Minkowski-metrics as a combination rule for digital-image-coding impairments

Huib de Ridder

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

ABSTRACT

The urge to compress the amount of information needed to represent digitized images while preserving perceptual image quality has led to a plethora of image-coding algorithms. At high data compression ratios, these algorithms usually introduce several coding artifacts, each impairing image quality to a greater or lesser extent. These impairments often occur simultaneously. For the evaluation of image-coding algorithms, it is important to find out how these impairments combine and how this can be described. The objective of the present study is to show that Minkowski-metrics can be used as a combination rule for small impairments like those usually encountered in digitally coded images. To this end, an experiment has been conducted in which subjects assessed the perceptual quality of scale-space-coded color images comprising three kinds of impairment, viz., 'unsharpness', 'phantoms' (dark/bright patches within bright/dark homogeneous regions) and 'color desaturation'. The results show an accumulation of these impairments that is efficiently described by a Minkowski-metric with an exponent of about two. The latter suggests that digital-image-coding impairments may be represented by a set of orthogonal vectors along the axes of a multidimensional Euclidean space. An extension of Minkowski-metrics is presented to generalize the proposed combination rule to large impairments.

1. INTRODUCTION

Interest in digital imaging systems is growing at a rapid pace. The reason for this interest is obvious: representing images in digital form allows visual information to be easily manipulated in useful and novel ways[1]. A potential problem with digital images is the large number of bits needed to represent them. For example, a black-and-white image sampled on a grid of 512 by 512 pixels with 8 bits/pixel requires already more than 2×10^6 bits. For storage and transmission of digital images it is often necessary to compress this amount of information considerably. All existing coding techniques aimed at compressing the amount of information needed to represent an image can do so only at the expense of distorting the image. At high compression ratios, distortions such as blurring or false contouring[2] degrade or impair perceived quality of a coded image. As most coded images are meant to be watched by human observers, these digital-image-coding impairments will play a crucial part in evaluating the applicability of a given coding algorithm.

In general, coding algorithms introduce several artifacts in an image. This is certainly valid for multiresolution coding techniques such as subband coding[3], Laplacian pyramid coding[4] or scale-space coding[5]. These techniques decompose the input image into different components that are coded and transmitted separately before being combined to reconstruct the image. Each component contains specific information about the input image. Hence, subsampling and/or gray-level quantization of different components may lead to different coding artifacts in the reconstructed image. Digital-image-coding impairments have been

 0-8194-0820-4/92/$4.00

found to accumulate when they occur simultaneously[2,6]. This implies that for the evaluation of image-coding algorithms it is not sufficient to know the perceptual consequences of each coding artifact separately. In addition, one needs to know how different impairments combine and how this can be described.

The objective of the present study is to show that Minkowski-metrics can be used as a combination rule for impairments, in particular for small impairments like those usually encountered in digitally coded images. Minkowski-metrics[7] have already been employed in many fields of human perception research. For example, as a distance function in multidimensional scaling[8] or as an efficient rule for combining responses from different spatial channels assumed to exist in the human visual system[9]. Results of recent experiments[6] suggest that Minkowski-metrics can also be used to describe the way digital-image-coding impairments are combined.

The present paper is organized as follows. First, Minkowski-metrics are introduced as a possible rule for combining small impairments (section 2). Subsequently, an experiment is described in which the applicability of the combination rule based on Minkowski-metrics is examined (section 3). In this experiment, subjects assess the perceptual quality of scale-space-coded color images[5] comprising one to three perceptually different impairments, viz., 'unsharpness' (blur), 'phantoms' (dark/bright patches within bright/dark homogeneous regions) and 'color desaturation'. Finally, an extension of Minkowski-metrics is presented to generalize the proposed combination rule to large impairments (section 4).

2. MINKOWSKI-METRICS AS A COMBINATION RULE FOR IMPAIRMENTS

The combination rule based on Minkowski-metrics is confined to impairments due to coding artifacts in appreciation-oriented images. In this case, numerical category scaling is the most efficient method to assess perceptual image quality[10]. The combination rule assumes a linear transformation of quality into impairment followed by a nonlinear addition of impairment scores. In a previous paper[6], the following equation has been proposed to derive impairment score I from average quality rating Q:

$$I = (Q_0 - Q)/(Q_0 - Q_{min}), \tag{1}$$

where Q_0 and Q_{min} are the highest and lowest quality observed in a rating experiment. Note that $Q_{min} \leq Q \leq Q_0$ and $0 \leq I \leq 1$. In general, Q_0 will be the quality of the original image comprising no coding artifacts. It is obvious that eq. (1) is not limited to images comprising only one impairment. It can also be used to transform quality rating Q_{tot} of an image with n impairments into overall impairment score I_{tot}.

A prerequisite for using Minkowski-metrics to combine impairments is that quality scales have the properties of at least an interval scale. Fortunately, this can easily be ensured by rescaling category ratings using models based on Thurstone's 'law of categorical judgment'[11]. The combination rule itself is given by

$$\hat{I}_{tot} = \left(\sum_{i=1}^{n} I_i^p \right)^{1/p}, \tag{2}$$

where $\hat{I}_{tot}$ denotes the predicted impairment score for an image with n impairments and I_i $(i = 1, \ldots, n)$ are impairment scores that have been derived from quality ratings of images comprising only one of these impairments. In this equation, exponent p is a free parameter. Recent results[6], however, suggest that this exponent will take only two values, viz., $p \approx 1$ when impairments are perceptually hardly distinguishable and $p \approx 2$ when impairments are perceptually clearly distinguishable. The former would imply linear addition, the latter nonlinear addition in accordance with Pythagoras' rule. These values of exponent p have

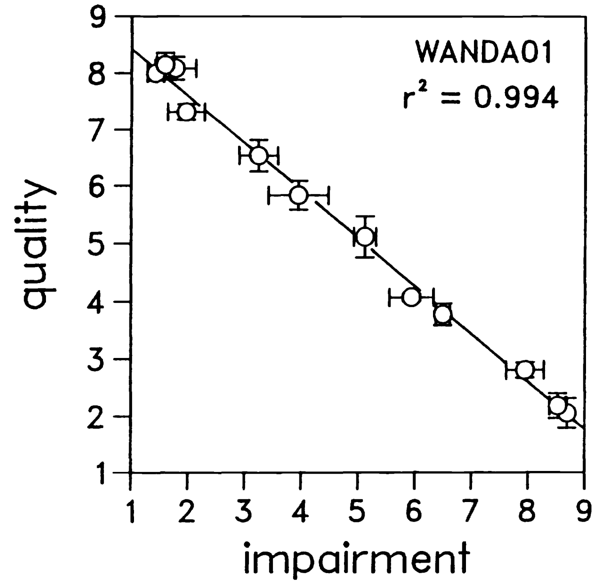

Figure 1: Perceptual quality as a function of perceived impairment. Stimuli are 12 scale-space-coded images of one natural scene: the portrait of a female model (WANDA01). Impairment has been introduced by systematically reducing the highest resolution information of the original image. Data have been taken from de Ridder[14]. In this and the following figures, bars denote twice the standard error of the mean.

been determined by means of a least-squares fitting procedure minimizing the sum of squared differences between predicted impairment scores $\hat{I}_{tot}$ of eq. (2) and the corresponding impairment scores I_{tot} obtained by putting quality ratings Q_{tot} into eq. (1). Whenever exponent p is known, predicted impairment scores $\hat{I}_{tot}$ can be transformed into predicted quality values $\hat{Q}_{tot}$ by means of eq. (1), leading to the following expression:

$$\hat{Q}_{tot} = Q_0 - \hat{I}_{tot}.(Q_0 - Q_{min}). \quad (3)$$

In general, impairment score $\hat{I}_{tot}$ will vary between zero and one. In extreme cases, however, it may become larger than one, leading to negative values of $\hat{Q}_{tot}$. In order to prevent this, an extension of eq. (2) is needed. A discussion on this topic can be found in section 4.

There is some experimental support for the assumption underlying eq. (1) that impairment and quality are linearly related[12,13]. More proof comes from an experiment that has recently been carried out at IPO[14]. In that experiment, subjects rated quality as well as impairment in a set of scale-space-coded images of one natural scene: the portrait of a female model (WANDA01). In this set of images, impairment was introduced by systematically reducing the highest resolution information of the original image. Experimental results are summarized in Figure 1 where perceptual quality has been plotted as a function of perceived impairment. Data show a linear relation between quality and impairment ($r^2 = 0.994$).

3. EXPERIMENTAL TEST OF MINKOWSKI-METRICS

2.1 Introduction

In order to test the applicability of Minkowski-metrics as a combination rule for (digital-image-coding) impairments, an experiment has been carried out in which subjects assessed the perceptual quality of scale-space-coded color images comprising one to three perceptually distinct impairments. This is, in fact, an extension of one of the experiments described by de Ridder[6]. In that experiment, images comprised two kinds of impairment, viz., 'unsharpness' and 'color desaturation'. To this, a third kind of impairment was added. It is denoted 'phantoms' as it consists of dark and bright patches within homogeneous bright and dark regions respectively.

2.2 Method

Stimuli

The coded pictures were two static, complex scenes: an outdoor scene (TERRASGEEL) and the portrait of a female model (WANDA01). RGB signals, obtained by scanning slides of these scenes, were digitized with 8 bits/pixel on a grid of 512 by 512 pixels and then transformed to luminance component Y and chrominance components U and V that are used in PAL color TV transmissions. Subsequently, scale-space-coding/decoding was applied to each component with the restriction that U and V components were always coded the same way. Finally, the coded YUV components were transformed to RGB signals again. During quality assessments, only part of the images (240 by 470 pixels) was displayed in the middle of the screen of a CONRAC 7211 high resolution color monitor.

For each component, five prediction error signals were generated that formed a scale-space pyramid[5,15]. These signals varied systematically in spatial resolution, the one on the lowest level or scale of the pyramid representing the highest-resolution information of the input image. Throughout this paper, these pyramids will be called Y- and UV-pyramids, while the lowest scale will be labelled s_0, the next scale s_1, etc. To create these prediction error signals, five images of reduced size were derived from the input image by means of sampling the latter with square lattices of 2D Gaussian kernels that were $T_i = 2^i$ pixels apart, with $i = 1, 2, \ldots, 5$. The ratio between sampling distance T_i and the standard deviation of the Gaussian weighting function on scale s_i was always equal to three. Subsequently, the reduced image on scale s_i was used to construct a prediction of the image on scale s_{i-1} by upsampling and interpolation. Subtraction of this predicted image from the image on scale s_{i-1} resulted in the prediction error signal on scale s_{i-1}. For the reconstruction of the input image, only the prediction error signals and the reduced image on scale s_5 needed to be transmitted.

Impairments were generated by uniform quantization of prediction error signals. Throughout this paper, the degree of quantization of the signal on scale s_{i-1} will be indicated by the size of quantization step q_{i-1}. A prediction error signal is said to be left intact when the step size is equal to one. That is, the signal is sampled with 8 bits/pixel. The complete deletion of the prediction error signal on scale s_{i-1} is indicated by $q_{i-1} = 0$.

Four levels of 'unsharpness' were introduced in the reconstructed image by deleting both prediction error signals on the two lowest scales of the Y component ($q_0 = q_1 = 0$), by deleting either one of them ($q_0 = 0, q_1 = 1$ and $q_0 = 1, q_1 = 0$) and by leaving both signals intact ($q_0 = q_1 = 1$). Three levels of 'phantoms' were produced by deleting the prediction error signal on scale s_4 of the Y component comprising low-resolution information ($q_4 = 0$), by compressing its information by a factor of four ($q_4 = 61$) and by leaving this signal intact ($q_4 = 1$). In all above-mentioned cases, the prediction error signals on scales s_2 and s_3 were left intact. Finally, four levels of 'color desaturation' were generated by deleting all prediction error signals of the U and V components ($q_0 = \cdots = q_4 = 0$), by deleting the signals on the four lowest scales ($q_0 = \cdots = q_3 = 0, q_4 = 1$) as well as on the three lowest scales ($q_0 = \cdots = q_2 = 0, q_3 = 1, q_4 = 1$) and by leaving all signals intact ($q_0 = \cdots = q_4 = 1$). All possible combinations of these impairments led to 48 different images per scene.

Procedure

The experiment was carried out by one female and two male subjects. Their age varied between 22 and 28 years. The subjects had normal or corrected-to-normal vision. Their visual acuity measured by a Landolt chart at 5 meter distance varied between 1.25 and 2.00. No color deficiencies were reported. Viewing conditions were in accordance with CCIR Recommendation 500[16] with the one exception that the peak luminance was increased to 110 cd/m^2. The subjects rated perceptual image quality on a 10-point numerical category scale, ranging from 1 (low quality) to 10 (high quality). There were two sessions per subject. In the course of a session all 96 images were displayed three times in a random order, except that the same scene never appeared on two consecutive trials. Stimuli were presented for 5 seconds, after which a 40 cd/m^2 adaptation field appeared on the screen. The resulting quality ratings were transformed to an interval scale using a class I, condition D model based on Thurstone's 'law of categorical judgment'[11]. For further details on experimental conditions, see de Ridder and Majoor[10].

2.3 Results and discussion

Figure 2 shows the experimental results for images comprising one or two impairments, viz., 'unsharpness' and 'color desaturation' (Fig. 2, upper panels), 'phantoms' and 'color desaturation' (Fig. 2, middle panels) and 'unsharpness' and 'phantoms' (Fig. 2, lower panels). The filled symbols in Figure 2 denote data that have been taken from another experiment[6]. These data have been linearly transformed to match the corresponding data of the present study (Fig. 2, open symbols). Despite the fact that different subjects participated in the two experiments, a good match was obtained between the two sets of data ($r^2 = 0.89$). This indicates that category ratings from different experiments can be combined by means of a simple linear transformation. The finding that perceptual quality is not affected by deleting the prediction error signals on the two lowest scales of the UV-pyramids, implies that high-resolution color information does not contribute to perceptual image quality.

Figure 2 shows that perceptual quality of images with two impairments is always less than that of images comprising only one of them. This implies that the impairments accumulate. To determine whether Minkowski-metrics can describe this accumulation, all quality values of Figure 2 were transformed into impairment scores by means of eq. (1). Subsequently, a least-squares fitting procedure was used for each pair of impairments to determine the value of exponent p that minimizes the sum of squared differences between predicted impairment scores $\hat{I}_{tot}$ of eq. (2) and corresponding impairment scores I_{tot}. Results of this procedure can be found in Table 1. In all cases, a good fit was obtained ($0.94 \leq r^2 \leq 0.98$), suggesting that Minkowski-metrics are indeed able to describe the observed accumulation of two impairments. To confirm this suggestion, the predicted impairment scores have been transformed into quality ratings $\hat{Q}_{tot}$ by means of eq. (3). In Figure 2, the resulting predicted quality ratings are indicated by dashed lines, demonstrating a good fit between predicted and measured quality.

Table 1 shows that exponent p has about the same value, irrespective of kinds of impairment involved. This suggests that all data can be described by one Minkowski-metric. This has been examined in a subsequent test, in which the fitting procedure was repeated, but now for all images including those comprising three impairments. The results of this procedure can be seen in Figure 3, where the predicted quality ratings are indicated by dashed lines. Again, a good fit was obtained between predicted and measured quality. Exponent p was found to vary around two: $p = 2.08 \pm 0.06$ ($r^2 = 0.96$) for TERRASGEEL and $p = 1.91 \pm 0.05$ ($r^2 = 0.97$) for WANDA01.

Based on a limited set of data, de Ridder[6] concluded that perceptually distinct impairments combine according to a Minkowski-metric with an exponent that is slightly above two. The results of the present study suggest that this exponent can be set equal to two. This would imply that perceptually distinguishable

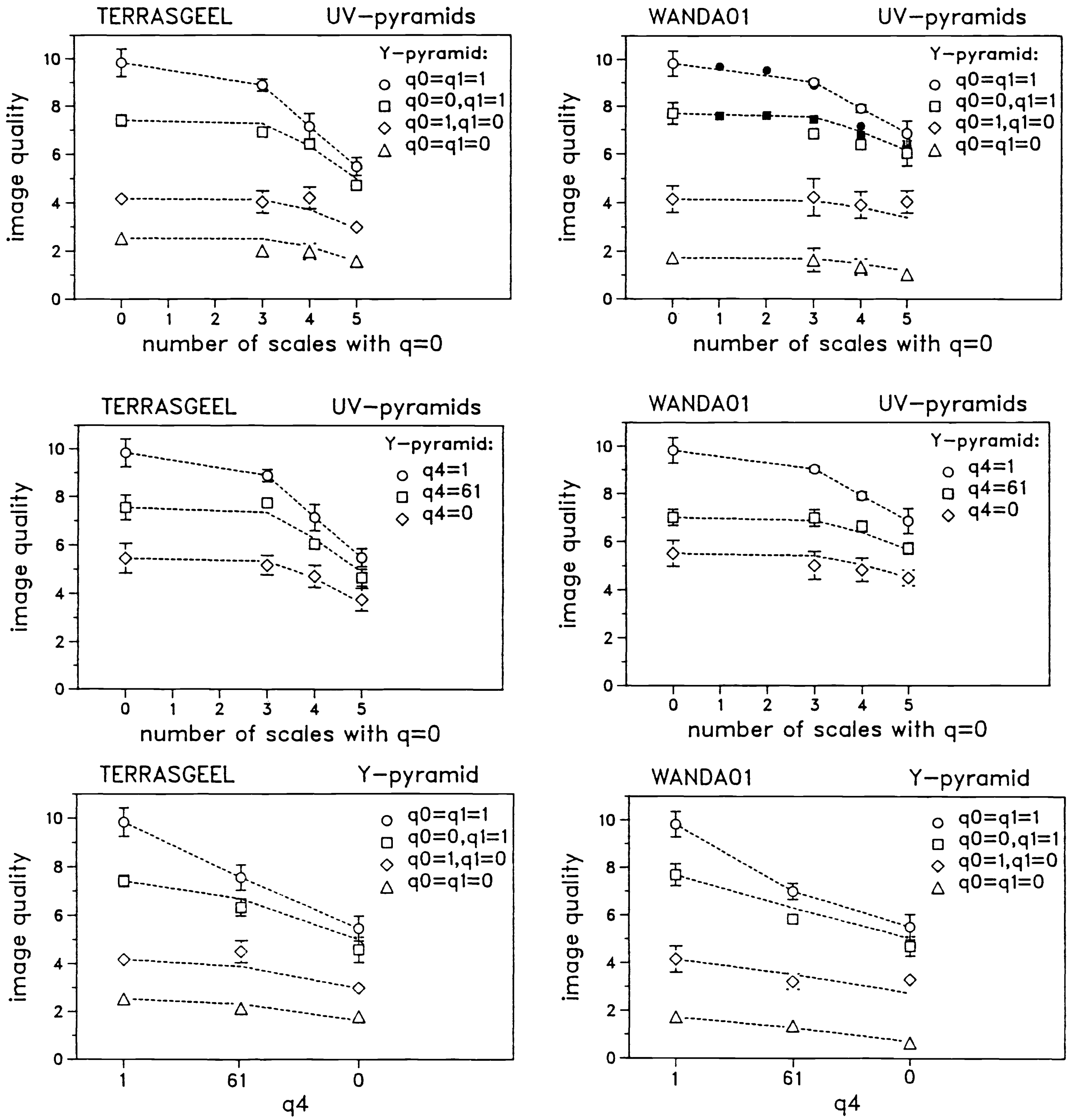

Figure 2: Perceptual quality of scale-space-coded color images comprising one or two impairments, viz., 'unsharpness' and 'color desaturation' (upper panels), 'phantoms' and 'color desaturation' (middle panels) and 'unsharpness' and 'phantoms' (lower panels). For details on how these impairments are generated, see section 2.2. Filled symbols denote data that have been taken from another experiment[6]. They have been linearly transformed to match the corresponding data of the present study (open symbols). Left-hand panels: TERRASGEEL, right-hand panels: WANDA01. Data have been averaged across three subjects. Dashed lines connect predicted quality ratings $\hat{Q}_{tot}$ of eq. (3). Values of these ratings are based on exponents that are presented in the upper three rows of Table 1.

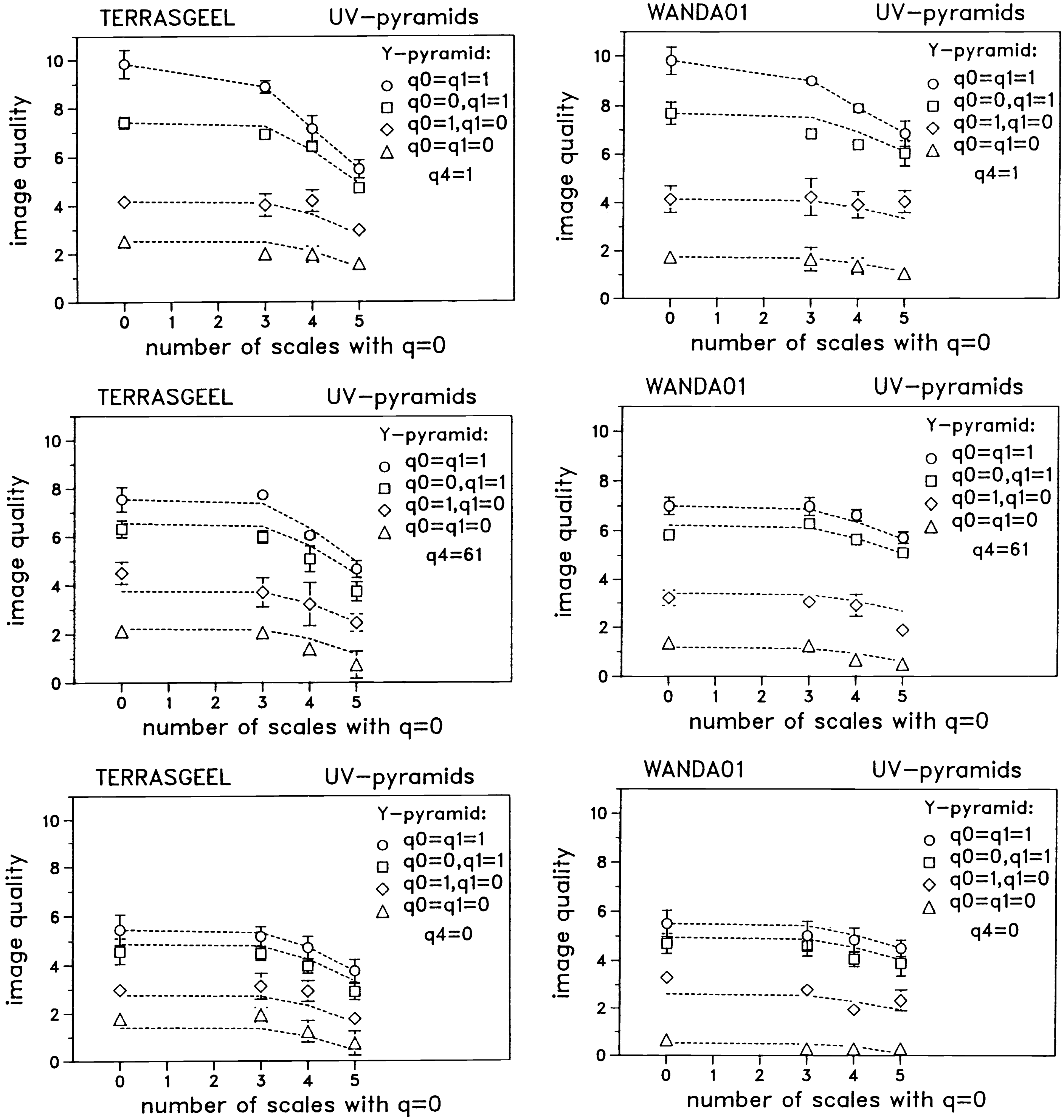

Figure 3: Perceptual quality of scale-space-coded color images comprising one to three impairments, viz., 'unsharpness', 'phantoms' and 'color desaturation'. For details on how these impairments are generated, see section 2.2. Left-hand panels: TERRASGEEL, right-hand panels: WANDA01. Data have been averaged across three subjects. Dashed lines connect predicted quality ratings $\hat{Q}_{tot}$ of eq. (3). Values of these ratings are based on exponents that are presented in the lower row of Table 1.

Table 1: Results of the least-squares fits of eq. (2) to impairment scores derived from the data shown in Figures 2 and 3. Impairment scores were determined by means of eq. (1).

kind of impairment	TERRASGEEL			WANDA01		
	p	95% confidence interval	r^2	p	95% confidence interval	r^2
'unsharpness' and 'color desaturation'	2.21	1.79-2.64	0.98	1.96	1.40-2.52	0.97
'phantoms' and 'color desaturation'	1.95	1.58-2.31	0.97	1.92	1.50-2.34	0.94
'unsharpness' and 'phantoms'	2.30	1.62-2.80	0.95	2.03	1.56-2.51	0.96
'unsharpness' and 'color desaturation' and 'phantoms'	2.08	1.97-2.20	0.96	1.91	1.81-2.01	0.97

digital-image-coding impairments can be represented by a set of orthogonal vectors along the axes of a multidimensional Euclidean space.

4. EXTENSION OF MINKOWSKI-METRICS

In the previous section, it was demonstrated that Minkowski-metrics can be used as a combination rule for small impairments. How about large impairments? Suppose, an image comprises two different artifacts, each producing an impairment that is so strong that an image comprising one of these impairments becomes almost unrecognizable. Then, the image comprising both impairments as well as images comprising one of these impairments will be classified into the lowest category of a quality scale. In other words, all images will receive an impairment score of one. This, however, implies that overall impairment score $\hat{I}_{tot}$ of eq. (2) becomes larger than one, unless exponent p approaches infinity. This example points out two important things. Firstly, there exists an upper bound for perceived strength of impairment. Secondly, Minkowski-metrics do not take into account the fact that upper bounds might exist.

For two impairments ($0 \leq I_1, I_2 \leq 1$), the following extension of eq. (2) is proposed to incorporate an upper bound in Minkowski-metrics:

$$\hat{I}_{tot} = \left(\frac{I_1^p + I_2^p}{1 + (I_1 . I_2)^p} \right)^{1/p} . \tag{4}$$

This equation is a generalization of Schönemann's 'Metric for Bounded Response Scales'[17] in the sense that eq. (4) is not confined to a Minkowski-metric with $p = 1$ (City-block metric). Experimental support for eq. (4) can be found in Figure 4, where it has been fitted to data extracted from an experiment in which, in extreme cases, the perceptual quality of the original image was degraded to such an extent that the image became almost unrecognizable. The interesting aspect of this experiment is that this strong degradation was already accomplished by either of the two impairments involved. These impairments were generated

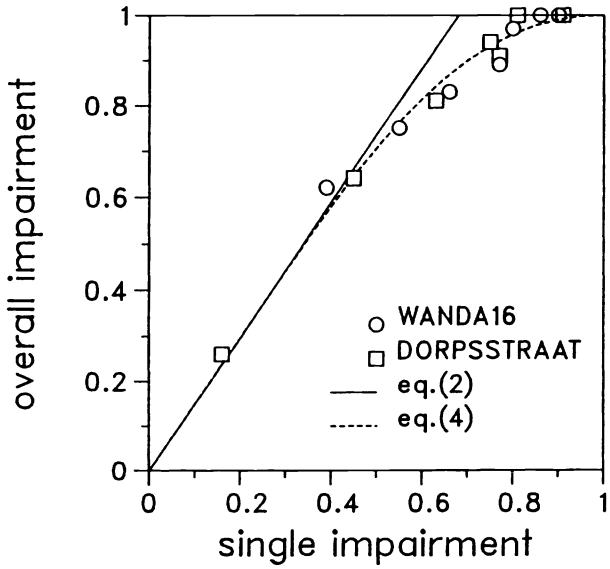

Figure 4: Impairment scores for images comprising two equally strong impairments as a function of the strength of these impairments. Impairment scores have been derived from quality ratings, averaged across six subjects, for two natural scenes: an old building (DORPSSTRAAT) and the portrait of a female model (WANDA16). Dashed line denotes the optimal fit of eq. (4) with $p = 1.77$. Solid line represents predicted impairment scores $\hat{I}_{tot}$ of eq. (2) using the same value of exponent p.

by optical low-pass filtering[18] and by adding normally-distributed spatial noise. In Figure 4, impairment scores are shown only for those images that comprise two equally strong impairments. That is, $I_1 = I_2$. The impairment scores have been derived from quality ratings, averaged across six subjects, for two natural scenes. Eq. (2) predicts that impairment scores lie on a straight line. Figure 4, however, shows that data systematically deviate from a straight line for impairment scores above 0.5. This deviation is correctly predicted by eq. (4) (Fig. 4, dashed line). Exponent p is again close to two, viz., 1.77 $\pm$ 0.18 ($r^2 = 0.98$).

The main difference between this experiment and the one described in section 3 is the perceptual quality range involved. Despite this difference, subjects used the whole category scale to rate image quality in both experiments. This tendency to adapt the category scale to the perceptual range under investigation is a well-known phenomenon in psychometrics[19]. Furthermore, impairment scores are always normalized by means of eq. (1). One way to allow for these properties of the evaluation technique used in the present study is to assume that eq. (4) describes how actually perceived impairments combine and that calculated impairment scores I are proportial to actually perceived impairments, or

$$I = K.I^* \; ; \; K \geq 1, \tag{5}$$

where I^* denotes actually perceived impairment on a scale from zero (original image) to one (unrecognizable image). Then, eq. (4) becomes

$$\hat{I}_{tot} = \left(\frac{I_1^p + I_2^p}{1 + \left(\frac{I_1.I_2}{K^2}\right)^p} \right)^{1/p} \quad ; \; K \geq 1. \tag{6}$$

For $K = 1$, this equation turns into eq. (4) again. In general, however, K will be much larger than one, because, under practical circumstances, impairments are rather small. This implies that the denominator in eq. (6) approaches one, meaning that eq. (6) turns into the original Minkowski-metrics described by eq. (2).

5. CONCLUSIONS

The main conclusions of the present study are:

- Minkowski-metrics can be used as a combination rule for small impairments like those usually encountered in digitally coded images
- Perceptually distinct impairments combine according to a Minkowski-metric with an exponent that is about two
- An upper bound for perceived strength of impairment has to be introduced into Minkowski-metrics to generalize the combination rule using this metric to large impairments.

6. ACKNOWLEDGMENTS

The author wishes to thank Paul Dirix for carrying out the experiment described in section 3. This research has been made possible by a fellowship of the Royal Netherlands Academy of Arts and Sciences.

7. REFERENCES

1. M. Rabbani and P.W. Jones, *Digital image compression techniques*, Tutorial Texts in optical engineering Vol. TT 7, SPIE Optical Engineering Press, Bellingham, WA, USA, 1991.

2. S. Ohtsuka and M. Inoue and K. Watanabe, "Quality evaluation of pictures with multiple impairments based on visually weighted errors", *Proceedings of the SID*, vol. 30, pp. 3-8, 1989.

3. J.W. Woods and S.D. O'Neil, "Subband coding of images", *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 34, pp. 1278-1288, 1986.

4. P.J. Burt and E.H. Adelson, "The Laplacian pyramid as a compact image code", *IEEE Transactions on Communications*, vol. 31, pp. 532-540, 1983.

5. J.B.O.S. Martens and G.M.M. Majoor, "The perceptual relevance of scale-space image coding", *Signal Processing*, vol. 17, pp. 353-364, 1989.

6. H. de Ridder, "Subjective evaluation of scale-space image coding", *Human Vision, Visual Processing, and Digital Display II*, B.E. Rogowitz, M.H. Brill and J.P. Allebach, Eds, Proc. SPIE 1453, pp. 31-42, 1991.

7. R.N. Shepard, "Metric structures in ordinal data", *Journal of Mathematical Psychology*, vol. 3, pp. 287-315, 1966.

8. P.E. Green and F.J. Carmone and S.M. Smith, *Multidimensional scaling: concepts and applications*, Allyn and Bacon, Boston, 1989.

9. L.A. Olzak and J.P. Thomas, "Seeing spatial patterns", In: *Handbook of perception and human performance; Vol. 1: Sensory processes*, K.R. Boff, L. Kaufman and J.P. Thomas, Eds, John Wiley and Sons, New York, 1986.

10. H. de Ridder and G.M.M. Majoor, "Numerical category scaling: an efficient method for assessing digital image coding impairments", *Human Vision and Electronic Imaging: Models, Methods, and Applications*, B.E. Rogowitz and J.P. Allebach, Eds, Proc. SPIE 1249, pp. 65-77, 1990.

11. W.S. Torgerson, *Theory and methods of scaling*, John Wiley and Sons, New York, 1958.

12. J.W. Allnatt and J.M. Corbett, "Comparison of category scales employed for opinion rating", *Proceedings of the IEE*, vol. 121, pp. 785-793, 1974.

13. Y. Yamamoto and Y. Ogata, "Subjective evaluation of white noise effect on picture quality of improved television", *Electronic Letters*, vol. 21, pp. 651-652, 1985.

14. H. de Ridder, "Comparison of combination rules for digital-image-coding impairments". To appear in: *IPO Annual Progress Report*, vol. 26.

15. J.B.O.S. Martens, "Application of scale space to image coding", *IEEE Transactions on Communications*, vol. 38, pp. 1585-1591, 1990.

16. 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.

17. P.H. Schönemann, "Some theory and results for metrics for bounded response scales", *Journal of Mathematical Psychology*, vol. 27, pp. 311-324, 1983.

18. M.R.M. Nijenhuis and F.J.J. Blommaert, "Perceptually optimal optical filters for spatially sampled imagery", *IPO Annual Progress Report*, vol. 25, pp. 66-73, 1990.

19. E.C. Poulton, "Models for biases in judging sensory magnitude", *Psychological Bulletin*, vol. 86, pp. 777-803, 1979.