Quality assessment of coded images using numerical category scaling

A.M. van Dijk[†] J.B. Martens[†] A.B. Watson[‡]

[†]Institute for Perception Research
P.O. Box 513, 5600 MB Eindhoven, The Netherlands

[‡]NASA Ames Research center

Abstract

The large variety of algorithms for data compression has created a growing need for methods to judge (new) compression algorithms. The results of several subjective experiments illustrate that numerical category scaling techniques provide an efficient and valid way not only to obtain compression ratio versus quality curves that characterize coder performance over a broad range of compression ratios, but also to assess perceived image quality in a much smaller range (e.g. close to threshold level).

Our first object is to discuss a number of simple techniques that can be used to assess perceived image quality. We show how to analyze data obtained from numerical category scaling experiments and how to set up such experiments. Second, we demonstrate that the results from a numerical scaling experiment depend on the specific nature of the subject's task in combination with the nature of the images to be judged. As results from subjective scaling experiments depend on many factors, we conclude that one should be very careful in selecting an appropriate assessment technique.

1 Introduction

oday's large variety of algorithms for data compression has created a growing need for methods to judge (new) ompression algorithms. It is generally agreed, however, that objective error measures such as the commonly used ISE, do not always correlate well with subjective quality ratings. Although more advanced measures are being eveloped, experiments with subjects are still the only reliable way of determining the perceived quality of coded nages.

. well known problem in papers on image compression, especially when very high compression ratios are involved, the lack of quality assessment. Usually, bitrates and signal-to-noise ratios are specified, but subjective quality valuations are not carried out. Probably, researchers are unfamiliar with quality assessment techniques or they onsider them to be too time-consuming. With this paper, we want to demonstrate that with limited additional fort, one can obtain valuable information on the subjective performance of image compression schemes.

1 this paper, we discuss a number of simple evaluation techniques that can be used to measure perceived image uality. We show how to analyze data obtained from such experiments and how to set up the experiments. In all our xperiments we use techniques that make use of numerical category scaling. Although past IPO research projects [7] ave indicated that numerical category scaling techniques provide an efficient and valid way to assess image quality, will be shown that the choice of an appropriate experimental technique is not always trivial.

'he subjective evaluation of a compression algorithm usually requires this algorithm to be compared with a 'standard' oder. In our presentation we compare a JPEG-based interpolative coding scheme with a standard JPEG baseline equential coder. The different nature of the artifacts introduced by these coders causes problems when evaluations re carried out using direct numerical category scaling. We will demonstrate this by using both a direct scaling method nd a scaling technique in which subjects have to determine quality differences between all possible combinations f coded images. The latter technique (based on Anderson's functional measurement theory) should be preferred in nis case.

0-8194-1802-1/95/$6.00

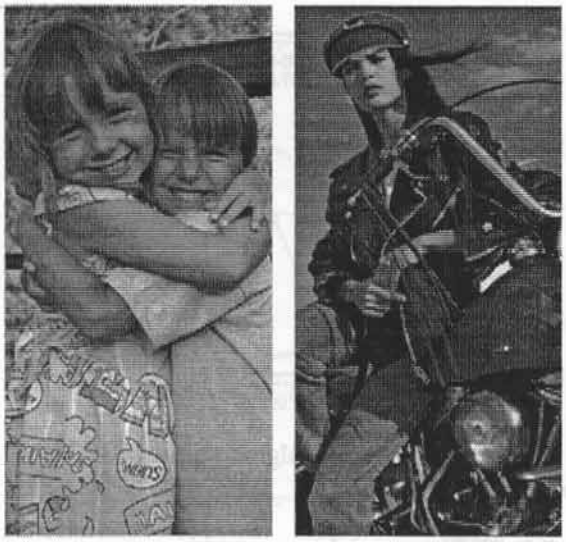

Figure 1: The original Kodak images 'girls' and 'motorbike'.

Figure 2: Fragments of 'girls' showing typical coding artifacts due to JPEG-DCT coding (C=28) and interpolative DCT coding (C=43).

All test scenes used in the experiments were acquired from a Kodak PhotoCD demonstration disc. The advantage of using these images is that they are varied and have high image quality. The original images with a resolution of 512x768 pixels were cut to 480x240 pixels in order to enable us to display two images simultaneously. For all experiments, we used 8-bit grayscale versions only. In figure 1, two of the original images ('girls' and 'motorbike') are shown.

In section 2, we present a JPEG-based interpolative image compression scheme. Subjective evaluations of this scheme are discussed sections 3, 4 and 5. In section 3, coder comparison experiments are carried out by means of direct numerical category scaling. In section 4, an assessment technique in accordance with Anderson's functional measurement theory is used. Both methods are discussed in some detail. In particular we will show how to average results over different subjects and how to interpret differences between subjects. In section 6, we show some results of subjective evaluations of a standard JPEG baseline sequential coder using scene-optimized quantization matrices.

2 JPEG-DCT versus Interpolative DCT coding

The first compression algorithm we chose to evaluate is the JPEG-based interpolative coding scheme proposed by Zeng and Venetsanopoulos in [11]. The idea of this scheme is as follows: First, the original image is low-pass filtered and subsampled. This subsampled image is then DCT-coded using a standard JPEG baseline sequential coder. The reconstructed image is obtained by DCT decoding and interpolation (upsampling and lowpass filtering). In our case we used a 2x2 average as the low-pass filter and a down-sampling factor of 2 in each direction to obtain the decimated image. A 4x4 binomial filter was used as the interpolator.

In order to test the interpolative scheme we used a standard JPEG baseline sequential coder as a reference [5]. As the interpolative coding scheme was designed to reduce blocking artifacts, the comparison was carried out with low quality images only (compression ratios approximately between 10 and 60). In figure 2, reconstructed fragments of the test image 'girls' are presented showing both the typical coding artifacts introduced by the reference coder (left image) and the impairments caused by the interpolative coder (right image). Note that JPEG-DCT coding typically introduces blocking artifacts, whereas the interpolative coding typically introduces unsharpness.

By scaling the standard DCT quantization matrix, we obtained a set of coded images for each coder. For each test scene, this resulted in twelve (2x6) coded images at various compression ratios. In figure 3, we plotted the PSNR versus the compression ratio curves obtained for two test images. Based on the PSNR performance curves, we found that the interpolative scheme typically outperforms the standard coder for compression ratios higher than 30. The

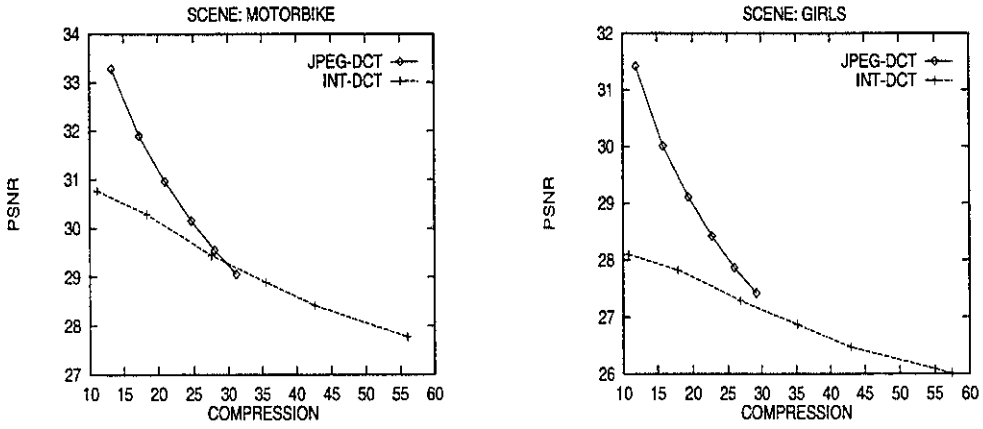

Figure 3: PSNR versus compression ratio curves for both compression algorithms. Results are shown for the test scenes 'girls' and 'motorbike'.

ne results were found in [11]. In order to get a more reliable impression of the coder performances, however, we :ried out a comparison based on subjective quality judgements. This subjective evaluation, using two different :egory scaling techniques, is discussed in the next three sections.

3 Quality assessment I (direct numerical category scaling)

ıe assessment of perceived image quality is often done by means of direct category scaling. In a category scaling periment, subjects are asked to classify images into a number of categories. For the measurement of perceived age quality, for example, the CCIR [2] has recommended a 5-point scale using the adjectives *bad*, *poor*, *fair*, *good* d *excellent* . In a similar way, descriptive impairment and comparison scales have been defined.

ı alternative way of identifying the points on a category scale is to label each category with a numerical value. It s been argued [6] that the use of numerical category scales is much more convenient as these scales are linear and n be easily adapted to specific ranges of image qualities. In this section we will use a direct numerical category ıling experiment to compare the performance of the image coders mentioned in section 2. We show how to set up e experiments and discuss how to analyze the data obtained from such experiments.

[ethod

ıbjective quality evaluations were carried out for the test scenes 'girls' and 'motorbike'. For stimulus display, a Hz non-interlaced BARCO CCID7351B monitor was placed in a dark room with a white, dimly lit 2.5 cd/m2 ckground. The subjects observed the coded images from a distance of 160 cm, corresponding to a ratio of 6:1 tween viewing distance and monitor height. All viewing conditions satisfied CCIR recommendation 500-3 [2]. On e monitor, each stimulus occupied an area of 25 x 12.5 cm, resulting in a horizontal size of 4.5 degrees of visual gle and a resolution of 53 pixels per degree of visual angle.

vo male and two female subjects between 24 and 28 years of age participated in the experiments. All subjects had perience with quality evaluations of processed natural images. They had normal or corrected-to-normal vision and visual acuity, measured on a Landolt chart, between 1.5 and 2.0. The subjects did not have any knowledge about e way the presented images were processed.

one session, all 12 reconstructed images of one test scene were presented 4 times in a random sequence (separate enes were handled in separate sessions). All 48 stimuli were displayed on the monitor during 5 seconds, followed by 11 cd/m2 adaptation field which lasted at least 2 seconds. The subjects were asked to rate the overall perceived ıality of each image using a numerical quality scale ranging from 1 to 10. Before starting a session, subjects had judge a training sequence in order to get an idea of the quality range of the images. This randomly displayed aining sequence contained all 12 coded images.

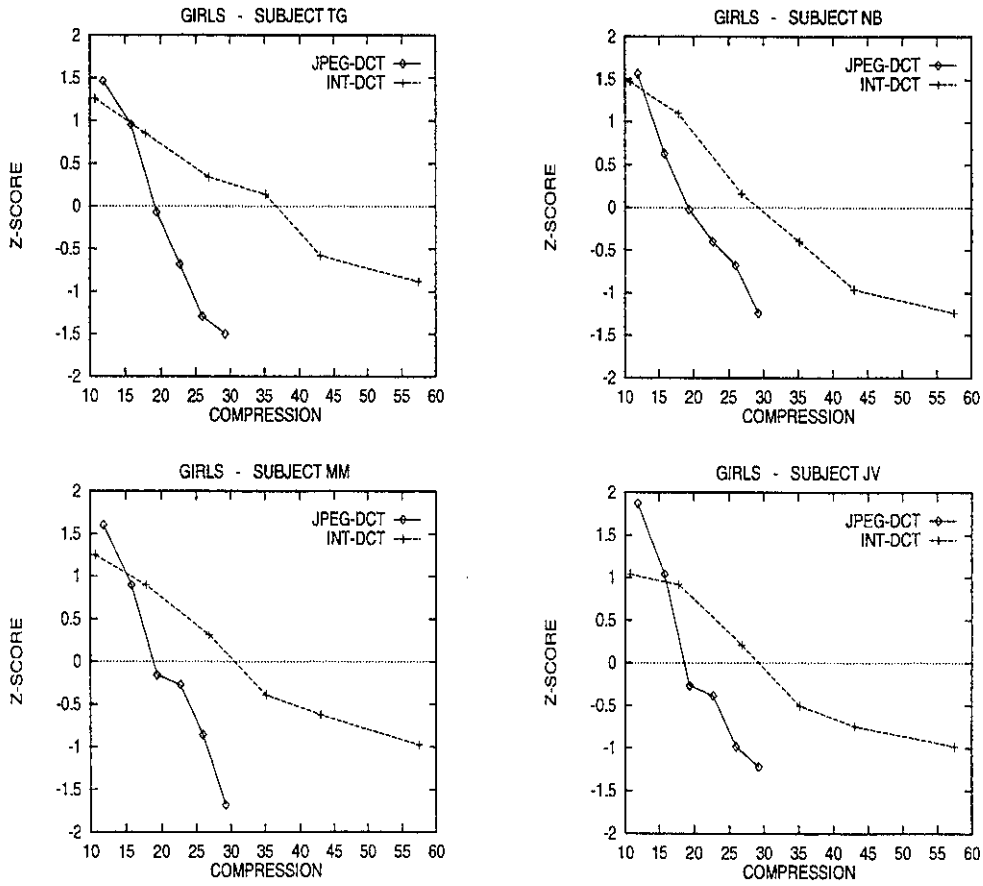

Figure 4: Individual z-scores, indicating the perceived image quality of the scene 'girls' as a function of the compression ratio.

Z-score results

Quality scores for each image were obtained by taking the mean of the four scores corresponding to that particular scene. Before averaging over all subjects, however, it is common to minimize the variation between the individual quality scores. This variation is caused by the fact that not all subjects use the full range of the numerical scale in classifying images. A simple method that normalizes individual scores is the z-score transform described by Hays [4]. This transform converts each score into a z-score that indicates the deviation from the mean score in standard deviation units (see appendix for the exact definition of z-scores). In figure 4 the z-scores of all individual subjects are shown for the scene 'girls'. For the 'motorbike' scene, similar results were obtained.

In figure 5, the averaged z-scores are plotted for both test scenes. Before being plotted, the averaged z-scores were rescaled so that a unit on the quality scale equals the average standard error of the mean. Comparing these results with the PSNR curves of figure 3 shows that for both scenes the equal-performance points (C=16 and C=15 for the scenes 'motorbike' and 'girls' respectively) have shifted considerably. From these plots, we would conclude that the interpolative coding scheme performs even better than could be expected based on the PSNR curves. In section 4, however, we will demonstrate that these results are heavily biased because of the difference in nature of the artifacts present in the reconstructed images.

Although averaging over individual z-scores does make sense, it implies that differences between individual subjects are not explicitly taken into account. Not only are all subjects given the same 'weight', averaging over subjects also removes all information about differences that exist between subjects. In section 4, we will see that these differences can be significant and meaningful. Therefore, we stress that analysis and interpretation of individual differences should always be part of a subjective evaluation experiment. In the next paragraph, we discuss a method

· data analysis based on Thurstone's category scaling model. This Thurstone scaling method does not only ease the erpretation of individual differences, but also provides a more perceptual view on the process of category scaling.

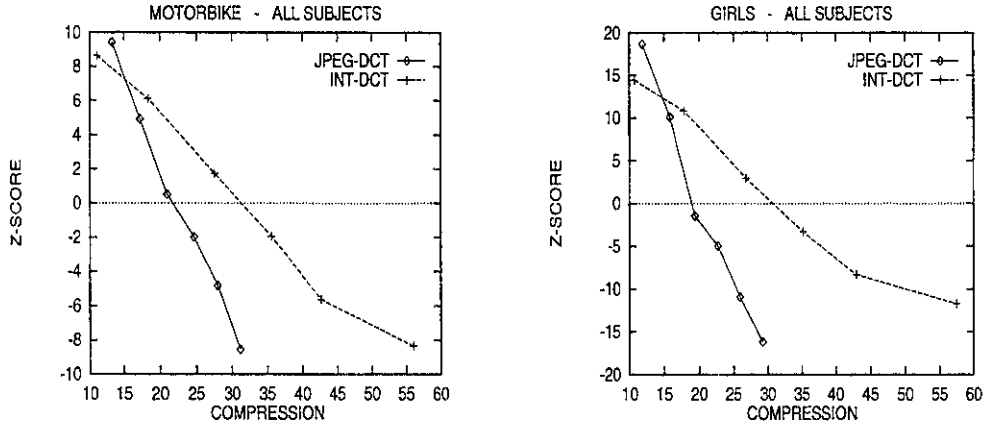

Figure 5: Perceived quality versus compression ratio curves for the two coders. The results are obtained by averaging over all individual z-scores and rescaling.

hurstone's law of categorical judgement

sults of numerical category scaling are not always presented in a raw (or normalized) form. An alternative way analyzing and presenting data from scaling experiments is to transform the data into an interval scale on a ychologically linear continuum. A technique for such a transformation, based on Thurstone's law of categorical dgement, is described by Torgerson in [8].

ıe ideas underlying Thurstone's law of categorical judgement can be describes as follows. Thurstone assumes that ery stimulus generates some momentary impression on a psychological continuum that is divided into a specified mber of ordered categories. The response to a given stimulus however is stochastic, as both the position of a mulus impression and the positions of the category boundaries are Gaussian distributed. In figure 6 this concept stimulus distributions and boundary distributions is visualized. Thurstone's law of categorical judgement now nsists of a set of equations that describe the judgement of an observer when asked to place stimuli into a number ordered categories.

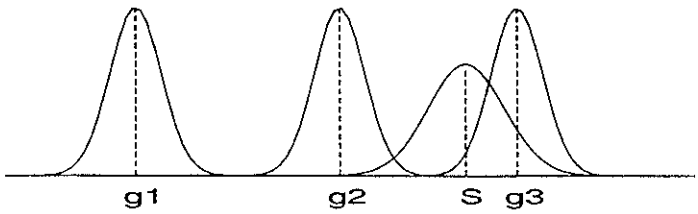

Figure 6: Position distributions on a psychological continuum for a stimulus S and category boundaries g1,g2 and g3.

nfortunately, the general law of categorical judgement is not solvable in its complete form, so applying this law ways requires additional constraints to be set. In this paper, we apply a model (Thurstone's model I) that involves plications over trials within one subject with constraints of condition D (for details see [8]). In condition D, the mber of parameters are limited by assuming that 1) the standard deviations of the category boundaries as well the stimulus locations are constant, and 2) the correlation between the momentary positions of a stimulus and a tegory boundary is also constant.

For all subjects, the raw scaling data obtained in the previously described experiment were transformed to a Thurstone scale using the in-house software package TurcatD. The output of this program includes estimates of the mean stimulus positions, the standard deviation of the stimulus distributions and some stress value indicating the goodness of fit of the scaling data (the standard deviation of the category boundaries is set to zero). As the experiment involved replications of stimuli over one subject (Torgerson calls this a model I experiment), each psychological scale calculated, reflects the behaviour of one particular subject for one particular test scene. The estimated stimulus standard deviation for a subject can be regarded as a measure for the consistency of that subject. It should be noted, however, that this standard deviation is also influenced by the range of the numerical scale values used by that subject (some subjects do not use the full range of the category scale).

As the psychological scale constructed by Thurstone scaling is considered to be a true interval scale, equal differences in the percept judged are reflected by equal distances on the scale. Direct interpretation of individual results, however, is difficult as it should be noted that a Thurstone scale is determined but for a linear transformation, and thus offset and scale can be chosen freely. In order to ease interpretation of individual subject behaviour, we normalized all Thurstone scales by subtracting the mean value of the Thurstone scores and rescaling such that one unit on the quality axis corresponded to the estimated stimulus standard deviation of a subject.

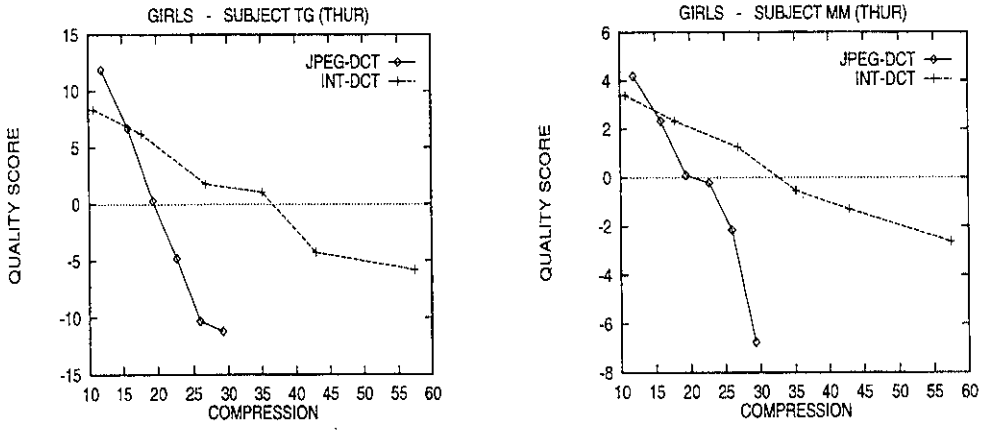

Figure 7: Normalized Thurstone scores expressed in stimulus standard deviation units for the subjects TG and MM. Results are shown for the test scene 'girls'.

On such a normalized psychological scale, differences between observers become clear. In figure 7, normalized individual results for subjects TG and MM are shown for the 'girls' scene. The difference in estimated stimulus standard deviation for the subjects TG and MM (0.244 and 0.375 respectively) is reflected by the larger range of quality scores calculated for subject TG. These standard deviations are influenced by scaling range and consistency of the subject. It can be seen that there is not much difference between the specific shapes of the curves obtained for both subjects.

Averaging over quality scores expressed in units of standard deviation implies that each subject is assigned a different weighting factor. The higher a subject's standard deviation, the smaller the weighting factor assigned. This corresponds to the idea that more consistent subjects are more reliable and should be assigned a higher weighting factor. Figure 8 shows this 'weighted' average of Thurstone scaling results for both test scenes. It can be seen that compared to the performance curves from figure 5, some small changes have occurred (especially for the 'motorbike' scene). Differences, however, are not very significant. In the next section, instead of using direct numerical category scaling, we will demonstrate a second quality evaluation method, based on functional measurement theory.

4 Quality assessment II (functional measurement theory)

The method of direct numerical category scaling has proved to be an efficient tool for image quality assessment [7]. One may doubt, however, the validity of quality scores derived from a direct scaling experiment when comparing the performance of two coders that cause artifacts of a very different nature. In image quality research, it is well known

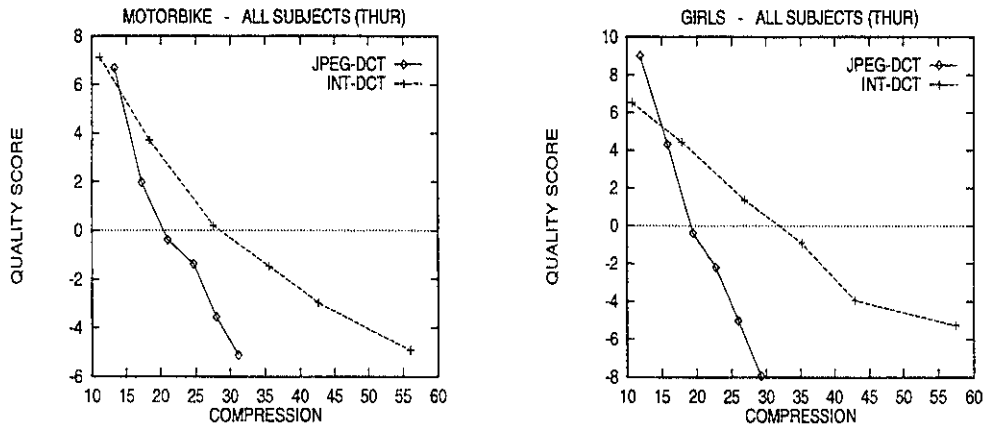

Figure 8: Averaged Thurstone scaling results indicating overall perceived quality of the test scenes 'motorbike' and 'girls'.

at if in a session of a direct category scaling experiment multiple scenes have to be judged, subjects tend to use a parate scale for each of those scenes. It is possible, however, that in the case where only one scene, but multiple pes of distortion are presented, subjects also tend to built up separate internal quality scales. If so, one would no ıger be able to compare direct quality ratings obtained for two different coders.

possible solution to this problem would be to force subjects to link the quality ratings for both coders. An approach at achieves this, is offered by Anderson's Functional Measurement Theory (FMT) [1]. In this approach, image alities are compared rather than separately evaluated. Although functional measurement theory was developed as broad framework for efficient description of non-observable psychological processes underlying the comparison of muli, it can be used quite efficiently for quality assessment of coded images [6].

[ethod

; functional measurement is rather time-consuming, and no significant differences were found between the test ɛnes, we only used the 'motorbike' scene in this experiment. The subjects were the same that participated in e direct scaling experiment. No changes were made in the viewing conditions, except for the fact that now each .mulus contained two images, simultaneously displayed on the left and on the right hand sides of the screen, with ɜpacing of 30 pixels (about 1.5 cm). The stimulus presentation time was extended to 6 seconds in order to give the bjects enough time to observe both images.

ll 12 stimuli (6 for each coder) were factorially combined to form 144 stimulus pairs. Subjects were asked to rate e difference in quality between the two images using a scale ranging from -10 to +10. The plus and the minus sign ɛre used to indicate whether the left or the right image was preferred. Again, a training session of 12 randomly lected image pairs was presented to each subject before the start of the actual experiment.

xperimental results

ɔr each subject, one 12x12-element matrix was obtained (one row and one column per stimulus), with elements $,j)$ representing the score given by the subject for the difference in quality between the pair of stimuli, stimuli i ıd j being displayed on the left and the right hand sides of the screen, respectively. In figure 9 the results averaged 'er all subjects are shown. It can be observed that approximately, the scores within the different rows and columns rm parallel curves.

he interaction between rows and columns has been examined statistically by means of two-way analysis of variance. ɔ significant interaction was found ($F_{121,288} = 0.87, p > 0.05$). According to FMT, it is now possible to determine quality score for each stimulus by averaging (with opposite sign) the row and column means of the matrix that ɪrrespond to that stimulus. This was done for each stimulus. A general quality score for all subjects can now be

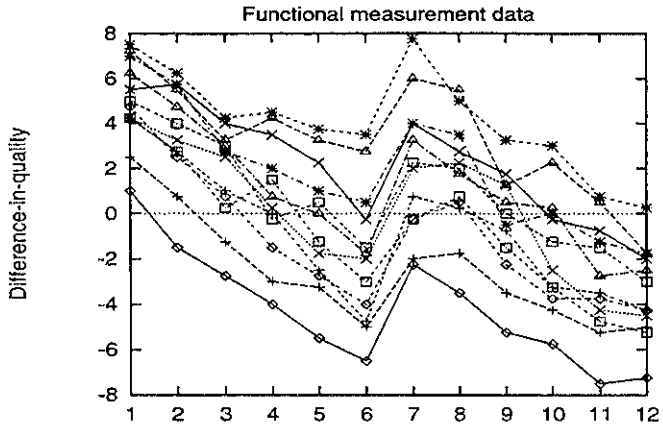

Figure 9: Difference-in-quality scores, averaged over all subjects. Note the parallelism of the curves.

obtained by averaging the individual quality scores. Before averaging, the individual scores were normalized using a z-score transform [4] (see appendix). The right part of figure 10 shows the averaged quality curves.

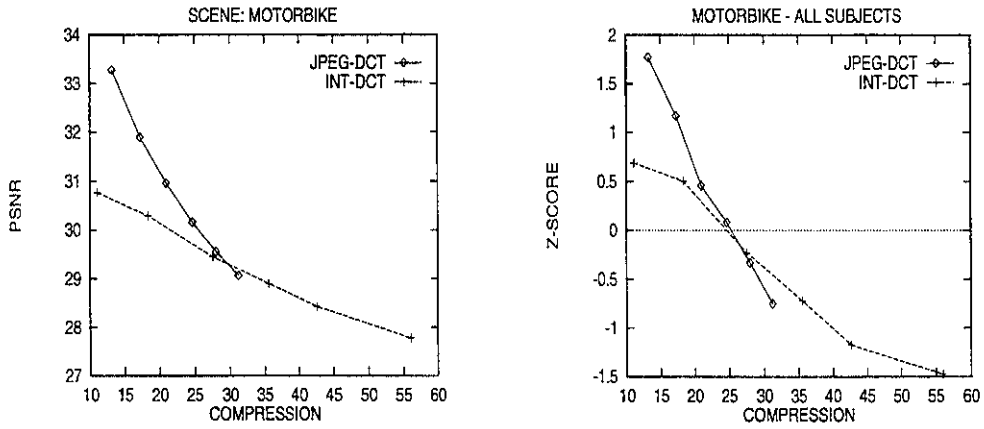

Figure 10: PSNR curves (left), and subjective quality curves obtained with Anderson's FMT.

Compared to the PSNR curves, the equal-quality point has shifted slightly towards the lower compression ratios. This shift, however, is much smaller than the effect that was found in the direct category scaling experiment. In table 1, the equal-performance points of the individual subjects are given for both the direct scaling experiment and the comparison-based experiment. It can be observed that these individual data all show the same effect: equal-performance points found for the two methods, differ significantly. There are also remarkable differences between the subjects. For instance, it is clear that some subjects (mm and tg) dislike blocking artifacts more than others. This is reflected by the much lower equal-performance points found for these subjects. Also, notice the similarity between the calculated PSNR curves and the quality curves obtained with the functional measurement method (figure 10). For the coded images used in this experiment, the PSNR performs quite well.

5 Functional measurement theory versus direct category scaling

The experimental results clearly show that in our case, both methods do not lead to comparable results. The problem is that reconstructed images from the two coders can be easily identified. This stimulus identification directly leads to problems in the sense that observers will use separate scales. By comparing figures 5 and 10, it can be seen that

	Compression at equal-quality point	
Observer	Direct Category Scaling	Functional Measurement Theory
mm	C=12	C=18
tg	C=12	C=19
nb	C=19	C=35
jv	C=18	C=28
average	C=16	C=27

Table 1: Equal-performance points for the 'motorbike' scene.

ı figure 5, for both test scenes, the quality curves are 'stretched' to fit into the full quality range. We showed that ıith Anderson's functional measurement theory, subjects can be forced to link both scales. Therefore, we think that irect scaling should not be used in a coder comparison experiment if the reconstructed images from the various ɔders can be easily identified.

)f course, this does not mean that the functional measurement based method should always be preferred. De Ridder nd Majoor [6], for instance, carried out some experiments to asses impairment of perceptual image quality due to uantization errors in a scale-space coding algorithm. They found that simple direct category scaling and the scaling rocedure in accordance with functional measurement theory gave rise to the same functional relationship between erceptual image quality and quantization error. This means that direct category scaling is very well applicable, for ıstance, in experiments in which different parameters of one specific coder are varied.

Ve already mentioned that direct scaling experiments are much less time-consuming than complete comparison ex-eriments. An intermediate solution when comparing different coders would be to perform a direct scaling experiment ɔ obtain several 'separate' scales, and to link these scales by carrying out an additional comparison experiment on a nall subset of the stimuli. All scales previously obtained by direct category scaling can then be linearly transformed ɔ that they can be compared to each other.

6 JPEG coding using optimized quantization matrices

ı JPEG-DCT coding, quality and bitrate of a compressed image are determined by the quantization matrix which efines the reconstruction levels for all transform coefficients. Usually, quality and bitrate are varied by scalar ıultiplication of some 'standard' JPEG quantization matrix. Proper design of the quantization matrix, however, an improve the performance of a DCT coder significantly. In this section, we will show some results of an experiment ı which the effect of scene-optimized quantization matrices on perceived image quality is evaluated. Our main goal ; to demonstrate that scaling experiments can also be carried out when differences between reconstructed images re fairly small. All scene-optimized matrices we used were produced by Watson's 'DCTune' technique [9, 10].

)ptimized quantization matrices can be constructed by applying models that describe the visibility of quantization rtifacts. The DCTune matrices calculated for our experiments were designed using a model which incorporated isibility of DCT frequencies, effects of display resolution and luminance characteristics, frequency summation, spatial ummation, and contrast masking. Although we only calculated optimized matrices for grayscale images, it should e noted that the DCTune technique has also been extended to color images [10]. In figure 12, both the standard PEG quantization matrix and a scene-optimized DCTune matrix for the test scene 'girls', are given. This latter ıatrix was designed for a bitrate of approximately 1 bit/pixel, a display resolution of 34 pixels per degree of visual ngle, and a mean luminance of 17 cd/m2.

or four test scenes, DCTune matrices were calculated for two different bitrates. One matrix for high-quality com-ression (about 0.8 - 1.0 bits/pixel) and one matrix for lower-quality compression (about 0.4 - 0.5 bits/pixel). All est scenes were then compressed with a standard JPEG baseline sequential coder using these DCTune matrices. ubsequently, for each coded image, five reference images were generated using quantization matrices, obtained by calar multiplication of the standard JPEG quantization matrix. The reference images were generated so that one f the reference images matched the compression ratio of the DCTune-coded image. In figure 12, PSNR versus ompression ratio curves in the lower quality range are plotted for all test scenes. It can be seen that for all scenes, ccording to the PSNR measure, the standard JPEG matrix outperforms the DCTune matrix.

16	11	10	16	24	40	51	61
12	12	14	19	26	58	60	55
14	13	16	24	40	57	69	56
14	17	22	29	51	87	80	62
18	22	37	56	68	109	103	77
24	35	55	64	81	104	113	92
49	64	78	87	103	121	120	101
72	92	95	98	112	100	103	99

11	10	11	17	28	63	143	255
10	13	12	15	22	41	93	255
11	12	24	30	41	69	178	255
17	16	31	60	85	145	255	255
31	24	46	103	255	255	255	255
65	47	78	163	255	255	255	255
169	95	138	255	255	255	255	255
255	255	255	255	255	255	255	255

Figure 11: Standard JPEG quantization matrix (right) and a DCTune matrix (left), designed for the test scene 'girls'. Notice that the DCTune matrix typically compresses the high frequency coefficients, while the DC coefficient is less severely quantized.

Method

In order to evaluate the perceptual quality of the DCTune-coded images experimentally, we used a so called 'constant stimulus' method. In this technique, stimuli are used that consist of two images: one image that occurs in every stimulus (in our case a DCTune-coded image), and one reference image (one of the corresponding JPEG-coded images). The first experiment we carried out involved the performance assessment of the DCTune matrices designed for the higher compression range. Each DCTune-coded image was used as an additional reference.

Except for the viewing distance and the stimulus presentation time, which were set to 100 cm and 9 seconds respectively, no changes were made in the viewing conditions. As each stimulus occupied an area of 25x26.5 cm, this resulted in 15 degrees of visual angle, and a ratio of 4:1 between viewing distance and stimulus height. The display resolution was 34 pixels per degree of visual angle. The mean luminance of the images displayed on the monitor, having a peak luminance of 60 cd/m2 and a gamma of 2.5, varied between 5 and 17 cd/m2.

Two male and two female subjects between 22 and 28 years of age participated in the experiments. All subjects had experience with quality evaluations of processed natural images. They had normal or corrected-to-normal vision, and a visual acuity, measured on a Landolt chart, between 1.5 and 2.0. Only one of the subjects had detailed knowledge of the set-up of the experiment and the kind of stimuli that were used.

The task of the subjects was to rate the difference in quality between the two displayed images, using a numerical category scale ranging from -5 to 5. The plus and minus sign indicated whether the left or the right image was preferred. The stimuli were presented in a random sequence, containing stimuli of all four test scenes. As each DCTune-coded image was combined with 6 reference images, and each stimulus combination was repeated 6 times, the number of image pairs to be compared in one session, was 144. The experiment was balanced in the sense that each DCTune-coded image occurred an equal amount of times on the left and on the right side of the screen. A training set of 16 randomly selected stimuli containing all test scenes, was presented to the subjects before the start of a session.

Results

For each of the test scenes, quality scores for the 6 reference images were obtained by taking the mean of the 6 scores indicating the distance between that particular image and the corresponding DCTune coded image. Before averaging over all subjects, these scores were z-score transformed to reduce the variation between the subjects. In figure 13, the averaged z-scores of all reference images are plotted. In each plot, the dotted horizontal line indicates the quality of the corresponding DCTune-coded image. As expected, the qualities rated for the DCTune-coded reference images, almost coincide with these lines.

The z-score plots clearly indicate that the rate/distortion performance of a DCT coder can indeed be improved by designing proper quantization matrices. For the test scene 'girls', the perceived quality of the DCTune-coded image exactly matches the quality of the best reference image. For the other test scenes, the perceived image quality of the DCTune-coded image even exceeds the quality of the best reference image. This means that it would have been better to select reference images in an even higher quality range. From these plots, one may additionally conclude that subjects are very well capable of rating images in a small quality range.

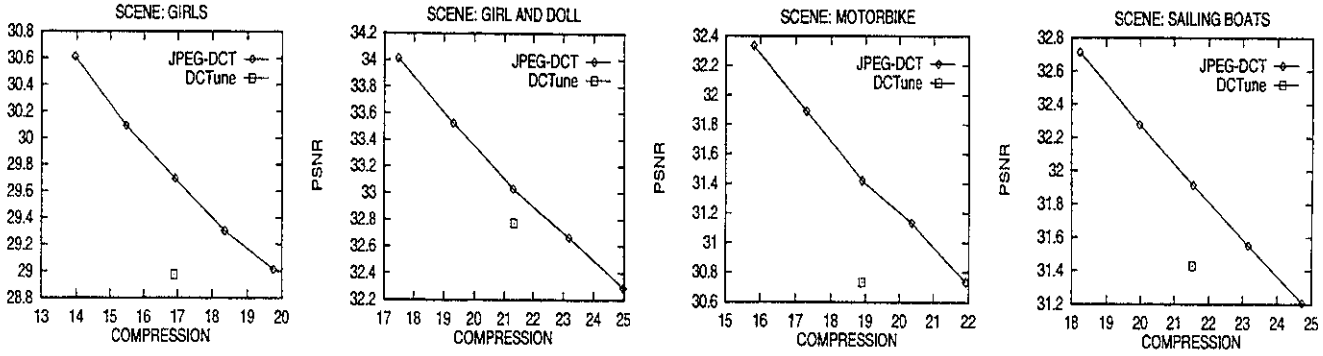

Figure 12: Compression ratio versus PSNR plots for all reference images.

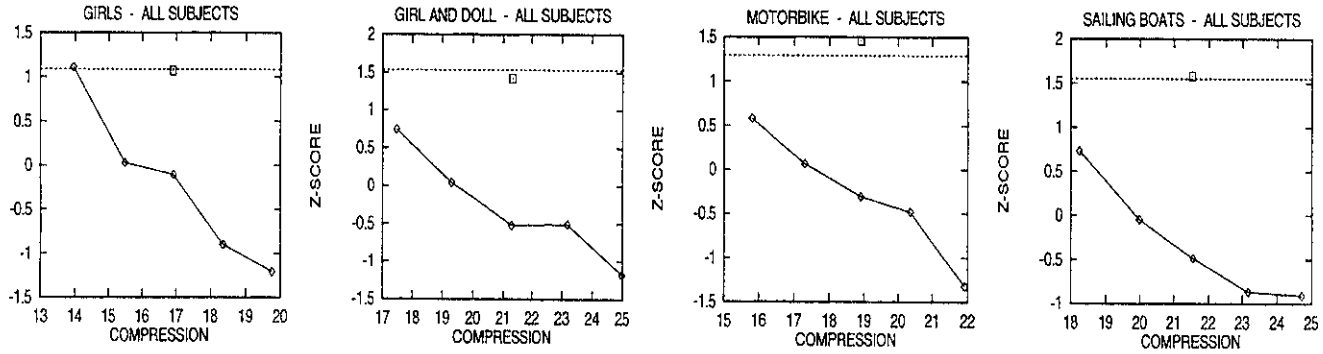

Figure 13: Averaged subjective quality ratings expressed in z-scores. The dotted lines indicate the quality of the DCTune-coded scenes.

similar comparison experiment was carried out using the DCTune matrices calculated for the high quality range. igure 14 shows both the PSNR's (left) and the subjective quality ratings (right) obtained for the test scene 'girl ıd doll'. The perceived image quality - expressed in z-scores - was averaged over 3 subjects. It can be observed ıat for these subjects, the effect of using a scene-optimized quantization matrix was found to be very small. One of ıe subjects clearly preferred the DCTune-coded image. Similar results were found for the other three test scenes.

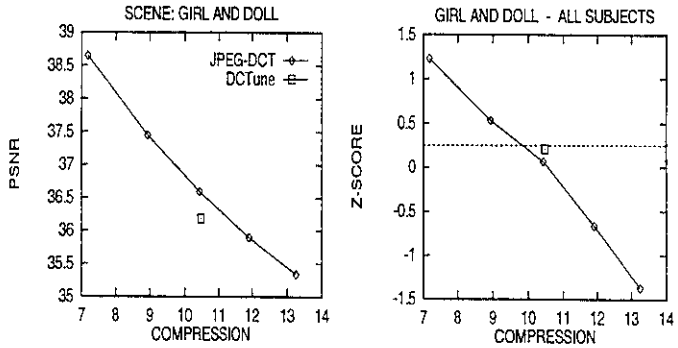

Figure 14: Averaged subjective quality ratings expressed in z-scores. Results are shown for for the test scene 'girl and doll'.

7 Conclusions

In this paper, we described the set-up of a number of experiments in which subjects had to judge image quality using numerical category scaling. By evaluating the performance of an image coder by using both a direct category scaling method and a scaling method based on Anderson's functional measurement technique, we showed how to analyze and interpret experimental data. In particular, we showed that results from direct category scaling experiments can be biased when image coders are evaluated that cause artifacts of a very different nature. Finally, by assessing the effect of using scene-optimized quantization matrices on the quality of JPEG-coded images, we demonstrated that category scaling can also be used if differences in image quality are fairly small.

Appendix

Each score obtained in a category scaling experiment can be converted into a standardized score, or z-score, expressing the deviation from the mean in standard deviation units [4]. The z-score z_j for a score x_j is given by

$$z_j = \frac{x_j - \overline{x}}{\sigma}$$

The z-score thus tells how many standard deviations away from the mean is x_j. The mean and standard deviation are given by:

$$\overline{x} = \frac{1}{N}\sum_{j=1}^{N} x_j \qquad \text{and} \qquad \sigma^2 = \frac{1}{N-1}\sum_{j=1}^{N}(x_j - \overline{x})^2$$

respectively. Notice that z-scores are normalized according to the mean and standard deviation of one particular subject.

References

[1] N.H. Anderson, Algebraic models in perception. *Handbook of perception, Voll.II: Psychophysical Judgement and Measurement* (E.C Carterette and M.P. Friedman, Eds.), pp 216-298, Academic Press, New York (1974).

[2] CCIR, Method for the subjective assessment of the quality of television pictures, Recommendation 500-3, In: *Recommendations and Reports of the CCIR*, International Telecommunication Union, Geneva, 1986.

[3] C. Chatfield, *Statistics for technology*, third edt., Chapman and Hall, London (1983).

[4] W.L. Hays, *Statistics*, fourth edt., Holt, Rinehart and Winston Inc., New York (1988).

[5] W.B Pennebaker and J.L Mitchell, *JPEG still image data compression*, Van Nostrand Reinhold, New York (1993).

[6] H. de Ridder and G.M. Majoor, Numerical category scaling: an efficient method for assessing digital image coding impairments. *Proceedings of the SPIE Conference on Human Vision and Electronic Imaging: Models, Methods, and Applications, SPIE-1249, Santa Clara, California,1990*, pp 65-77.

[7] J.A.J Roufs, Perceptual image quality: concept and measurement, Philips Journal of research, vol 47(1), pp 35-62 (1992).

[8] W.S. Torgerson, *Theory and methods of scaling*, John Wiley & Sons, New York (1958).

[9] A.B. Watson, DCT quantization matrices visually optimized for individual images, Human Vision, Visual Processing, and Digital Display IV. Rogowitz ed. 1993 SPIE. Bellingham, WA.

[10] A.B. Watson, Perceptual optimization of DCT color quantization matrices, Proc. ICIP-94, vol I, pp 100-104.

[11] B. Zeng and A.N. Venetsanopoulos A JPEG-based interpolative image coding scheme, Proc. ICASSP 1993, vol V, pp 393-396.