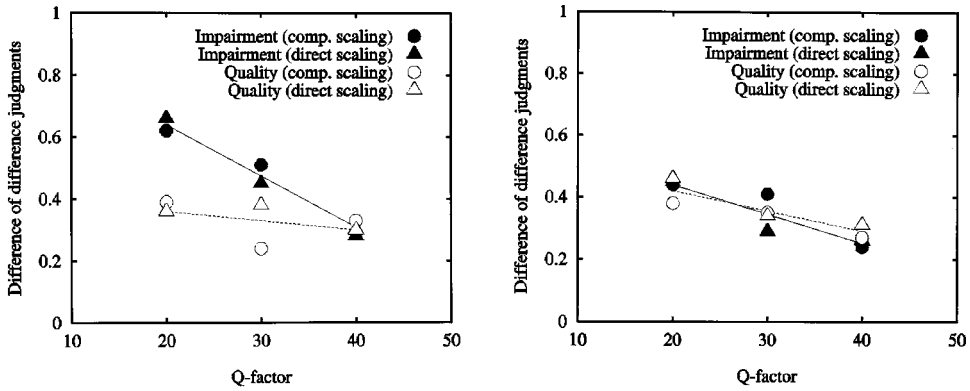

Fig. 6 *Q* factor and the difference of difference judgments for perceived quality (open symbols) and impairment (filled symbols). Data have been averaged across scenes and subjects. Difference judgments were obtained either directly via comparison scaling (circles) or indirectly via subtracting the category ratings of the reference and test images (triangles). In each condition, measured/calculated difference judgments were divided by the largest obtained difference before determining differences of difference judgments. Left-hand panel: Comparison scaling followed by single stimulus scaling. The average standard error of the mean is 0.07. Right-hand panel: Single stimulus scaling followed by comparison scaling. The average standard error of the mean is 0.06.

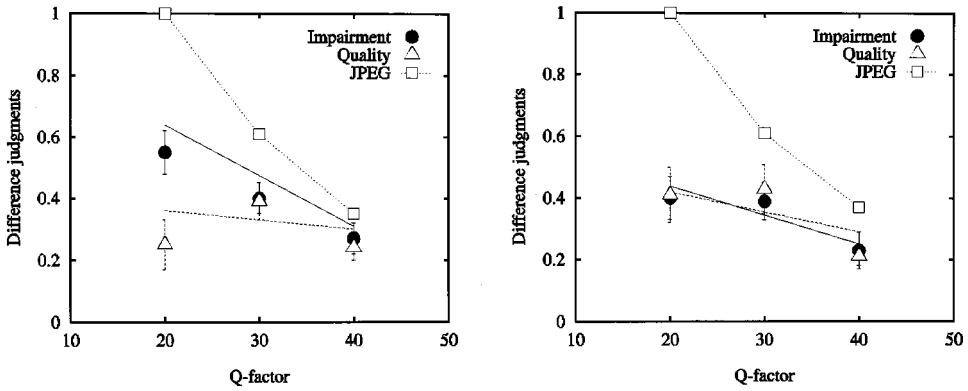

Fig. 7 Difference judgments of pairs consisting of standard and manipulated JPEG-coded images. Solid and dotted lines have been taken from Fig. 6. Vertical bars denote twice the standard error of the mean. For comparison, the difference judgments of pairs consisting of reference and standard JPEG-coded images have been included (open squares). Left-hand panel: Difference judgments measured before single stimulus scaling. Right-hand panel: Difference judgments measured after single stimulus scaling.

lus scaling (Fig. 6, right-hand panel). Interestingly, the judgment strategy induced by the first employed scaling procedure transfers to the second procedure. This process appears to be independent of the nature of the first scaling procedure.

In Fig. 7 the difference judgments of the stimulus pairs consisting of standard and manipulated JPEG-coded images are compared with the calculated differences shown in Fig. 6. The fair agreement between these two data sets implies that transitivity holds for comparison scaling. The results also suggest that ringing has a relatively strong influence on overall judgments, in particular during comparison scaling of overall impairment, despite the fact that blockiness is the most prominent artifact. This is in line with the results of scaling experiments with similar image material showing that at *Q* factors varying between 20 and 40 ringing is not the most prominent artifact but still an artifact that is difficult to neglect.[29]

4 General Discussion

The research described in this paper was meant to demonstrate the influence judgment strategies induced by experimental procedures can have on quality assessment. This was shown for contextual effects due to stimulus spacing (experiment 1) as well as for instructions given to subjects (experiment 2). In both cases a profound effect of scaling procedure was found. There was even an effect of the order in which scaling procedures are used. These findings clearly indicate that one should be cautious with interpreting quality judgments as a direct reflection of the viewers' quality impression.

One way to tackle this problem is by modeling judgment strategies. Recently, Hamberg and de Ridder[30] applied this approach to the overall judgments of MPEG-2-coded video material with time-varying image quality. In this case, the quality impression could be measured by means of a continuous assessment procedure. By this method, subjects continuously indicate the instantaneously perceived quality by moving a slider along a graphical scale.[13,16,31] Hence, the problem became the relation between the instantaneously perceived quality and the overall quality ratings. It appeared that this could be understood by assuming two judgment strategies, namely, a strong emphasis on the worst parts of a sequence and a recall advantage for the most recently presented material[3,32,33] (''recency effect''). These strategies were modeled by nonlinear averaging using a Minkowski-power weighting procedure[34] and an exponentially decaying weighting function. The good fit between model predictions and overall judgments supports the usefulness of this approach.

In conclusion, the research described in this paper underscores the important role judgment strategies play in the psychophysical evaluation of image quality. Ignoring this influence on quality judgments will undoubtedly lead to invalid conclusions about the viewers' impression of image quality. This implies that, in general, knowledge about judgment strategies is indispensable in designing and improving evaluation techniques.

Acknowledgments

The first experiment was carried out as part of the European RACE project MOSAIC (Methods for Optimization and Subjective Assessment in Image Communications) and the European ACTS project TAPESTRIES (The Application of Psychological Evaluation to Systems and Technologies in Remote Imaging and Entertainment Services). The author wishes to thank Martijn Willemsen and Lydia Meesters for their contributions to the second experiment.

References

1. H. Ohzu and K. Habara, ''Behind the scenes of virtual reality: Vision and motion,'' *Proc. IEEE* **84**, 782–798 (1996).
2. D. Harrison and N. Lodge, ''Broadcasting presence: Immersive television,'' in *Human Vision and Electronic Imaging V*, B. E. Rogowitz and T. N. Pappas, Eds., *Proc. SPIE* **3959**, 540–547 (2000).
3. D. E. Pearson, ''Viewer response to time-varying video quality,'' in *Human Vision and Electronic Imaging III*, B. E. Rogowitz and T. N. Pappas, Eds., *Proc. SPIE* **3299**, 16–25 (1998).
4. J. A. J. Roufs and H. Bouma, ''Towards linking perception research and image quality,'' *Proc. SID* **21**, 247–270 (1980).
5. J. W. Allnatt, *Transmitted-Picture Assessment*, Wiley, New York (1983).
6. J. A. J. Roufs, ''Perceptual image quality: Concept and measurement,'' *Philips J. Res.* **47**, 35–62 (1993).
7. B. E. Rogowitz, T. N. Pappas, and J. P. Allebach, ''Building bridges

between human vision and electronic imaging: A ten year retrospective,'' in Ref. 3, pp. 2–15.
8. ITU/R Recommendation BT.500-7, 10/1995. Internet address http://www.itu.ch/
9. W. Y. Zou, ''Performance evaluation: From NTSC to digitally compressed video,'' *SMPTE J.* **103**, 795–800 (1994).
10. J. C. Baird, *Sensation and Judgment: Complementary Theory of Psychophysics*, Erlbaum, Mahwah, NJ (1997).
11. E. C. Poulton, *Bias in Quantifying Judgments*, Erlbaum, Hove (1989).
12. G. E. Gescheider, ''Psychophysical scaling,'' *Annu. Rev. Psychol.* **39**, 169–200 (1988).
13. R. Hamberg and H. de Ridder, ''Continuous assessment of time-varying image quality,'' in *Human Vision and Electronic Imaging II*, B. E. Rogowitz and T. N. Pappas, Eds., *Proc. SPIE* **3016**, 248–259 (1997).
14. B. A. Mellers and M. H. Birnbaum, ''Loci of contextual effects in judgment,'' *J. Exp. Psychol.* **8**, 582–601 (1982).
15. A. Parducci and D. H. Wedell, ''The category effect with rating scales: Number of categories, number of stimuli, and method of presentation,'' *J. Exp. Psychol.* **12**, 496–516 (1986).
16. R. Hamberg and H. de Ridder, ''Continuous assessment of perceptual image quality,'' *J. Opt. Soc. Am. A* **12**, 2573–2577 (1995).
17. M. R. M. Nijenhuis and F. J. J. Blommaert, ''A perceptual error measure for sampled and interpolated complex color images,'' *Displays* **17**, 27–36 (1996).
18. M. R. M. Nijenhuis and F. J. J. Blommaert, ''Perceptual error measure for sampled and interpolated images,'' *J. Imaging Sci. Technol.* **41**, 249–258 (1997).
19. A. Parducci, ''Category judgment: A range-frequency model,'' *Psychol. Rev.* **72**, 407–418 (1965).
20. A. Parducci and L. F. Perrett, ''Category rating scales: Effects of relative spacing and frequency of stimulus values,'' *J. Exp. Psychol. Monograph*, **89**, 427–452 (1971).
21. H. N. J. Schifferstein, ''Contextual effects in difference judgments,'' *Percept. Psychophys.* **57**, 56–70 (1995).
22. ''Studies toward the unification of picture assessment methodology,'' ITU-R Report 1082-2, Annex 3 (1997).
23. H. de Ridder, ''Contextual effects in quality judgments,'' in Workshop on quality assessment in speech, audio and image communication, pp. 56–61, Darmstadt, Germany, March 11–13 (1996).
24. H. N. J. Schifferstein and J. E. R. Frijters, ''Contextual effects on judgments of sweetness intensity,'' *Percept. Psychophys.* **52**, 243–255 (1992).
25. M. C. Boschman and J. A. J. Roufs, ''Text quality metrics for visual display units: II. An experimental survey,'' *Displays* **18**, 45–64 (1997).
26. W. B. Pennebaker and J. L. Mitchell, *JPEG: Still Picture Data Compression Standard*, Van Nostrand Reinhold, New York (1993).
27. J. C. Russ, *The Image Processing Handbook*, 2nd ed., CRC, London (1995).
28. M. C. Willemsen, ''Subjective evaluation of JPEG-coded images: Quality versus impairment judgments,'' Master's thesis, Eindhoven University of Technology, IPO Report 1171 (1997).
29. H. de Ridder and M. C. Willemsen, ''Percentage scaling: A new method for evaluating multiply impaired images,'' Ref. 2, pp. 68–77.
30. R. Hamberg and H. de Ridder, ''Time-varying image quality: Modeling the relation between instantaneous and overall quality,'' *SMPTE J.* **108**, 802–811 (1999).
31. H. de Ridder and R. Hamberg, ''Continuous assessment of image quality,'' *SMPTE J.* **106**, 123–128 (1997).
32. R. Aldridge, J. Davidoff, M. Ghanbari, D. Hands, and D. Pearson, ''Measurement of scene-dependent quality in the assessment of digitally-coded television pictures,'' *IEE Proc. Vision Image Signal Process.* **142**, 149–154 (1995).
33. A. Aldridge, J. Davidoff, M. Ghanbari, D. Hands, and D. Pearson, ''Recency effect in the subjective assessment of digitally-coded television pictures,'' in *Proc. 5th IEE International Conf. on Image Processing and its Applications*, pp. 336–339, Edinburgh, UK (1995).
34. H. de Ridder, ''Minkowski-metrics as a combination rule for digital-image-coding impairments,'' in *Human Vision, Visual Processing, and Digital Display III*, B. E. Rogowitz, Ed., *Proc. SPIE* **1666**, 16–26 (1992).
35. N. Anderson, ''Functional measurement and psychophysical judgment,'' *Psychol. Rev.* **77**, 153–170 (1970).

Huib de Ridder holds his MSc degree in psychology from the University of Amsterdam and his PhD degree from Eindhoven University of Technology, The Netherlands. From 1982 until 1998, he has been affiliated with the Vision Group of the Institute for Perception Research (IPO), Eindhoven, The Netherlands, where his research focused on both fundamental and applied psychophysics. From 1987 until 1992 his research on the fundamentals of perceptual image quality metrics was supported by a personal fellowship from the Royal Netherlands Academy of Arts and Sciences. In November 1998, he was appointed associate professor of Informational Ergonomics at the Department of Industrial Design, Delft University of Technology, The Netherlands. His current research interests include image quality metrics, stereoscopic displays, augmented reality, user-product interaction, form perception, intention tracking in user-product interaction, integration of information streams, and interaction with embedded intelligence. Recently, he became full professor at the same department.