

Predicted and perceived quality of bit-reduced gray-scale still images

Citation for published version (APA):

Meesters, L. M. J. (2002). *Predicted and perceived quality of bit-reduced gray-scale still images*. [Phd Thesis 1 (Research TU/e / Graduation TU/e), Industrial Engineering and Innovation Sciences]. Technische Universiteit Eindhoven.

Document status and date:

Published: 01/01/2002

Document Version:

Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.tue.nl/taverne

Take down policy

If you believe that this document breaches copyright please contact us at:

openaccess@tue.nl

providing details and we will investigate your claim.

Predicted and perceived quality of bit-reduced gray-scale still images

L.M.J. Meesters

The work described in this thesis has been carried out at
IPO, Center for Research on User-System Interaction,
Eindhoven, the Netherlands.

Printing: Eindhoven University Press Facilities

© L.M.J. Meesters, 2002.

CIP-DATA LIBRARY TECHNISCHE UNIVERSITEIT EINDHOVEN

Predicted and perceived quality of bit-reduced gray-scale still images/
by L.M.J. Meesters. -

Eindhoven: Technische Universiteit Eindhoven, 2002. -

ISBN 90-386-1727-5

NUGI 832

Keywords: Instrumental single-ended measure / Predicted image quality / Perceived image quality

Predicted and perceived quality of bit-reduced gray-scale still images

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Technische Universiteit Eindhoven, op gezag van de
Rector Magnificus, prof.dr. R.A. van Santen, voor een
commissie aangewezen door het College voor
Promoties in het openbaar te verdedigen
op maandag 22 april 2002 om 16.00 uur

door

Lydia Maria Johanna Meesters

geboren te Lemiers

Dit proefschrift is goedgekeurd door de promotoren:

prof.dr. A.J.M. Houtsma
en
prof.dr. A. Kohlrausch

Copromotor:
dr.ir. J.B.O.S. Martens

Acknowledgement

The author wishes to acknowledge the support of the ACTS AC055 project 'Tapestries'.

Contents

1	Introduction	1
1.1	Motivation	1
1.2	Image compression	2
1.2.1	Lossless compression	2
1.2.2	Lossy compression	3
1.3	Measuring perceptual image quality	4
1.3.1	Subjective assessment	4
1.3.2	Experimental conditions	5
1.4	Instrumental image quality measures	6
1.5	Scope of this thesis	7
2	Classification of instrumental measures	11
2.1	Introduction	12
2.2	Image set	13
2.3	Instrumental quality measures	14
2.3.1	Monitor characteristic correction stage	14
2.3.2	Image analysis stage	15
2.3.3	Combination stage	16
2.3.4	Instrumental quality measures used in the clustering analysis	19
2.4	Classification method	19
2.4.1	Normalization	21
2.4.2	Distance measure	22
2.4.3	Multidimensional scaling	23
2.4.4	Ward's hierarchical clustering	23

2.5	Classification of instrumental quality measures	24
2.5.1	MDS stimulus configuration	25
2.5.2	Groups of instrumental quality measures	28
2.6	Scene content	30
2.6.1	Stimulus selection for subjective testing	32
2.6.2	Scene content as selection criterion	36
2.7	Conclusions and discussion	39
3	Quality scaling across scenes and processing methods	41
3.1	Introduction	42
3.2	Subjective quality scaling	43
3.2.1	Quality scale uses	43
3.2.2	Experimental procedure	46
3.3	Experiments: processing methods	47
3.3.1	Stimulus sets	48
3.3.2	Method	49
3.3.3	Results	49
3.3.4	Discussion	55
3.3.5	Conclusions	56
3.4	Experiments: scene content	57
3.4.1	Stimulus sets	57
3.4.2	Method	58
3.4.3	Results	59
3.4.4	Discussion	64
3.4.5	Conclusions	66
3.5	Concluding remarks	66
4	A Single-Ended Blockiness Measure for JPEG-coded Images	67
4.1	Introduction	68
4.2	Experiments	69
4.2.1	Image quality and its underlying attributes	69
4.2.2	Blockiness in natural images	75

Contents

4.3	Blockiness model	76
4.3.1	Front end processing	78
4.3.2	Block boundary estimation	80
4.3.3	Integration	83
4.4	Model evaluation	84
4.4.1	Block boundary estimation on natural images	85
4.4.2	Integration of the estimated block boundaries	87
4.5	Summary	96
5	Evaluation of instrumental quality measures	99
5.1	Introduction	100
5.2	Selection of quality measures and scenes	101
5.2.1	Quality measure selection	102
5.2.2	Scene selection	102
5.3	Experiment 1: attribute scaling within a scene	106
5.3.1	Stimulus set	108
5.3.2	Procedure	108
5.3.3	Results and discussion	108
5.4	Experiment 2: attribute scaling across scenes	113
5.4.1	Stimulus set	113
5.4.2	Procedure	114
5.4.3	Results	114
5.5	Performance of instrumental quality measures	115
5.5.1	Performance within a scene	115
5.5.2	Performance across selected scenes	119
5.5.3	Performance across scenes in general	119
5.6	Performance of the single-ended blockiness measure	121
5.7	Stimulus configuration based on quality predictions and experimental data	126
5.8	Conclusions	127
6	Epilogue	131

	<i>Contents</i>
Bibliography	135
Appendix A	141
Summary	149
Samenvatting	151
Biography	153

Chapter 1

Introduction

1.1 Motivation

The ability of people to communicate over long distances is arguably one of the most important achievements in “present-day” life. Television and telephone have already become indispensable to the average person. If new and upcoming media services (such as video conferencing and internet) want to secure a comparable status they have to provide the quality to which people have become accustomed.

The overall quality of a service is dictated by many aspects, including transmission speed, ease of operation and correct reproduction of sound and imagery. Visual representation of information is essential in most forms of communication. Therefore, good image quality, e.g. a realistic or truthful reproduction of a “real-world” scene, is one of the first requirements that needs to be satisfied. A service that cannot provide good image quality is often deemed by the consumer as less powerful. An extreme example is if parts of the image information are lost so that the message becomes incoherent. But less serious defects, which do not or hardly affect the communicated message, can also affect its quality. For instance, an impaired image is not a realistic reproduction of a “real-world” scene and thereby less interesting to watch because it does not fulfill the consumer’s expectation. Besides, the quality of an image can be affected such that it is strenuous to watch and physically experienced as tiring eyes.

One of the problems in engineering is to reproduce an image as true as possible with limited resources (such as storage or channel capacity). Even though the capacity of storage devices has increased during the years this problem remains a serious one. Consider, for example, the storage capacity needed to store an image sequence of 1 hour, with a frame-rate of 25 frames per second. A typical computer frame comprises 1024x768 pixels and each pixel is represented by three bytes, one for each color channel. This means that one hour of video will take up a storage capacity of approximately 233 GB which adds up to a staggering pile of 300 CD-roms.

Also the channel capacity, the rate at which data can be transmitted over a communication path, is limited. Thus it becomes a problem to host the vastly growing number of TV-channels or computer network users. Since the introduction of the internet, the data put

through computer networks has been growing exponentially. A substantial part of this data load is caused by imagery.

During the years much effort has been spent on using resources as efficient as possible. In the context of imagery this can only be realized with image compression techniques. Nowadays lossy image compression has become unavoidable to achieve the required bit reduction. This implies that image information is lost during the compression process to the extent that it can become noticeable. Therefore lossy image compression above the visual threshold is always a compromise between bit-reduction and image quality.

Future prospects for the internet and certainly wireless communication systems require increasingly more services that operate on low bit rates. Therefore, retaining image quality for still and certainly for moving images is a problem which is expected to grow for some time to come. This thesis addresses some issues in regard to that problem.

1.2 Image compression

Image compression algorithms transform images into less bit-intensive representations. Hence resources like storage and channel capacity can be used more efficiently. In general two methods can be distinguished: lossless and lossy compression. Lossless compression algorithms preserve all image information. Through lossy compression, on the other hand, some image information is lost. The defining feature of lossy compression is that the signal representation of the coded signal is different from the original. We can distinguish between two forms of lossy compression: perceptually lossless coding, where the coded image is perceptually indistinguishable from the original (transparent coding), and perceptually lossy coding, where the coded image visibly differs from the original (non-transparent coding). In the former case, the image represented with the smallest number of bits is assumed to contain only the information that a human perceives (Watson, 1987). Forms of lossless and lossy compression are described in sections 1.2.1 and 1.2.2, respectively.

1.2.1 Lossless compression

The basic principle of lossless image compression is to exploit redundant image information. A gray-scale image is conventionally represented as a 2D array of pixel values. Such a representation contains several forms of redundant data. Two forms of redundancy are: intensity and spatial redundancy also known as coding and interpixel redundancy, respectively (Wandell, 1995; Gonzales and Woods, 1992).

Run-length coding schemes, such as Huffman coding or arithmetic coding, remove coding redundancy by eliminating the restriction that all gray-scale levels in an image are represented by the same amount of bits. For instance with Huffman coding the gray-scale histogram of an image can be used to assign less bits to more frequently occurring gray-scale values than to those less frequently occurring. On average this reduces the number of bits needed to describe an image.

1.2. Image compression

Interpixel redundancy is removed by making use of the spatial correlation of pixel values. The gray-scale values change gradually from one pixel to the other. Therefore the value of any pixel can be predicted from the values of its neighbors. Especially in video sequences this correlation of adjacent pixel values in succeeding frames is used to obtain high compression ratios.

1.2.2 Lossy compression

Lossless compression algorithms exploit the redundancy in image data to obtain more efficient image representations. The process is error-free and reversible, thus the original signal can be recovered. This contrasts with lossy compression which represents a non-reversible process. However, image information can be discarded without being noticeable. Images can be perceptually the same although the physical signals are different. Hence, a third kind of redundant image data can be defined, namely psychovisual redundancy (Gonzales and Woods, 1992). This is image information that is not relevant for human perception.

Human visual processing does not respond with the same sensitivity to all visual information. Frequency sensitivity and contrast masking are properties of the human visual system that can be used to remove perceptually redundant data. With transform coding, such as the discrete cosine transform (DCT), images are transformed from the spatial or pixel domain into the frequency domain. The transform coefficients are products of cosines in two orientations at different spatial frequencies. Although the decomposition is not the same as assumed in the human visual system, the understanding is that high spatial frequencies can be quantized without losing much image quality. It is mainly the quantization process which achieves compression. Other coding methods, which are more similar to the properties of the human visual system, use a pyramid decomposition of the image to achieve a higher compression ratio by quantization of the error images. Transform coding can be used to change the original signal such that bit-reduction is achieved without producing a signal perceptually different from the original.

In present-day applications, removing also non-redundant information seems unavoidable to achieve the necessary high compression ratios. Therefore, the trend is that images are more often becoming compressed above the perceptual threshold. In particular on the internet highly compressed images are no exceptions. For instance JPEG-coded images are highly quantized with unavoidable introduction of disturbing image features such as blockiness and blur. Image quality of highly compressed images is also a problem of interest for television broadcasting. Here it becomes unavoidable to use compression above the perceptual threshold (Falkus, 1996). Therefore understanding the image quality of highly compressed images has become increasingly important. Furthermore, studies of the relationship between the physical parameters of compression algorithms and the resulting image quality are needed to develop or enhance compression algorithms.

The compression methods currently used for still images are JPEG (Pennebaker and Mitchell, 1993) and wavelet coding (Said and Pearlman, 1996). For broadcasting, MPEG-1 and MPEG-2 standards are used while MPEG-4 is used for multi-media (Mitchell *et al.*,

1997). Compression is mainly achieved by quantizing the transform coefficients. The algorithms operate in several modes from lossless to lossy compression. The drawback of most lossy compression modes is that it is up to the user to set the parameters which achieve the compression ratio. The relationship between these physical parameters and image quality is often badly understood and, it is difficult, especially for inexperienced users, to tune the compression parameters such that a certain image quality is achieved. Therefore it is essential that the perceived image quality can be measured and quantified. A definition of the relationship between the physical parameter settings and the perceived image quality would thus be a valuable contribution to help users to set the compression ratio according to the desired image quality.

In this thesis two ways of measuring the image quality of visibly distorted images are considered: subjective image quality measurements and instrumental image quality measurements. In the next sections we will expand on both measuring methods.

1.3 Measuring perceptual image quality

Perceptual image quality is expressed as a gradation of subjective impressions of how well the image information is transmitted to an observer. The observer's criterion of good transmission of image information depends on the application. Roufs (1992) differentiates between two types of perceptual image quality: performance-oriented and appreciation-oriented image quality.

Performance-oriented image quality is applicable whenever the purpose of the images is to facilitate detection tasks. Medical diagnosing for instance is facilitated by MRI or CT images. The purpose of such images is to give accurate information. Therefore if a lesion can be detected by means of a noisy MRI image the image quality satisfies the purpose.

In appreciation-oriented applications, such as television, the goal is to generate images that are as "pleasing" as possible. The emphasis is on the visual comfort associated with the images. For instance, it is strenuous to watch a noise-impaired television program. Watching such a program requires a great deal of effort and viewers experience this as unpleasant. In this thesis we will focus on appreciation-oriented quality.

1.3.1 Subjective assessment

In the ITU-R 500-7 recommendation (ITU-R-500-7, 1997), experimental methods are described to assess perceived image quality of impaired still images and image sequences for television applications. In general three different approaches are proposed: the double-stimulus-continuous-quality-scale method (DSCQS), single-stimulus methods and stimulus-comparison methods.

In DSCQS observers assess the overall image quality for a series of images pairs. Each pair consists of an unimpaired image (reference) and an impaired image (test). For both images (reference and test) observers assess the overall picture quality separately. Eventually the DSCQS assessment results are differences of scores between the reference and test image.

1.3. Measuring perceptual image quality

Table 1.1: ITU-R 500-7 recommendation rating scales

single stimulus quality scale	DSIS and single stimulus impairment scale	comparison scale
5 excellent	5 imperceptible	-3 much worse
4 good	4 perceptible but not annoying	-2 worse
3 fair	3 slightly annoying	-1 slightly worse
2 poor	2 annoying	0 the same
1 bad	1 very annoying	1 slightly better
		2 better
		3 much better

In single-stimulus scaling the overall picture quality of each image in the stimulus set is assessed individually. In stimulus-comparison scaling, again, a series of images pairs is used. These image pairs can include all possible combinations of two images in the stimulus set or just a sample of all possible image pairs in order to restrict the number of observations. In this procedure, observers assign a relation between the two images for each image pair. The same single-stimulus and stimulus-comparison methods can be used to assess impairment. In the double-stimulus-impairment scale method (DSIS) again a series of image pairs (reference and test) is presented. However, the assessors are asked to judge only the test image, “keeping in mind the reference” (ITU-R-500-7, 1997).

The scaling methods impose different grading scales to assess the perceived image quality. In DSCQS, a continuous graphical scale is used to avoid quantization errors. The scale is often labeled with verbal terms such as *excellent*, *good*, *fair*, *poor*, and *bad*, to guide the observer. For single-stimulus scaling, stimulus-comparison and DSIS the usually applied rating scales such as verbal or numerical categories are given in Table 1.1. The subjects express the perceived image quality, the impairment, or the relation between two images by placing the presented stimuli in one of these categories.

Average observer’s quality judgements can be obtained by a number of different analysis methods. Methods such as averaging the judgements across observers by defining a confidence interval indicating the individual differences are specified by the ITU. More complex judgment models were proposed by Torgerson (1958). At the IPO a lot of effort has been spent on developing such models underlying the rating mechanisms of observers (Boschman, 2001). In this thesis mainly one of these analysis methods is used, namely DifScal.

1.3.2 Experimental conditions

Evaluation methods as described above are used to measure the input-output relationship between manipulated imagery and human visual sensations. The sensation is expressed as a response of image quality gradations using qualitative terms, such as excellent or bad image quality. Unlike in threshold experiments where the unit of the rating scale can be defined as just-noticeable difference, the image quality degradation scale as used in supra-

threshold experiments is an ill-defined scale. The image quality judgements can be affected by contextual effects such as image content, presentation order and stimulus spacing (de Ridder, 2001; ITU-R-JWP10-11Q, 1998).

Threshold experiments have mainly been conducted with simple stimuli such as sinusoidal grating patterns. These stimuli have been useful in perceptual studies, such as measuring display fidelity. However, image quality in terms of appreciation can not be addressed with such simple stimuli. The trend in image quality studies is towards using complex natural scenes. The effect of a specific degree of impairment on image quality is not necessarily the same for images with different content. For example, it depends on the information that is lost or on how annoying the distortions are in a particular region of an image.

1.4 Instrumental image quality measures

A virtue of developing image quality models is to get a better understanding of image quality. This is, for example, essential for improving existing and developing new compression algorithms. Several approaches to obtain a quantitative measure of image quality can be used. In this section we discuss the approaches that are based on 1) a mathematical function to express the loss of information in a physical signal, 2) the transformations in the peripheral human visual pathways, 3) identifying and quantifying the impairment strengths, and 4) knowledge of human visual information processing.

Engineers often use an objective fidelity criterion to express the loss of information in an image. The information loss is expressed as a mathematical function of the original image and a processed version of it. Often used functions are the root mean square error (*RMSE*) or the mean-square signal-to-noise ratio (*SNR*) (Gonzales and Woods, 1992). The simple calculations needed to express the loss of image information have led to a large number of related measures (Eskicioglu and Fisher, 1995). Objective fidelity criteria are probably satisfactory within certain constraints but are not always suited as image quality measures. For instance the image quality of a particular scene processed at several levels with the same processing method can probably be quantified by these objective fidelity criteria. However, applied across scenes or different types of distortion their reliability is most questionable. Daly (1993) showed that differently impaired images with similar *RMSE* can be of different subjective quality.

The lack of taking the visual system into account is probably one of the serious drawbacks of the above mentioned measures. Instrumental image quality measures that include properties of the human visual system (HVS) are more likely to approximate subjective image quality.

HVS-based quality measures model the path an image passes through the human visual system, including the optics of the eye, the retina, and the primary visual cortex. Several variations of implementing these stages of the visual system are possible (Ahumada, 1993; Watson, 1987; Daly, 1993; van den Branden Lambrecht, 1996; Winkler, 1999). A typical HVS measure is described in detail by Lubin (1993). First the optics of the eye and the sampling by the cones is modeled. As for most HVS measures in the field of image-coding, the next

step is to decompose the image in a multiresolution image and the human visual contrast sensitivity as well as the sensitivity to spatial patterns are modeled. At this level, a spatial map of distances is computed between the model output of the reference image (original) and the test image (coded). Finally, the distances are converted and, for instance, summed to a probability representing the probability that a human observer can discriminate between the reference and the test image. The distances can also be converted to perceptual differences between the reference and test image quantified in units of Just Noticeable Differences (JND) and integrated into a single scalar value expressing the perceived image quality.

A different technique to model image quality is based on identifying the underlying attributes of image quality and quantifying the perceived strengths of each attribute. For this approach, descriptions of the subjective attributes, such as noise, blur or blockiness, as well as their technical characterization are needed (Karunasekera and Kingsbury, 1995; Kayargadde and Martens, 1996c; Libert and Fenimore, 1999). To relate the attribute strengths to overall image quality, different combination rules can be used (de Ridder, 1992; Allnatt, 1983). The visibility of the attribute strengths can be quantified from the reference image, usually the original, and a processed version of it (Karunasekera and Kingsbury, 1995). At present, much effort is spent on developing single-ended measures, which quantify the degree of impairment directly from the processed image and do not require an original image. For example, in Kayargadde and Martens (1996d) estimation algorithms based on the Hermite transform were used to estimate the perceptual strength of blur and noise directly from the processed image.

Another current approach is to consider image quality in terms of the adequacy of the image to enable humans to interact with their environment. In this concept image quality is attributed to terms like usefulness and naturalness, expressing the precision of the internal image representation and its match to the description stored in memory, respectively. To quantify these image quality attributes usefulness and naturalness, measures of discriminability and identifiability were used (Janssen and Blommaert, 2000).

1.5 Scope of this thesis

One of the key questions in the field of image quality measurement is: how does one indicate the difference between existing instrumental quality measures and what is or should be the added value of newly developed measures? First of all, factors have to be identified that can be used to discriminate between quality predictions of different measures. For a picture, image quality is determined by the distortions, introduced by, e.g., image acquisition, transmission, processing and display, in combination with the variety of scenes. Human observers are able to judge image quality independent of scene content or impairment type. Since instrumental quality measures are intended to be used as a substitute for human observers they should be able to cope with different scene content and impairment types. These two factors (scene content and impairment type) can therefore probably be used as discriminants for the quality predictions of instrumental measures. More particular, the prediction should correspond with across-scene and across-impairment quality

judgements.

The major aim in this thesis is to enhance our understanding of how human observers assess image quality across scenes and impairment types, and how such judgements and quality predictions can be used to discriminate between the instrumental quality measures available today. The second aim is to develop a single-ended instrumental blockiness measure for sequential baseline coded JPEG images that is robust enough to predict the image quality across scenes. The studies in this thesis are limited to gray-scale still images containing degradations above the perceptual threshold, with the emphasis on JPEG-coded images.

In *Chapter 2* a method is demonstrated to classify instrumental quality measures without the need for subjective testing. The measures will be classified on the basis of their quality predictions only. The advantage of such an initial classification is that the differences between instrumental quality measures can be investigated for a large image set since only computer resources are needed. In the same chapter we will also show that images can be selected which discriminate between the classes of quality measures. The methods introduced in this chapter are not meant to replace the evaluation of instrumental quality measures by means of subjective data, but merely to complement it.

Chapter 3 presents an investigation of how comparison scaling can be used to obtain reliable subjective quality judgements across scenes or distortion types. In comparison scaling subjects judge the quality difference of image pairs. Both images are usually of the same scene content and manipulated by the same processing method although at different levels of compression. This means that only the difference in processing level is compared explicitly. When the stimulus set contains several scenes it is assumed that the subjects apply the same rating scale across scenes even though they are not compared explicitly. The question is whether subjects calibrate their quality scale for each identifiable class of images in a stimulus set. If this is the case reliable subjective quality judgements can only be obtained with an explicit comparison across scenes. The same would hold for a stimulus set containing images of different distortion types.

In *Chapter 4* subjective testing will be used to identify the underlying attributes of image quality for JPEG-coded images. In spite of the fact that several distortions are visible (blockiness, ringing and blurring) it will be shown that the strengths of these distortions are linearly related to the perceived image quality. As a result, the image quality of JPEG-coded images can be modeled by a single attribute. Therefore a single-ended instrumental blockiness measure for JPEG-coded images will be developed. In this model blockiness is derived from the magnitude of horizontal and vertical edges that do not occur in the original image. The edge amplitudes of these artificial horizontal and vertical edges are estimated by means of Hermite coefficients. The estimated edge amplitudes are collapsed into a single value indicating the overall blockiness in a JPEG-coded image. It will be shown that the predicted blockiness correlates highly with the perceived image quality of JPEG-coded images.

Finally in *Chapter 5*, the pre-classification of instrumental quality measures by means of their predictions only (*Chapter 2*) will be used to select quality measures that are essentially different in their quality predictions for JPEG-coded images. As suggested in *Chapter 2*, a

small set of scenes will be selected that discriminates between these measures. These scenes will first be used to obtain subjective image-quality data. The quality judgements will be obtained by explicitly comparing the image quality of different scenes and will then be used to evaluate the performance of the presented instrumental quality measures, including the single-ended blockiness measure derived in *Chapter 4*. It will be demonstrated that quality judgements of selected scenes, obtained from a cluster analysis, are indeed suited to discriminate between the quality measures. Furthermore, it will be investigated whether for each of the selected scenes the linear relationship between the perceived attribute strengths and the perceived image quality is the same.

Chapter 2

Classification of instrumental measures

Abstract

In this chapter various instrumental quality measures are classified on the basis of their quality predictions. Usually, the performance of instrumental quality measures is evaluated by means of subjective data. Due to the time-consuming nature of subjective testing, this can only be done for a limited stimulus set. In contrast a mutual comparison of quality measures by means of their predictions allows to use a large image set with a variety of scene contents and distortion types. In this way the effect of scene content and type of distortion on predictions of quality measures can be explored. Furthermore, it will be demonstrated in this chapter how a small image set of, e.g. 4, scenes can be selected for the purpose of discriminating between the predictions of instrumental quality measures. Using such a selection procedure, the usefulness of quality measures can then be ascertained from a small, well chosen set of images.

2.1 Introduction

In the past years many instrumental quality measures¹ have been proposed for processed and compressed imagery. Nevertheless, ongoing development still increases the number of such measures. Most measures base their quality predictions on the difference between a processed image and its original. The detailed computational approach can be diverse. Some measures are, for example, simple mathematical functions such as the root-mean-squared error while other measures use complex methods to simulate the human visual system (HVS). Nevertheless, all measures aim at modeling the relationship between imagery parameters and the assessment of perceived image quality. Therefore, traditionally the usefulness of instrumental quality measures is evaluated by means of subjective quality data. Since subjective testing is time consuming, an extensive evaluation which includes quality judgements for a wide range of impairments (perceived artifacts introduced due to e.g. image processing) and scenes is definitely hard to achieve. A public image bank and a database of subjective quality judgements can be a solution to this problem (Carney *et al.*, 1999, 2000; Rohaly *et al.*, 2000b,a; Corriveau *et al.*, 2000). For example, the video quality expert group (VQEG) performed intensive subjective tests on a number of image sequences degraded by various distortions. These sequences and subjective data are freely accessible to encourage the video community to test and compare instrumental quality measures. Yet, the database is still limited, subjective quality ratings were obtained for test sequences compressed at a bit-rate of 768 kbs up to 50 Mbs. Furthermore, the quality assessments were performed at a single viewing distance and with a single monitor size. The VQEG evaluation of instrumental quality measures, including the *RMSE*, showed that with such a limited database it is hardly possible to differentiate between the measures. The performance of instrumental quality measures were not fully tested and therefore it is to be expected that more complicated measures will indeed outperform the *RMSE* if for example the range of viewing conditions and video material is extended.

In this chapter we describe a technique to compare and *classify* instrumental quality measures. This classification is performed on the basis of the proximity of their quality predictions. Instead of evaluating the usefulness of instrumental quality measures we address the question whether measures are essentially similar or not. Since only computer resources are consumed and no time-consuming subjective tests are needed, a large image set with varying scene content and a wide range of distortions can easily be used, in such a classification.

The second point discussed in this chapter is how a clustering analysis of a large number of scenes can be used to select a limited number of scenes that allow to discriminate between the predictions of instrumental quality measures. We will also investigate if the scene content can be used as selection criterion for such a representative image set. Predefined classes of scene content will be compared to the groups resulting from the hierarchical cluster analysis of scenes.

The image set used for the classification of instrumental quality measures is described

¹Usually such measures are indicated by the term objective quality measures. We prefer to use the term instrumental quality measures instead since in our opinion the term ‘objective’ cannot be attributed to image quality measures.

in section 2.2. This image set consists of a representative sample of 164 scenes including for instance representations of *portraits*, *objects* and *landscapes*. Diverse types and degrees of distortions are introduced through DCT-coding, wavelet coding and low-pass filtering (Pennebaker and Mitchell, 1993; Watson, 1993; Said and Pearlman, 1996; Gonzales and Woods, 1992). The instrumental quality measures that are analyzed in this thesis are introduced in section 2.3. The following two sections describe the classification: the proposed method to classify instrumental quality measures by means of their predictions (section 2.4) and the resulting groups of instrumental measures which give similar predictions (section 2.5) for the large image set. Finally, in section 2.6 a subset of scenes is selected which discriminate optimally between the groups of instrumental quality measures.

2.2 Image set

The image set used for the categorization of instrumental measures consists of 3936 images. This set is obtained by manipulating 164 scenes by four processing methods, each at six different levels. This collection of 164 scenes represents a considerable range of scene contents among which *portraits*, *objects*, *buildings* and *landscapes* (see Appendix A). A more detailed description of the contents is given in section 2.6. The effect of processing method on the predicted image quality is studied for low-pass filtering and two coding methods, namely DCT-coding and wavelet coding. DCT-coded images are obtained by means of the standard JPEG coding algorithm as well as by means of DCTune which uses an optimized quantization table for each scene.

The processing methods and levels applied to the set of 164 natural scenes are:

- Sequential baseline JPEG coding with Q-parameter: 15, 20, 25, 30, 40 and 60 (Pennebaker and Mitchell, 1993).
- DCTune coding with perceptual error: 4, 3.5, 3, 2.5, 2 and 1.5 (Watson, 1993).
- Wavelet coding at bit-rates: 0.15, 0.2, 0.3, 0.4, 0.5 and 0.6 bits per pixel (bpp) (Said and Pearlman, 1996).
- Low-pass filtering with kernel length ² of 9, 7, 6, 5, 4 and 3.

Each processing method introduces specific distortions whereby the image quality deteriorates with increasing perceptual-error (DCTune coding) or blur kernel length (Low-pass filtering) and decreasing Q-parameter (JPEG) or bit-rate (Wavelet coding). JPEG and DCTune coding are both block-based DCT-coding algorithms. This implies that the introduced distortions are mainly blockiness, ringing and blur. Although JPEG and DCTune images both contain these distortions, the proportion between these distortions can be different for each

²The low pass filters are normalized binomial NxN filters. Due to the even filter kernels the pixels of the low pass-filtered images filtered with kernel length 4 and 6 are not in registration with the original pixel values. Therefore these low-pass filtered images are bilinearly interpolated (Gonzales and Woods, 1992). Next they are shifted by 1 pixel horizontally and vertically and downsampled to remove the pixel shift such that they are registered with the original.

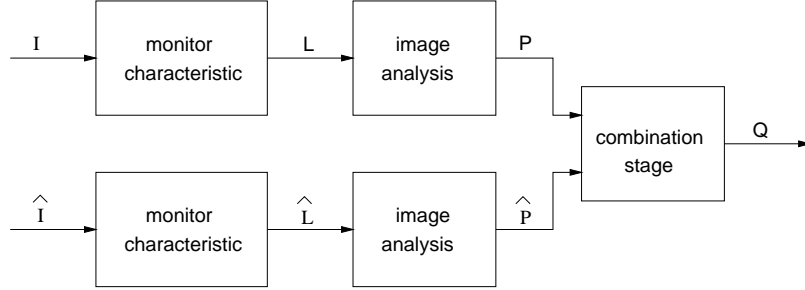


Figure 2.1: Double-ended instrumental quality measures calculate a distance between the original image, I , and a processed version, \hat{I} . Three stages of modeling can be identified. The first stage is a monitor correction stage in which the original image and a processed version of it are transformed in a luminance image, L and \hat{L} , respectively. The images P and \hat{P} resulting from an image analysis stage contain image information, e.g. image edges. Next a difference image is obtained and in the combination stage collapsed into a single scalar value, Q , representing the distance between the original image I and its processed version \hat{I} . Not all three stages are necessarily present in each instrumental quality measure.

coding method. Moreover, in JPEG the strengths of blockiness and blur increase monotonically with decreasing Q-parameter while ringing tends to saturate for high data compression (de Ridder and Willemsen, 2000). Wavelet coding introduces mainly blur which occurs at image-dependent positions. This is in contrast with the uniformly distributed blur in the low-pass filtered images.

2.3 Instrumental quality measures

The majority of instrumental measures used to assess image quality compute a distance between the original and a processed version of it (Ahumada, 1993). In this section a particular group of such quality measures is described. These measures consist of up to three computational stages: a monitor characteristic correction stage, an image analysis stage and a combination stage (see figure 2.1). Not all three stages are necessarily present in each instrumental quality measure. The separate stages are described in sections 2.3.1, 2.3.2 and 2.3.3, respectively. Different instrumental measures are obtained by varying the computational approach in each stage. Finally, the measures as used in the classification of section 2.5 are listed in section 2.3.4.

2.3.1 Monitor characteristic correction stage

The input images I and \hat{I} are 2-dimensional digital representations of gray-scale values. These gray-scale values are internal values but the monitor characteristic can affect the per-

2.3. Instrumental quality measures

ceived energy or emitted light. Therefore a gray-value-to-luminance characteristic of the monitor is modeled as:

$$L = \max[L_{min}, L_{max} (g/g_{max})^\gamma], \quad (2.1)$$

where the minimum luminance is $L_{min} = 0.2 \text{ cd/m}^2$, and the maximum luminance³ is $L_{max} = 60 \text{ cd/m}^2$. The gray-scale values g lie between 0 and g_{max} with a maximum gray-scale value of $g_{max} = 255$, and the exponent γ equals 2.5 (Poynton, 1993).

2.3.2 Image analysis stage

In this stage specific information is extracted from the images. As described in section 2.2 various processing methods were applied on the images, namely DCT-coding, wavelet-coding and low-pass filtering. These processing methods introduce distortions such as blockiness, ringing and blur. The artifacts manifest themselves in an image as added or lost edges. Therefore an image analysis technique is used to extract the amount of spatial information in an image (Beerends, 1997; ITU-WP-2/12, 1995). The *Sobel* edge filter is used to calculate the edge magnitude for each pixel (i, j) in an image \mathbf{I} in two orientations namely in the horizontal and the vertical direction (Gonzales and Woods, 1992). The filter kernel in the horizontal direction is:

$$\begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix},$$

and in the vertical direction:

$$\begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}.$$

The filter kernels are centered on each pixel (i, j) in an image \mathbf{I} in a 3x3 pixel neighborhood. The filter coefficients are multiplied by the pixel values and subsequently added. For each pixel the magnitudes in horizontal direction $Sobel_x(i, j)$ and in vertical direction $Sobel_y(i, j)$ are combined into the gradient amplitude:

$$Sobel_r(i, j) = \sqrt{Sobel_x(i, j)^2 + Sobel_y(i, j)^2}. \quad (2.2)$$

Thus each pixel $I(i, j)$ is replaced by the derived edge magnitude $Sobel_r(i, j)$. The distance measures as described below in section 2.3.4 are based on the edge magnitude only.

³ L_{min} and L_{max} are chosen for a specific monitor. These values are based on the monitor calibration as obtained for the experiments in chapter 3 and chapter 5.

2.3.3 Combination stage

The image analysis stage extracts edge information for a processed image and its original. If these quantities of information are subtracted a difference map is obtained. For a human, this difference map often shows in a glance where differences between both images are detected. The combination stage is used to actually compute a single scalar value which indicates a distance between two images. It is obvious that if these maps are collapsed into a single scalar value spatial information is lost. On the other hand, the difference maps themselves mainly reveal the positions where differences are detected but do not give a quality indication or an indication how perceptually different they are.

The results presented by the VQEG (Video Quality Expert Group) showed that the *RMSE* performed well for moving images (Corriveau *et al.*, 2000). For that reason several derivatives of this simple measure will be used as combination rules for the measures that are classified in section 2.5. Similar combination rules are used in most existing double-ended measures (Eskicioglu and Fisher, 1995).

The first two stages, as shown in figure 2.1, result in transformed images \mathbf{P} and $\hat{\mathbf{P}}$. The distance resulting from the combination, given as $S_k(\mathbf{P}, \hat{\mathbf{P}})$, indicates how dissimilar the images are. We will look at three different classes of combination rules, namely those based on: 1) a measure of correlation, 2) Minkowski summation, and 3) threshold weighting.

Correlation measures

Two images, \mathbf{P} and $\hat{\mathbf{P}}$, are considered similar if the corresponding pixel values (i, j) in both images are equal up to a linear transformation. Two statistical measures to obtain such a relationship are the Pearson correlation coefficient, r_p , and the inner-product correlation, r_i . The corresponding combination rules are S_1 and S_2 , respectively (see Table 2.1). In both cases a correlation of ± 1 results in a distance of zero. An instrumental measure using combination rule S_1 cannot discriminate between two images for which the pixel values differ up to a scaling factor and offset. For measures using S_2 , two images are the same if their pixel values differ up to a scaling factor. Therefore a distance value of 0 does not necessarily mean that the two images are perceptually indistinguishable.

Minkowski summation

An often-used combination rule is the Minkowski summation represented as S_3 in Table 2.1. In this case it is assumed that large pixel differences have a large impact on the perceived image quality (de Ridder, 1992). The exponent α is used to attribute a higher weight to large pixel differences. Four instances of α were used, namely $\alpha = 1, 2, 3$ and ∞ . An increasing value of α increases the contribution of large pixel differences to the overall distance. A Minkowski summation with $\alpha = 1$ is the average absolute difference between two images (city-block distance) and if $\alpha = 2$, the *RMSE* is calculated. The Minkowski summation with $\alpha = \infty$ is equal to the maximum of the absolute difference between both images \mathbf{P} and $\hat{\mathbf{P}}$.

Table 2.1: The combination rules collapse the pixel differences between the original, P_{ij} , and a processed version of it, \hat{P}_{ij} , into a single scalar value, S_k . The pixel differences are denoted by $F_{ij} = P_{ij} - \hat{P}_{ij}$.

$$\begin{aligned}
 S_1(\mathbf{P}, \hat{\mathbf{P}}) &= 1 - r_p^2 \text{ with } r_p = \sqrt{\frac{\left[\sum_{ij}^N (P_{ij} - P_{avg}) \cdot (\hat{P}_{ij} - \hat{P}_{avg}) \right]^2}{\sum_{ij}^N (P_{ij} - P_{avg})^2 \cdot \sum_{ij}^N (\hat{P}_{ij} - \hat{P}_{avg})^2}} \\
 S_2(\mathbf{P}, \hat{\mathbf{P}}) &= 1 - r_i^2 \text{ with } r_i = \sqrt{\frac{\left[\sum_{ij}^N P_{ij} \cdot \hat{P}_{ij} \right]^2}{\sum_{ij}^N P_{ij}^2 \cdot \sum_{ij}^N \hat{P}_{ij}^2}} \\
 S_3(\mathbf{P}, \hat{\mathbf{P}}, \alpha) &= \left[\sum_{ij}^N |P_{ij} - \hat{P}_{ij}|^\alpha \right]^{\frac{1}{\alpha}} \text{ with } \alpha \in \{1, 2, 3, \infty\} \\
 S_4(\mathbf{P}, \hat{\mathbf{P}}) &= \frac{\sum_{ij}^N (P_{ij} - \hat{P}_{ij})^2}{\sqrt{\sum_{ij}^N P_{ij}^2 \cdot \sum_{ij}^N \hat{P}_{ij}^2}} \\
 S_5(\mathbf{P}, \hat{\mathbf{P}}, \sigma) &= \sum_{ij}^N \text{Perona}(F_{ij}, \sigma) \text{ with } F_{ij} = P_{ij} - \hat{P}_{ij} \text{ and} \\
 &\quad \text{Perona}(F_{ij}, \sigma) = \sigma^2 \log \left[1 + \frac{1}{2} \left(\frac{(F_{ij})^2}{\sigma^2} \right) \right] \\
 S_6(\mathbf{P}, \hat{\mathbf{P}}, \sigma) &= \sum_{ij}^N \text{Tukeky}(F_{ij}, \sigma) \text{ with } F_{ij} = P_{ij} - \hat{P}_{ij} \text{ and} \\
 &\quad \text{Tukeky}(F_{ij}, \sigma) = \begin{cases} \frac{(F_{ij})^2}{\sigma^2} - \frac{(F_{ij})^4}{\sigma^4} + \frac{(F_{ij})^6}{3 \cdot \sigma^6} & , \text{ if } |F_{ij}| \leq \sigma \\ \frac{1}{3} & , \text{ if } |F_{ij}| > \sigma \end{cases} \\
 S_7(\mathbf{P}, \hat{\mathbf{P}}, \sigma) &= \sum_{ij}^N \text{Huber}(F_{ij}, \sigma) \text{ with } F_{ij} = P_{ij} - \hat{P}_{ij} \text{ and} \\
 &\quad \text{Huber}(F_{ij}, \sigma) = \begin{cases} \frac{(F_{ij})^2}{2 \cdot \sigma} + \frac{\sigma}{2} & , \text{ if } |F_{ij}| \leq \sigma \\ |F_{ij}| & , \text{ if } |F_{ij}| > \sigma \end{cases}
 \end{aligned}$$

Table 2.2: In the summation rules S_5 , S_6 and S_7 three threshold values σ are used. The threshold values are pixel values, averaged across scenes, taken at 75%, 90% and 95% of a cumulative histogram. Different threshold values were obtained for gray, *Sobel*-filtered gray, luminance and *Sobel*-filtered luminance difference images.

Threshold values σ for S_5 , S_6 , and S_7			
	75%	90%	95%
gray image	7.68	13.67	18.47
<i>Sobel</i> -filtered gray image	35.94	65.91	90.21
luminance image	1.66	3.45	5.14
<i>Sobel</i> -filtered luminance image	8.10	17.24	25.62

Also a normalized version of the Minkowski summation with exponent $\alpha = 2$ is considered (Eskicioglu and Fisher, 1995). The normalized root-mean-squared error, S_4 , takes the total variation in the original and processed image into account. If the variation of gray values in an image is large then the differences between the original and the processed image is perceptually less visible. On the other hand if the variation of gray values in an image is small then the difference between the original and the processed images is probably more visible.

Threshold weighting

In the combination rules S_3 and S_4 large pixel differences are assumed to contribute more to the overall distance between two images. Three additional combination rules with a threshold parameter will be considered. Pixel differences above a particular threshold are weighted differently than those below the threshold. The three functions given in Table 2.1 are: the Perona transform, S_5 , the Tukey transform, S_6 , and the Huber transform S_7 (Black and Marimont, 1998).

The threshold value σ was determined by means of an image set which consisted of 79 scenes processed by four processing methods at six levels ⁴. The difference images, taken between the processed images and their original, were then used to compute a threshold value. For each difference image a pixel value was taken at 75%, 90% and 95% of the cumulative histogram. The threshold value σ was the average across all difference images at a particular level. Since the combination rule is applied after the first two stages, σ is determined for each of the four optional concatenations. The threshold values σ are given in Table 2.2 for gray-scale, *Sobel*-filtered gray-scale, luminance and *Sobel*-filtered luminance images.

⁴The processing and the scenes are described in section 2.2. A subset of the 164 scenes was used to derive the threshold values.

2.3.4 Instrumental quality measures used in the clustering analysis

The instrumental quality measures as considered in the classification of section 2.5 are a combination of the previously described three stages. By realizing all possible combinations of these stages, we obtain 64 different instrumental measures, listed in Table 2.3. The columns indicate the monitor correction stage and the applied image analysis, whereas the rows represent the combination rules. The name of each instrumental quality measure is given in the separate cells of the table.

In addition three instrumental quality measures, based on the human visual system (HVS), will be used as reference measures. For this purpose two implementations of the Sarnoff model (Lubin, 1995) are used, the full Sarnoff model with all orientation filters and a simplified version of it (Martens and Meesters, 1998). In this chapter we refer to these vision models as *sarnoff* and *sarnoff-s*, respectively. An extended description of the model implementations used in this chapter can be found in Martens and Meesters (1998). Furthermore, a vision model⁵ proposed by CCETT (a joint research center of France Telecom) is used.

2.4 Classification method

In this section we discuss a method to classify instrumental quality measures on basis of the mutual correlation between their quality predictions. Thus, we investigate which instrumental measures give similar outputs, without analyzing the quality of the model predictions with respect to subjective data.

The instrumental quality measures as described in section 2.3 predict the image quality on a continuous scale. Moreover, the double-ended quality measures interpret the image quality as a distance between the original image and a processed version of it. Therefore, the instrumental quality measures are considered to produce quality predictions on a ratio scale (Stevens, 1951; Luce and Krumhansl, 1988). This implies for double-ended instrumental measures that the distance between identical pictures equals zero. Although this is true for each quality measure the individual range of quality predictions may be different. Therefore, the quality predictions are normalized per measure so that the individual ranges of predictions are comparable. Furthermore, we assume that within each instrumental measure the distances of all scenes are measured on the same scale. This implies that within each instrumental measure the distances between scenes are comparable.

A multitude of alternative proximity measures and clustering methods can be used to sort the collection of instrumental quality measures into a number of groups. The choice of proximity measures such as, for example, the Euclidean distance, the City Block distance or the Pearson correlation coefficient can affect the resulting groups of quality measures (Cox and Cox, 1994). The same holds for alternative clustering concepts, such as linkage methods, centroid methods or variance methods, though most methods will give similar results (Anderberg, 1973). In the scope of this chapter the inner-product correlation is chosen

⁵The model was developed in the framework of the European research project TAPESTRIES.

Table 2.3: Overview of the 64 instrumental quality measures obtained by varying the three computational stages. The columns indicate the monitor correction stage and the applied image analysis whereas the rows represent the combination rules. The name of each dissimilarity measure is given in the cells.

monitor characteristic correction stage	luminance		gray	
image analysis stage	none	<i>Sobel</i>	none	<i>Sobel</i>
combination stage				
s_1	ddot	sddot	gddot	gsddot
s_2	dcor	sdcor	gdcor	gsdcor
$s_3, \alpha = 1$	mink1	smink1	gmink1	gsmink1
$s_3, \alpha = 2$	mink2	smink2	gmink2	gsmink2
$s_3, \alpha = 3$	mink3	smink3	gmink3	gsmink3
$s_3, \alpha = \infty$	dmax	sdmax	gdmax	gsdmax
s_4	nrmse	snrmse	gnrmse	gsnrmse
s_5, σ at 75%	per75	sper75	gper75	gsper75
s_5, σ at 90%	per90	sper90	gper90	gsper90
s_5, σ at 95%	per95	sper95	gper95	gsper95
s_6, σ at 75%	tuk75	stuk75	gtuk75	gstuk75
s_6, σ at 90%	tuk90	stuk90	gtuk90	gstuk90
s_6, σ at 95%	tuk95	stuk95	gtuk95	gstuk95
s_7, σ at 75%	hub75	shub75	ghub75	gshub75
s_7, σ at 90%	hub90	shub90	ghub90	gshub90
s_7, σ at 95%	hub95	shub95	ghub95	gshub95

2.4. Classification method

as a proximity measure to characterize the relation between quality measures. This proximity measure is chosen to substantiate the assumption that the predictions of instrumental quality measures are defined up to a scaling factor. Even though the analysis is carried out on the basis of this strong supposition the methodology proposed in the next sections can be applied in the same way if proximity measures are used that require less assumptions. As an example, when using the Spearman rank-order correlation coefficient as a proximity measure, instrumental quality measures are already considered the same if they agree in the rank-order of quality predictions. Another choice has to be made with respect to the cluster analysis procedure. Also in this case the chosen method can affect the clustering of instrumental quality measures to some extent. In consequence, the specific quality measures obtained as representation of each cluster are subject to the applied clustering method. Ward's hierarchical clustering, which is used in our analysis, has the property of generating compact clusters (Everitt, 1993). Finally, we want to emphasize that even though particular choices were made the procedure as described in this chapter does not critically depend on them.

In section 2.4.1 a transformation is given to normalize the quality measure's predictions. A distance measure to express the proximity between instrumental quality measures is described in section 2.4.2. A classification of measures by means of multi-dimensional scaling and Ward's hierarchical cluster analysis is described in sections 2.4.3 and 2.4.4, respectively.

2.4.1 Normalization

The range of numbers indicating quality predictions is not the same for each quality measure. If one would compare the predictions of quality measures with different ranges, those measures with the largest range would contribute most to the overall difference. Therefore prior to computing the proximity between quality measures their predictions are standardized by normalizing for each measure the overall RMSE to unity. A normalized prediction, \hat{Q}_{mspl} , for quality measure m , scene s processed with method p at level l is given by:

$$\hat{Q}_{mspl} = \frac{Q_{mspl}}{\text{RMSE}_m}$$

$$\text{RMSE}_m = \sqrt{\sum_s \sum_p \sum_l Q_{mspl}^2}$$

where Q_{mspl} is the non-normalized quality prediction.

The normalization of quality measure's predictions has no effect if the proximity of quality measures is the inner-product correlation (see section 2.4.2). However, in section 2.6 the relation from one scene to the other is expressed by the city-block distance. In that case it is appropriate to standardize the quality measure's predictions to guarantee the same contribution of each quality measure to the overall scene distance.

2.4.2 Distance measure

Comparing instrumental quality measures means that we have to define a measure of proximity which indicates the relation between two different instrumental measures (Cox and Cox, 1994). Since the measures are assumed to give quality predictions on a ratio scale, a proximity measure is chosen which preserves the scale properties. The distance between instrumental quality measures is assumed to be the same if their predictions are equal up to a scaling factor. Therefore, the inner-product correlation is used as a measure of similarity between the predictions of two instrumental measures m and n :

$$C_{mn} = \left[\frac{\left[\sum_s \sum_p \sum_l \hat{Q}_{mspl} \cdot \hat{Q}_{nsp} \right]^2}{\sum_s \sum_p \sum_l \hat{Q}_{mspl}^2 \cdot \sum_s \sum_p \sum_l \hat{Q}_{nsp}^2} \right]^{\frac{1}{2}}, \quad (2.3)$$

where \hat{Q}_{mspl} is the normalized predicted quality calculated using instrumental measure m for scene s , between an image processed by method p at level l , and its original.

The similarity between two instrumental measures is transformed into a dissimilarity measure in the following way:

$$D_{mn} = 1 - C_{mn}^2. \quad (2.4)$$

Two instrumental measures m and n are identical in their quality prediction if $D_{mn} = 0$. The most extreme dissimilarity between two measures is given as $D_{mn} = 1$.

D_{mn} represents an overall dissimilarity taken across scenes, processing methods and processing levels. The obtained dissimilarity between two instrumental measures can be due to any of these three factors. Because of this, instrumental measures can be grouped as being similar while the similarity can be attributed to either of these factors. For instance if the instrumental measures give different predictions for the various processing methods this can be lost in the overall dissimilarity measure. Therefore, the mutual correlation of the quality measures is also investigated for each processing method separately. The inner-product correlation between two instrumental measures m and n for a particular processing method p is obtained in the following way:

$$C_{mn}(p) = \left[\frac{\left[\sum_s \sum_l \hat{Q}_{msl} \cdot \hat{Q}_{nsl} \right]^2}{\sum_s \sum_l \hat{Q}_{msl}^2 \cdot \sum_s \sum_l \hat{Q}_{nsl}^2} \right]^{\frac{1}{2}}, \quad (2.5)$$

where \hat{Q}_{msl} is the normalized predicted quality calculated using instrumental measures m for scene s between an image processed with method p at level l and its original.

Again the similarity coefficient is transformed into a dissimilarity in the following way:

$$D_{mn}(p) = 1 - C_{mn}^2(p). \quad (2.6)$$

The above described measure of association, D_{mn} , is calculated for every pairwise combination of instrumental measures, resulting in a $N \times N$ dissimilarity matrix that is used in the following section to group instrumental measures.

2.4. Classification method

2.4.3 Multidimensional scaling

The above described dissimilarity matrices are used in a multidimensional scaling analysis (*MDS*) to find groups of similar instrumental measures and to determine the underlying distance measure.

Given a matrix of distances between a number of objects, multidimensional scaling is a technique used to find the coordinates of these objects in a low-dimensional space such that the distances between these points fit the original matrix of distances as closely as possible. A classical example in multidimensional scaling is to reconstruct a geographical map of cities from measured distances between cities only. This technique is valuable since it is easier to interpret a map of the city locations than a matrix with distances between the cities (Cox and Cox, 1994). Besides, the obtained dimensionality and distance measure that best fits the underlying distance matrix can give us a better understanding of the complexity of the data.

The program *xgms* is a multi-dimensional scaling analysis tool which is used to determine classes of instrumental measures (Martens, 1999). *Xgms* allows to interactively alter the parameters of the *MDS* model and to view and manipulate the resulting stimulus configuration. This is a valuable tool to explore and get a better understanding of the original dissimilarity data. With *xgms* instrumental measures are grouped in the same class if their predictions are similar. Below, the *MDS* algorithm is explained briefly.

Let D_{mn} denote the dissimilarity between two instrumental measures m and n . If these instrumental measures are comparable they can be mapped to similar coordinates in an N -dimensional space. The dissimilarities can be modeled as an N -dimensional stimulus configuration $\{\mathbf{x}_1, \dots, \mathbf{x}_m\}$ with \mathbf{x}_i a stimulus position in an N -dimensional space. The dissimilarities are transformed by a power function, T . The monotonically non-linear transformed distances TD_{mn} are linearly related to the distances between the stimulus coordinates $d_{mn} = \|\mathbf{x}_m - \mathbf{x}_n\|_l$, where the Minkowski metric with power l is used to calculate this distance. An N -dimensional stimulus configuration is determined such that the stress formula is minimized:

$$Stress(\mathbf{x}_1, \dots, \mathbf{x}_N) = \frac{\sum_{(m,n)} (TD_{mn} - a \cdot \|\mathbf{x}_m - \mathbf{x}_n\|_l)^2}{\sum_{(m,n)} TD_{mn}^2}. \quad (2.7)$$

An optimally fitting stimulus configuration with the lowest dimensionality will be used as a representation of the dissimilarity matrix indicating the groups of instrumental measures. The dimensionality is chosen by means of the 'elbow principle'. The stress is plotted as a function of the number of dimensions and the optimal dimensionality is chosen where the 'elbow' appears (Cox and Cox, 1994).

2.4.4 Ward's hierarchical clustering

Using multidimensional scaling, an optimal fit is established for a specific distance function which is in our case the Euclidean distance. If for example a two-dimensional stimulus con-

figuration will be obtained, then a plot of the stimulus positions gives a better understanding of the relation between the instrumental quality measures than the distance matrix D . Nevertheless, proceeding from the stimulus configuration it is difficult to actually group the quality measures. Therefore, an agglomerative method is used to cluster the quality measures according to their estimated Euclidean distances.

Ward's hierarchical clustering method is used to find groups of similar instrumental measures within the stimulus configuration (Anderberg, 1973). The concept is that one by one instrumental measures with the smallest within group variance are merged. The detailed procedure is described below.

A matrix, d , is constructed containing the Euclidean distances between all pairwise combined quality measures m and n . In Ward's clustering method the initial distances between the instrumental measures m and n are $\frac{1}{2}d_{mn}^2$. At first, each of the 67 instrumental measures is considered as a separate cluster. Next the following two steps are repeated until all measures are grouped into one cluster.

1. The two clusters, m and n , with the minimum distance in the matrix, d , are merged, resulting in a new cluster v .
2. The distance matrix is updated. A new distance is calculated from the merged cluster v towards all other existing clusters u . The new distances are the Euclidean distances between the centroids of each cluster and are calculated in the following way:

$$d_{uv} = \frac{1}{N_u + N_v} [(N_u + N_m) \cdot d_{um} + (N_u + N_n) \cdot d_{un} - N_u \cdot d_{mn}]$$

where N_u , N_v , N_n and N_m are the number of instrumental measures in the clusters u , v , m and n . The distance d_{mn} , d_{um} and d_{un} are given in the distance matrix d .

The resulting hierarchical cluster tree shows the aggregation of the instrumental measures into groups. Only the major groups are identified as means to determine the difference between measures. This is done in the following way. The distance between the merged clusters is proportional to the increase in the within-group error. This distance between the merged clusters is plotted as a function of the number of clusters and an optimal number of instrumental quality measure groups is chosen where the 'elbow' appears.

2.5 Classification of instrumental quality measures

In this section, instrumental quality measures are clustered according to their image quality predictions. In comparison with evaluating the usefulness of quality measures, we investigate whether their predictions are essentially similar or not for a large image set. The 67 instrumental quality measures which are used for this purpose were discussed in section 2.3.4. These measures will be clustered by means of their image quality predictions obtained for a large image set, namely 3936 pictures. The image set contains 164 different scenes, each processed by four algorithms (JPEG, DCTune, wavelet coding and low-pass filtering) at six levels. This collection of 164 scenes represents a considerable range of

2.5. Classification of instrumental quality measures

scene contents, including *portraits*, *objects*, *buildings*, and *landscapes*. To investigate the relationship between the 67 instrumental measures two analyses are performed: multidimensional scaling and Ward's hierarchical clustering. In section 2.5.1 the resulting MDS stimulus configuration is discussed. Thereupon, in section 2.5.2 groups of instrumental quality measures which compute similar image quality are presented.

2.5.1 MDS stimulus configuration

The relationship between 67 instrumental quality measures is explored by means of the multidimensional scaling tool *xgms*. For each instrumental measure, image quality predictions were obtained for 164 scenes processed with four methods (JPEG, DCTune, wavelet coding and low-pass filtering) at six levels. Next, the proximity measure of section 2.4.2 is applied to express in a single scalar value the relationship between the quality predictions of all pairwise combinations of instrumental measures. Thus, from the (164x4x6) quality scores per quality measure a 67x67 dissimilarity matrix was computed, representing the overall dissimilarity between the instrumental quality measures.

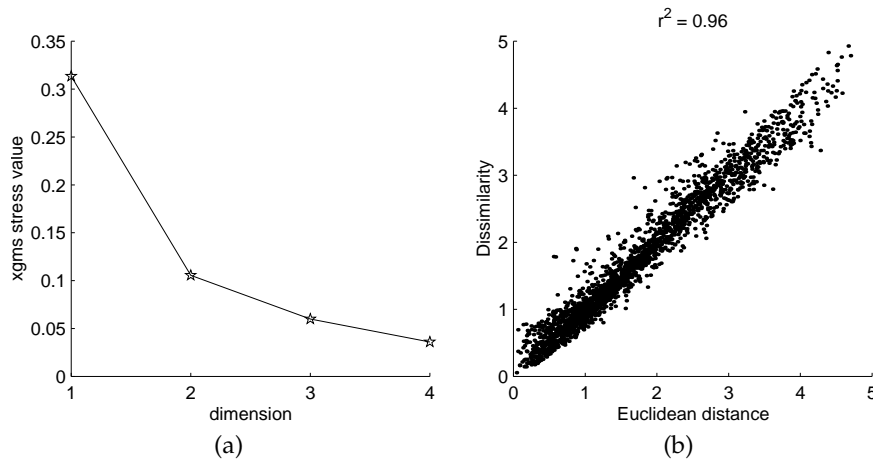


Figure 2.2: (a) The stress for stimulus configurations resulting from MDS analyses up to four dimensions is plotted as a function of the number of dimensions. The depicted stress function was obtained for stimulus configurations derived for quality predictions of all processing methods. (b) The Euclidean distances between the coordinates representing the instrumental quality measures in a 2-dimensional stimulus configuration are plotted versus the original dissimilarity obtained between the quality predictions of 67 instrumental quality measures.

This dissimilarity matrix was input to *xgms* to explore which stimulus configuration, with the lowest dimensionality, fits the original dissimilarity data best. Stimulus configurations up to 4 dimensions were tried. In figure 2.2(a) we show the stress obtained for the dimensions 1 up to 4. As can be seen the "elbow" appears if the dissimilarities are approximated

by a two-dimensional Euclidean space. The fit of the estimated distances increases from one dimension to two dimensions. However additional dimensions, such as in a 3D or 4D configuration, add less to the fit. In figure 2.2(b) the estimated Euclidean distances are plotted versus the original dissimilarities. This figure shows that a two-dimensional stimulus configuration with a Euclidean distance function is an acceptable model for the dissimilarity matrix. The original dissimilarities correlate highly with the approximated Euclidean distances.

In figure 2.3 the resulting 2-dimensional *MDS* stimulus configuration of instrumental quality measures is shown. The 67 instrumental measures are depicted in the following way:

1. the three symbols at the top of the legend characterize the three vision models;
2. the remaining 16x4 instrumental quality measures are represented by 16 sets of four symbols that are connected by lines. The symbols indicate measures applying the 16 different combination rules and:
 - (a) a gray-scale-to-luminance transformation (the 16 unique symbols at the bottom of the legend);
 - (b) a gray-scale-to-luminance transformation in combination with *Sobel* filtering, +;
 - (c) neither gray-scale-to-luminance transformation nor *Sobel* filtering, *;
 - (d) *Sobel* filtering only, ×.

This 2-dimensional graph, representing the 67 instrumental measures, is easier to interpret than the complex 67x67 dissimilarity matrix. Quality measures that compute a similar image quality are located near to each other whereas measures computing a different image quality are far away. It can be seen that the instrumental quality measures are not uniformly distributed over the 2-dimensional space, but they are organized in subgroups. In order to derive groups Ward's clustering will be used in the next section. In figure 2.3 six major clusters resulting from such an analysis are indicated by dotted circles.

The quality measures were also classified for each processing method separately: JPEG, DCTune, Wavelet coding and low-pass filtering. For each processing method a 67x67 dissimilarity matrix was computed from quality scores obtained for 164x6 image pairs per quality measure. In accordance with the previously found dimensionality, where the quality measures were classified for all processing methods, a two-dimensional stimulus configuration gave the best results. The dissimilarities correlate highly with the approximated distances in a two dimensional space, $r^2 \geq 0.96$.

In *xgms*, stimulus configurations are determined up to a rotation and scaling factor. A Procrustes analysis is used to analyze whether the configuration obtained for each processing method is similar to the configuration obtained for quality predictions of all processing methods (Cox and Cox, 1994). The analyses show that an *MDS* stimulus configuration of the separate processing methods correlates highly with the configuration obtained for quality predictions of all processing methods ($0.86 \leq r^2 \leq 0.94$). In the following section we use the latter two-dimensional stimulus configuration for a hierarchical clustering analysis.

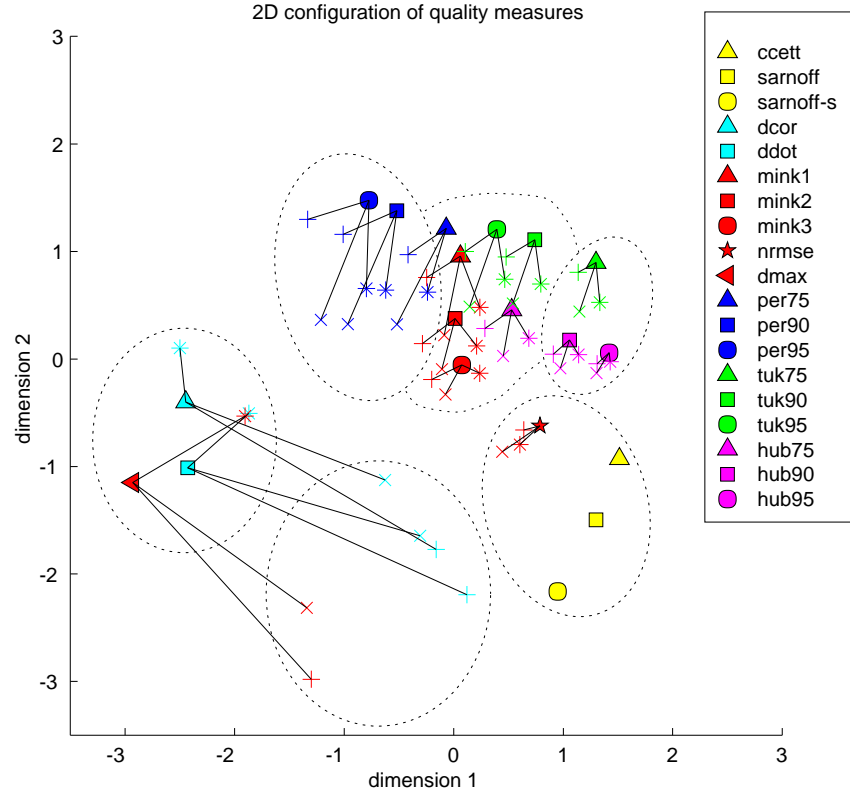


Figure 2.3: Two dimensional stimulus configuration of the quality measures obtained by *xgms* for all quality predictions. The three symbols at the top of the legend characterize the three vision models. The remaining 16x4 instrumental quality measures are represented by 16 sets of 4 symbols that are connected by lines. The symbols indicate measures applying the 16 different combination rules together with 1) a gray-scale-to-luminance transformation (the 16 symbols at the bottom of the legend); 2) a gray-scale-to-luminance transformation together with *Sobel* filtering, +; 3) neither gray-scale-to-luminance transformation nor *Sobel* filtering, *; 4) *Sobel* filtering only, x. The dotted circles indicate the clusters that are obtained by Ward's hierarchical cluster analysis.

2.5.2 Groups of instrumental quality measures

The *MDS* stimulus configuration obtained in the previous section shows that groups of quality measures that predict similar image quality can be identified. In this section, Ward's hierarchical clustering will be used to identify, systematically, groups of similar instrumental quality measures. The distances, $\frac{1}{2}d_{mn}^2$, as obtained by the *MDS* analyses were input to Ward's hierarchical clustering. The resulting aggregation of the quality measures is shown by the hierarchical cluster tree in figure 2.4. In figure 2.5 the distance between the merged clusters is plotted as a function of the number of clusters. The optimal number of groups which contain instrumental measures that compute different quality predictions is chosen where the 'elbow' appears. The figure suggests that the largest dissimilarity between the instrumental quality measures is caused by six main groups.

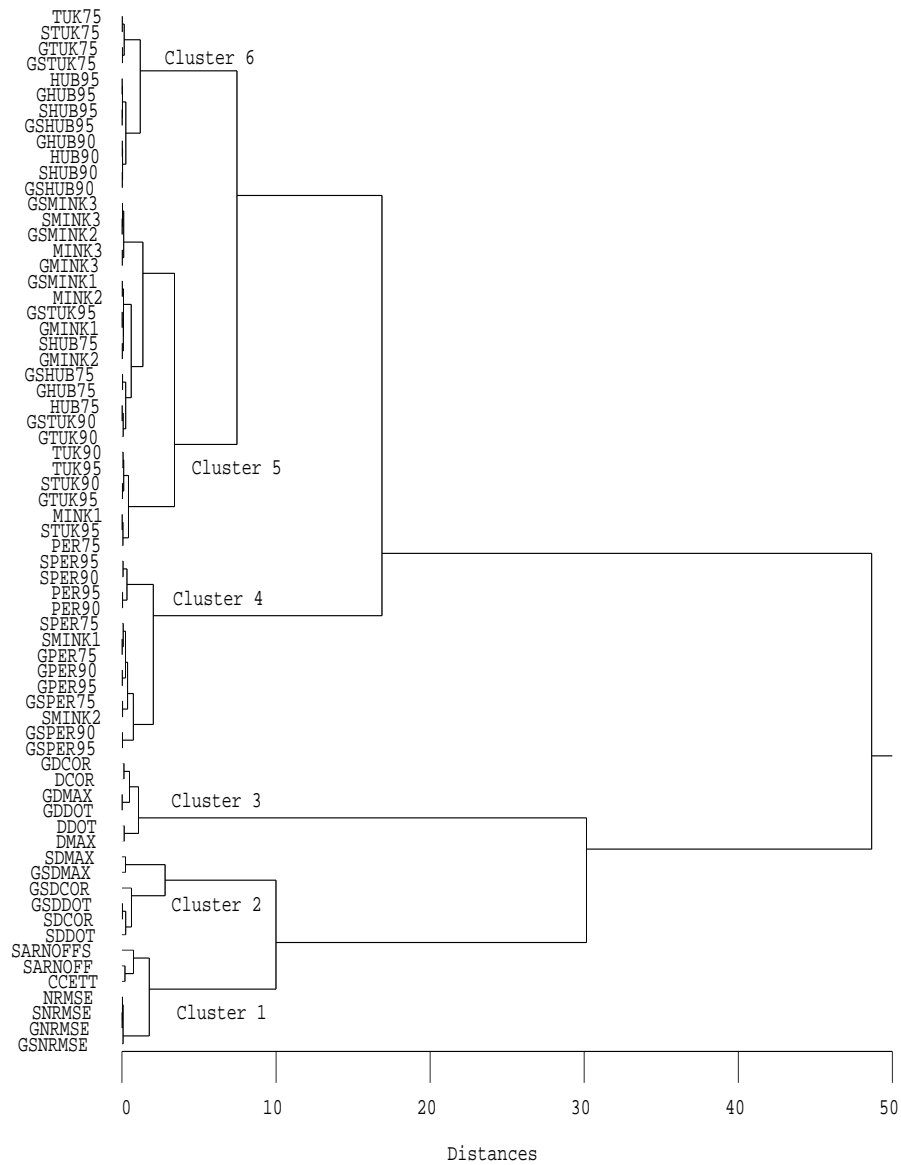
Figure 2.6 shows the two-dimensional stimulus configuration obtained by the *MDS* analysis with the six main clusters of quality measures indicated by different colors. In each panel the effect of one of the computational stages of the quality measures is illustrated.

Figure 2.6(a) shows the effect of the monitor characteristic correction stage on the groups of quality measures. Quality measures in which a gray-scale-to-luminance transformation is applied and those without a transformation are depicted as \diamond and \star , respectively. Corresponding measures are connected by a line. Figure 2.6(b) shows the effect of the image analysis stage on the groups of quality measures. Quality measures in which *Sobel* filtering is applied are indicated by a \diamond , and those which do not apply *Sobel* filtering by \star . The three vision models are depicted as \circ . The two versions of each quality measure with the same monitor characteristic correction and combination stage (with and without *Sobel* filtering) are connected by a line. Figure 2.6(c) shows the effect of the combination stage on the groups of quality measures. The symbols show the quality measures in which the combination rules, S_1 up to S_7 , are applied. The vision models are depicted as \diamond .

Instrumental quality measures compute similar image quality whether a gray-scale-to-luminance transformation is applied or not. This is demonstrated in figure 2.6(a). In this figure the lines connecting the quality measures which apply a gray-scale-to-luminance transformation or not are short such that they mainly belong to the same cluster. From the total 32 pairs of quality measures the two quality measures of 29 pairs belong to the same cluster. A few of the measures are spread across two clusters namely cluster 4 (magenta) and cluster 5 (green), though the distance between them remains small.

Not all quality measures which apply *Sobel* filtering or not compute similar image quality. In figure 2.6(b) the lines connecting these quality measures indicate that for two clusters the between cluster distances of quality measures is larger than the within cluster distances. Considering these two types of quality measures (applying *Sobel* filtering or not), measures with the summation rules S_1 , S_2 and S_3 with $\alpha = \infty$ belong to different clusters.

The main effect leading to the differences between the instrumental quality measures is demonstrated in figure 2.6(c). This figure shows that the main difference between the 67 quality measures can be attributed to the combination rules ($S_1 \cdots S_7$). An interesting observation is that the HVS measures are similar to simple statistical measures which use the normalized root-mean-squared error (S_4) as combination rule. However the root-mean-



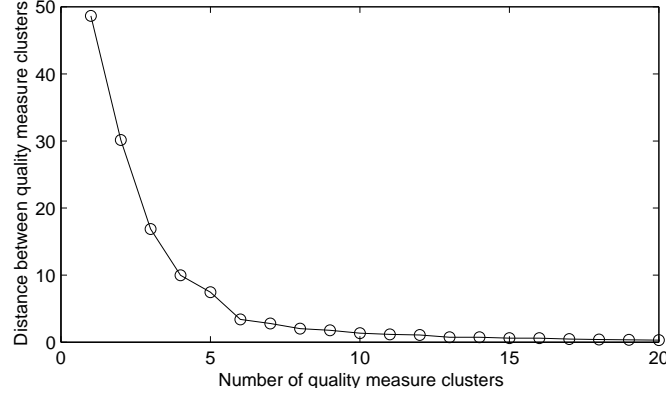


Figure 2.5: The distance between the merged clusters is plotted as a function of the number of clusters.

squared error ($S_3, \alpha = 2$), often used to determine image quality gives dissimilar results.

The largest difference between the quality measures can thus be attributed to the different combination rules. Integrating pixel differences into an overall image quality indication is probably one of the most critical stages of a quality measure. The translation of the perceived differences in an image into an overall quality judgement is in most quality measures simplified to mathematical combination rules as described in this chapter. However whether these rules are comparable to the rules applied by human beings is still not fully understood.

2.6 Scene content

In this section we will investigate the effect of scene content on the predictions of quality measures. Moreover, a method is described to select scenes for subjective testing.

In the previous section, six main groups of quality measures which compute the image quality differently were identified. However, the effect of scene content cannot be deduced from that analysis. In that sense, two questions remain. First of all, whether the quality predictions within each quality measure are similar for each scene. And secondly, whether the difference between groups of quality measures is related to the scene content. If this is the case then scenes can be identified that discriminate between quality measures.

In a subjective evaluation only a limited number of scenes can be incorporated. For example in the experiments conducted in this thesis (see *Chapters 3, 4 and 5*) three or four scenes were used. Therefore it is most useful to select those scenes which discriminate best within and between quality measures. If one would evaluate the six main groups of quality measures with images selected on the bases of their scene content, e.g. *outdoor* or *head-and-shoulder* scenes, it is not necessary that these scenes discriminate best between the quality

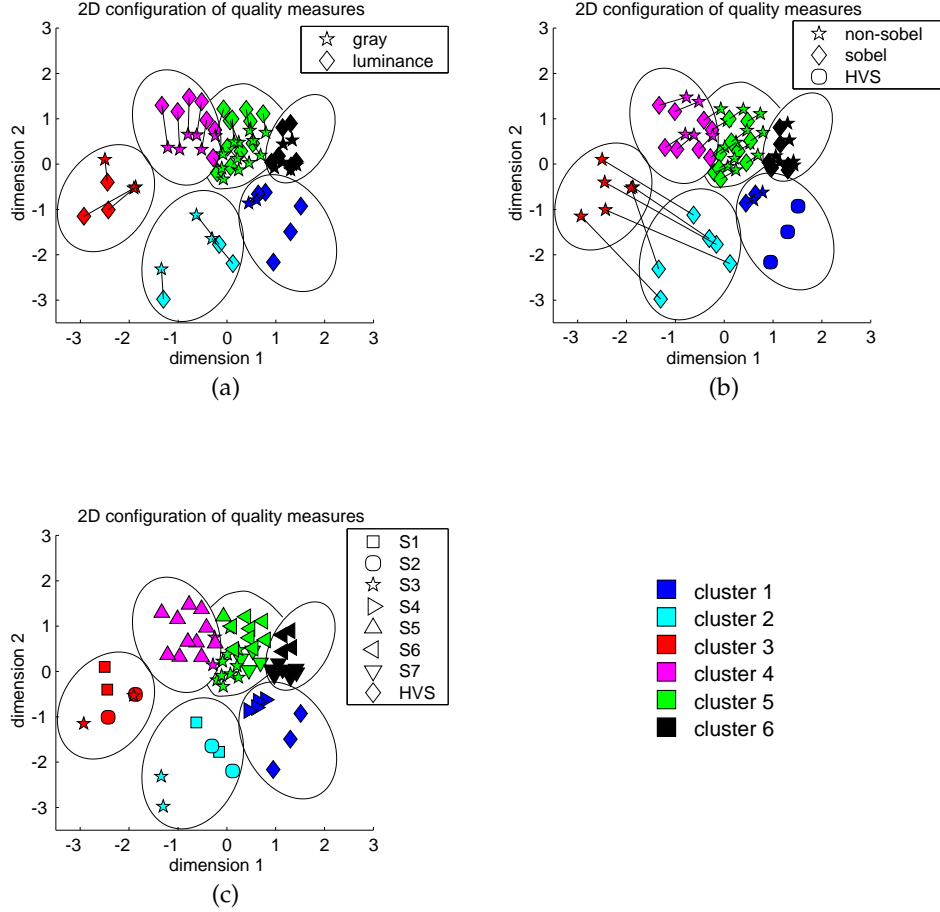


Figure 2.6: Two-dimensional stimulus configurations of the quality measures. The six main clusters are indicated by different colors. a) illustrates the effect of the monitor characteristic correction stage (measures applying no gray-scale-to-luminance transformation, \star , versus measures applying a gray-scale-to-luminance transformation, \diamond). The lines connect the quality measures with the same image analysis and combination stage. b) illustrates the effect of the image analysis stage (measures applying no *Sobel* filtering, \star , versus those applying *Sobel* filtering, \diamond). The lines connect the quality measures with the same monitor characteristic correction and combination stage. c) demonstrates the effect of the combination stage. The symbols indicate the quality measures with different combination rules.

measures. Moreover, the scenes could be such that no difference is obtained between the quality measures and as a result they would perform similarly. This issue will be investigated in detail in *Chapter 5*.

In the following section we will discuss how scenes for subjective testing can be selected systematically. A procedure is described to select scenes which is based on two criteria: 1) each quality measure yields different results for each of the scenes and, 2) scenes yield different results for each of the quality measures. It is suggested to use scenes in a subjective evaluation that meet both criteria.

In section 2.6.2 this procedure is compared to a clustering of scenes based on a priori defined scene content classes.

2.6.1 Stimulus selection for subjective testing

In section 2.5 we showed that the 67 quality measures can be grouped into six different clusters. Thus the predicted image quality, Q_{mspl} , is on average similar for a scene, s , processed by a particular method p , and level l , for all quality measures m , within a cluster. On the other hand quality measures of different clusters predict the image quality, on average, differently. Therefore, to study the effect of scene content on the groups of quality measures it is sufficient to take one quality measure of each cluster. These are selected as follows. First, for each cluster, c_k , a centroid coordinate (x_{c_k}, y_{c_k}) , is calculated as:

$$(x_{c_k}, y_{c_k}) = \frac{1}{2} \left(\max_n \{x_n\} + \min_n \{x_n\}, \max_n \{y_n\} + \min_n \{y_n\} \right),$$

where quality measure n from cluster c_k is represented as a coordinate (x_n, y_n) in a 2 dimensional stimulus space.

Next, for each cluster the quality measure with the maximum summed distance to all other cluster centroids is selected:

$$d_{max} = \max_m \sum_{k=1}^5 \sqrt{(x_{c_k} - x_m)^2 + (y_{c_k} - y_m)^2},$$

where m is a quality measure not belonging to cluster k . The resulting six quality measures are given in Table 2.4. These six selected measures will be used to cluster the 164 scenes.

Table 2.4: The selected instrumental quality measures, one from each of the six clusters. In the left column the cluster colors as used in figure 2.6 are given.

cluster	selected quality measure
1 (blue)	<i>sarnoff-s</i>
2 (cyan)	<i>sdmax</i>
3 (red)	<i>dmax</i>
4 (magenta)	<i>sper95</i>
5 (green)	<i>tuk90</i>
6 (black)	<i>tuk75</i>

Scenes differentiating between predictions within quality measures

In the following a procedure is described to obtain clusters of scenes for which the image quality is predicted differently by each of the six selected quality measures of Table 2.4. In brief, 24 instances per scene were obtained by processing the original with four methods at six levels. For the six selected quality measures the relationship between the scenes is investigated by means of these image quality predictions.

The range of numbers indicating the quality predictions is not the same for each quality measure. If one would compare the predictions of quality measures with a different range then those measures with the largest range would contribute most to the overall difference. Therefore to guarantee the same contribution of each quality measure to the overall scene distance the normalized predictions, \hat{Q}_{mspl} , of section 2.4.1 were used.

In section 2.5 the inner-product correlation was chosen to indicate the relation between instrumental quality measures. However this is not an appropriate proximity measure to indicate the relation from one scene to the other because this measure does not preserve the relation between scenes. After all the scene predictions within a particular quality measure are assumed to be given on the same scale. Therefore the city-block distance is chosen as a measure of proximity to indicate the relation from one scene to the other. The distance between two scenes, s_i and s_j , is computed in the following way:

$$D_{\text{scenes}}(s_i, s_j) = \sum_m \sum_p \sum_l \|\hat{Q}_{mspl} - \hat{Q}_{jspl}\| \quad (2.8)$$

where m is a quality measure, p a processing method and l the level of processing. A 164x164 distance matrix, D_{scenes} , is obtained for all combinations of scenes (s_i, s_j) .

Next, Ward's hierarchical clustering is applied on the distance matrix D_{scenes} . The hierarchical cluster tree is given in figure 2.7. The main groups in the cluster tree are chosen by means of figure 2.8. This figure suggests that five main groups of scenes can be identified. Quality measures predict a similar image quality for scenes of the same cluster. On the other hand quality measures predict dissimilar image quality for scenes from different clusters. Thus to obtain scenes that discriminate within the quality measures, one scene is selected from each scene cluster.

The selection of these representative scenes is done in the same way as the selection of a representative quality measure by maximizing the distance to all other cluster centroids. This analysis shows that a set of five scenes can be selected that discriminates within the quality measures. Such a small set of scenes is feasible and most recommendable to use in subjective testing if the goal is to evaluate the performance within quality measures.

Scenes which differentiate within and between quality measures

Up to now the scenes are grouped together if the quality predictions within each quality measure are similar. Resulting from this, five scenes were selected for which the image quality is predicted differently within a quality measure. Next it will be studied if

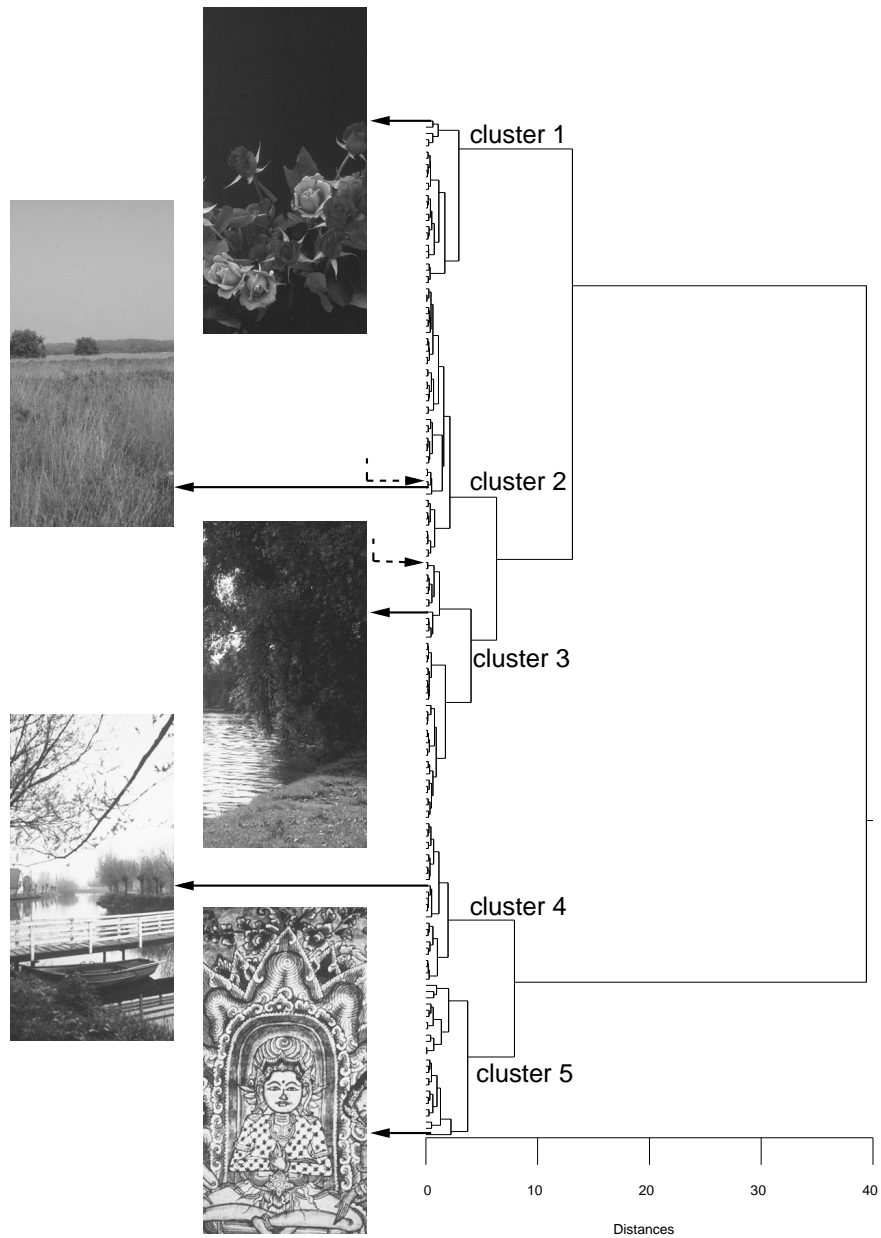


Figure 2.7: The hierarchical cluster tree shows the aggregation of 164 scenes. Five main clusters can be identified. For each cluster the selected scene is shown. The distances between two scenes were obtained by summing their absolute quality difference across 24 image versions and six quality measures. For cluster 2 and 3 different scenes would be selected if only the first criterion is used. These scenes are indicated by dotted arrows.

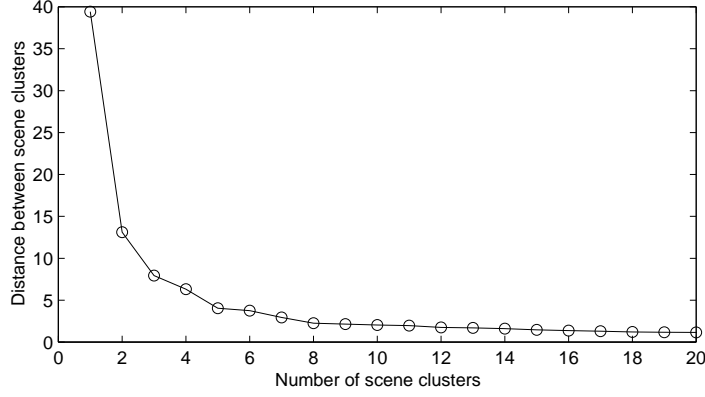


Figure 2.8: The number of scene clusters versus the distance between scene clusters obtained by Ward's hierarchical cluster method when 164 scenes are grouped.

these scenes also discriminate *between* quality measures. For this purpose another proximity measure is taken. Scenes that allow to differentiate between quality measures should have very different predictions of the different quality measures. This discriminability is expressed in a single number in the following way.

The procedure described next is repeated for the scenes of each of the five scene clusters, separately. For each scene in a particular cluster a 6x6 proximity matrix is obtained. This matrix, \mathbf{D}_s , indicates the relation between the predictions from quality measure m to another quality measure n for a particular scene s :

$$\mathbf{D}_{mns} = \sum_{p=1}^4 \sum_{l=1}^6 \|\hat{Q}_{mspl} - \hat{Q}_{nsp}\| \quad (2.9)$$

where \hat{Q}_{mspl} is the quality prediction of quality measure m for processing method p at level l , and \hat{Q}_{nsp} is the quality prediction of quality measure n for the same processing method p , and level l .

For each scene s the proximity matrix is summed as follows:

$$\hat{d}_s = \sum_{m=1}^5 \sum_{n=m+1}^6 \mathbf{D}_{mns} \quad (2.10)$$

where \hat{d}_s is an overall proximity between the quality measures for a particular scene s and \mathbf{D}_{mns} is the proximity between quality measure m and n . The higher the value of \hat{d}_s , the more different are the predictions of the quality measures under consideration for scene s .

For each cluster the scene with the largest value, \hat{d}_s , is selected. These five scenes are shown in figure 2.7. For three clusters the same scene is selected that already resulted from the earlier clustering on the bases of the first criterion (cluster 1, 4 and 5). For the other two

Table 2.5: The selected quality measures if the distance between them is calculated for 164 scenes and for 5 scenes. The cluster numbers are related to the selected quality measures from 164 scenes.

cluster nr	164 scenes	5 scenes
1	<i>sarnoff-s</i>	<i>sarnoff-s</i>
2	<i>sdmax</i>	
3	<i>dmax</i>	<i>dmax</i>
4	<i>sper75</i>	
5	<i>tuk90</i>	
6	<i>tuk75</i>	<i>tuk75</i> and <i>gtuk75</i>

clusters (cluster 2 and cluster 3), different scenes were found for both criteria. In the figure, the scenes that discriminate between quality measures are depicted. The location in the hierarchical cluster tree of the scenes that were selected according to the first criterion are indicated by the dashed arrows. The distance between these images and those chosen according to criterion 2 is small.

Classification of quality measures using a limited number of scenes

Since the clustering of quality measures depends on the scenes under consideration and vice versa, we repeated the clustering of the 67 quality measures described in section 2.5, but now using the 5 representative scenes selected above.

These five scenes were used to test whether they discriminate between the 67 quality measures in an identical way as when 164 scenes are used. From the analysis a similar 2-dimensional stimulus configuration of quality measures was obtained. However, only four main groups of quality measures could be identified by means of five scenes. It follows that a subset of scenes changes the quality measure clusters. Apparently a small set of scenes differentiates less between the quality measures. Next, four quality measures, one from each cluster, were selected. The 4 selected quality measures were compared to those obtained from the original clustering in Table 2.5. So, even though less clusters were found the difference between the selected quality measures obtained by 164 or 5 scenes is small.

Consequently, five scenes can be selected which differentiate between and within quality measures. However they differentiate less between all quality measures than if 164 scenes were used. Therefore, in subjective testing a tradeoff has to be made between a feasible number of scenes for the observers and their power to differentiate between quality measures.

2.6.2 Scene content as selection criterion

In the previous section an objective procedure was described to select a small set of representative images that can be used in a subjective evaluation. This procedure will be applied in *Chapter 5* to select scenes for subjective testing of JPEG coded images. However, in

2.6. Scene content

practice scenes as used in subjective evaluation are often chosen by considering particular image properties. It is known that perceived image quality depends on image properties such as for example brightness, contrast and spatial information. This results most often in choosing images which are described by their scene content. Often used descriptions are *head-and-shoulder*, *outdoor* and *indoor* scenes. Since the image content determines which information is available it is not likely that images of different scene content but with the same degree of impairment are of similar quality. The perceived image quality probably depends on the importance of the lost information. Therefore in this section we will study whether the scene clusters as described in the previous section are related to a priori defined scene content classes. If this is the case then the selection of scenes for subjective testing can also be based on the scene content.

The 164 scenes as used throughout this chapter, see Appendix A, were divided into various a priori defined scene content classes according to the study of Klein Teeselink *et al.* (2000). The authors asked subjects to order a large number of scenes into a predefined number of categories (2,3,4,5 and 6 categories). This study indicated that subjects group a set of images according to the image information such as for example *portrait* or *landscape*. The 164 scenes were divided into similar categories as described by Klein Teeselink *et al.* (2000). The six categories were referred to as: 1) *close-up people*, 2) *no close-up people*, 3) *close-up objects*, 4) *no close-up objects*, 5) *buildings*, and 6) *landscapes*. Figure 2.9 shows the categories and the number of images in each category.

These six a priori defined scene classes were compared to the resulting five scene clusters of the previous section. It was assumed that if the scene content can be used to make an adequate selection for testing quality measure differences then the clusters can be labeled by the a priori defined scene classes. Figure 2.10 gives the distribution of scenes from five clusters over the a priori defined scene classes. The a priori defined scene classes are given on the x-axis and the number of scenes for each of the clusters on the y-axis. In this figure it can be seen that each a priori defined scene category is represented by different scene clusters. Although there is no one-to-one mapping of the a priori defined scenes classes onto the scene clusters the following trend can be observed. The a priori defined scenes classes *no close-up people* and *no close-up object* are mainly distributed over the clusters 1 and 2 while the classes *close-up people* and *close-up objects* are mainly present in the scene clusters 4 and 5. The third scene cluster is predominantly represented by the a priori scene classes *buildings* and *landscapes*. Nevertheless, in this case the relation between the scene content and the five scene clusters is weak. If one selects scenes that are predicted to have different quality within and between the measures at least one knows that not all of the measures perform well, despite their inherent differences. Furthermore, it is not guaranteed that the difference between quality measures can be shown if scenes are selected on the basis of their scene content. In that case it can happen that all measures perform well. In Chapter 5 scenes chosen by scene content as well as discriminability are used to evaluate instrumental quality measures.

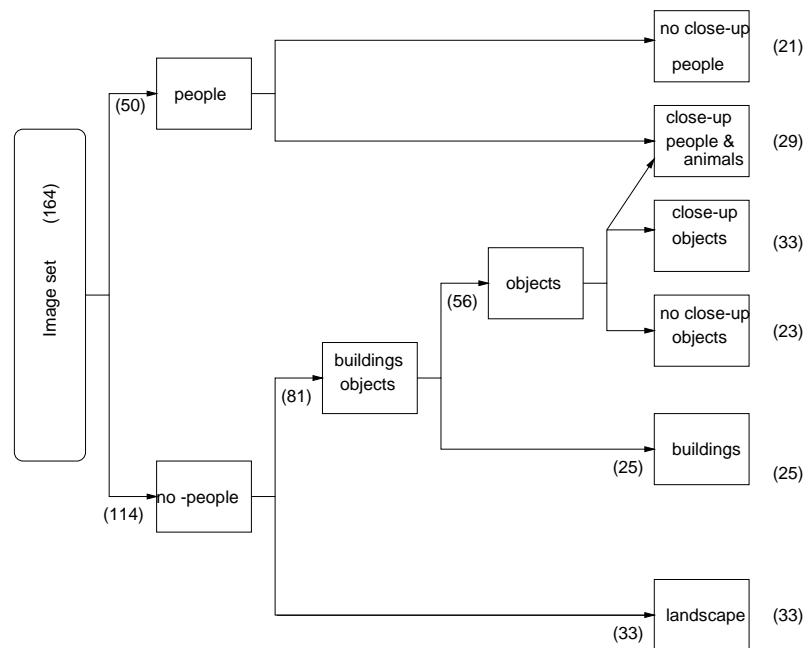


Figure 2.9: The 164 scenes are divided into six categories of similar scene content. In this diagram also the classification for fewer categories is given (Klein Teeselink *et al.*, 2000).

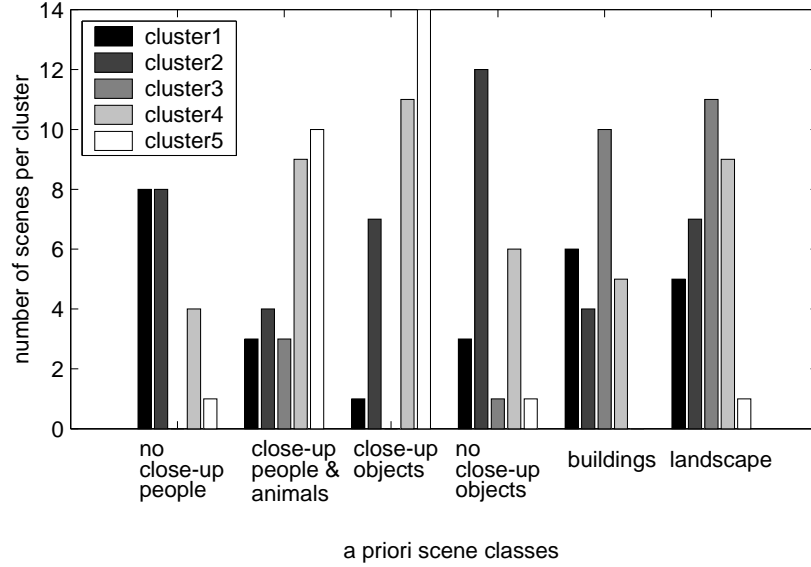


Figure 2.10: The distribution of scenes from five clusters over the a priori defined scene classes.

2.7 Conclusions and discussion

It was demonstrated that the difference between instrumental quality measures can be investigated by means of their predictions only. One advantage is that a large image set can be used since only computer resources are needed. Moreover, it was shown how a highly critical sample of images can be selected such that the differences in quality predictions within and between instrumental measures can be revealed.

The classification of instrumental quality measures depends on the applied proximity measure and clustering analysis. In the scope of this chapter we decided upon the inner-product correlation and Ward's hierarchical cluster analysis, respectively. For these particular choices the results of the classification showed that a large set of 67 quality measures can be reduced to six groups of measures which are essentially different. The difference between these groups of quality measures could mainly be attributed to the applied combination rules. Moreover, the three vision models *sarnoff*, *sarnoff-s* and *ccett* seem to predict similar quality as a normalized version of the *RMSE*. Furthermore, it was shown that a similar two-dimensional configuration of quality measures is obtained if quality predictions are compared for four processing methods together as well as for each processing method separately. This indicates that for the applied processing methods JPEG, DCTune, wavelet coding and low-pass filtering similar groups of quality measures can be identified which predict the image quality differently.

Although a classification on the basis of quality predictions cannot substitute subjective

testing it is a method that can be used to compare newly developed instrumental measures with existing ones. In that perspective, a measure which has proven to compute quality predictions that correlate highly with subjective judgements can be used as a reference to test new measures.

The performance of instrumental quality measures is hard to generalize if only a limited set of images is used to compare quality predictions with quality judgements. Therefore a method was introduced to select by means of Ward's hierarchical cluster analysis a highly critical sample of images consisting of a small number of scenes. These selected scenes discriminate between the quality predictions of different measures as well as between the quality predictions within a measure. Thus using such an image set in subjective testing guarantees that the differences in performance will be revealed. This contrasts to a selection of scenes on the basis of scene content. In this case, the discriminability of the scenes within or between the measures is not known beforehand. For that reason it may be that with such a stimulus set no difference can be demonstrated in the performance of instrumental measures.

Chapter 3

Quality scaling across scenes and processing methods

Abstract

Quality metrics are intended to predict the perceived image quality independent of variations in processing method and scene content. However, one can imagine that in practice this requirement is not trivial to meet. Moreover, reliable subjective data are needed to test quality metrics. Observers might have difficulties in judging the image quality across processing methods or across scenes if the various processing methods or scenes are not compared explicitly. In order to test whether subjects use separate quality rating scales for each identifiable scene and processing method or whether they use a joint rating scale, we conducted experiments in which the image quality was judged with and without explicit comparisons. The results show that subjects use separate quality rating scales for identifiable processing methods or scenes. Moreover, quality judgements with or without explicit comparisons are not always comparable. This implies that reliable subjective quality judgements across different processing methods or across different scenes can only be obtained by explicitly comparing the various processing methods or scenes.

3.1 Introduction

The goal of lossy image coding techniques is to balance the reduction in bit-rate and the unavoidable loss of image quality against each other. Image quality is affected by coder-specific distortions which can vary in strength and location. The perceptual distortion strength and the location of the distortions are not only determined by the selected bit-rate but also by the scene. For instance, JPEG coding introduces blockiness that is most easily perceived in uniform regions. Thus the visibility of these distortions depends on the location and size of such regions in the original image. Therefore we cannot expect different scenes, coded at the same bit-rate, to evoke the same subjective image quality sensation. Hence, instrumental quality metrics should reflect the image quality for a specific coder across scenes. Furthermore, an overall quality measure should also reflect the quality of images coded by different coders, since each coding method introduces typical distortions. For instance, DCT-block based coding introduces mainly blockiness, while wavelet coding introduces blur. The difference in appearance of these distortions may cause dissimilar image quality sensations for a particular scene. Therefore, to determine the overall image quality evaluation tools are needed which take the scene and distortion type into consideration.

Several image quality evaluation metrics have been proposed (Ahumada, 1993; Eskicioglu and Fisher, 1995; Lubin, 1995). They range from simple statistical operations performed on the difference between the original and the coded image to more complex algorithms based on the human visual system. Their principal goal is to predict the image quality as judged by an average human observer during subjective tests. How well the different methods meet this goal, and how different the image quality predictions from the various methods are, is not very well understood. Part of this problem is that subjective data, which are used as reference for these methods, may not be as reliable as is usually assumed.

The acquisition of subjective image quality data is not as straightforward as it seems. De Ridder (1998) pointed out that "quality judgements are affected by the judgement strategies induced by the experimental procedure". Experimental conditions such as stimulus set composition, instructions and scaling technique may influence the image quality responses. This was demonstrated for the effect of scaling technique and stimulus spacing on perceived sharpness judgements. Three scaling techniques: single-stimulus, double-stimulus and comparison scaling, were evaluated for a positively and negatively skewed stimulus set. Only for comparison scaling, comparable results were obtained for both sets. This suggests that for this scaling technique the judgements are hardly affected by stimulus spacing. The issue of influence of scaling procedure on subjective judgements was also addressed by van Dijk and Martens (1996). Single-stimulus and comparison scaling lead to different results for subjective quality evaluation between different codecs. They argued that typical distortions introduced by the different codecs can be easily identified and, consequently, subjects are inclined to use separate rating scales for each coder in single-stimulus scaling. In order to link these subjective scales, an explicit comparison between images from the different coders is required.

If this hypothesis is indeed valid, the same reasoning could be applied to scene content. A stimulus set composed of various scenes makes it possible for the subjects to recognize

each scene. Subjects therefore may choose to use a separate quality scale for each scene.

The above reasoning indicates that quality metrics, tested by means of subjective image quality responses measured on separate scales, mainly predict the level of distortion and not the overall image quality. Thus in order to really evaluate a codec performance it is necessary to obtain quality judgements linked across distortion types and scene content.

In this chapter, psychophysical experiments are conducted to obtain such quality judgements across distortion types and scenes. One aspect of the experiments is to find out whether these quality judgements can be obtained without explicitly comparing different distortion types and scenes (see the following section 3.2). In section 3.3, the effect of distortion is investigated. Two scenes are used to study quality ratings induced by four processing techniques: JPEG coding, DCTune coding, wavelet coding and low-pass filtering. Two experiments are conducted, one for each scene. In section 3.4 the effect of scene content on quality judgements is investigated. Again, two experiments are conducted, one containing JPEG coded images and the other containing wavelet coded images.

3.2 Subjective quality scaling

Several scaling techniques for quantifying the image quality assessments of human observers are available. The most accepted methods, such as single-stimulus scaling double-stimulus scaling or comparison scaling are described in the ITU 500-7 recommendation (ITU-R-500-7, 1997). These subjective test methods are often used to measure the perceived image quality of images degraded by a particular coder. Their reliability concerning image quality measured across images of different scene content or images impaired by different distortion types is hardly investigated. In section 3.2.1 we explain two issues of subjective testing in these cases. In section 3.2.2 the experimental procedure is described.

3.2.1 Quality scale uses

The purpose of subjective quality tests is to measure the quality sensation evoked by a stimulus. Before a subjective test starts observers are given specific instructions how to assess the image quality. Depending on these instructions a stimulus generates a sensation which is expressed by a response on a quality scale. The quality scores are, for instance, categories on a five-point rating scale labeled by "excellent, good, fair, poor, bad" or numerical values such as "5, 4, 3, 2, 1". The scale establishes a relation between the response and the sensation generated by a stimulus (Roufs, 1992). An implicit assumption is often that subjects apply such a rating scale in a similar way to different distortion types or scenes. However, subjects do not always use a quality scale as expected, namely as a single rating scale across all attributes underlying image quality (van Dijk and Martens, 1996). As described in the introduction, quality responses can be evoked by various impairments in a scene if the image quality is degraded by different distortions. In such a case, the complexity of the scaling task increases and it is not obvious that the perceived differences are translated to a global quality impression. As discussed in the introduction a similar increase in scaling complexity can also be caused by differences in scene content.

Scaling behavior can be generalized in the following way: *observers may try to facilitate the scaling task by using separate rating scales for each identifiable class 1) if stimulus categories can be identified, and 2) when no explicit comparisons between the categories are made.*

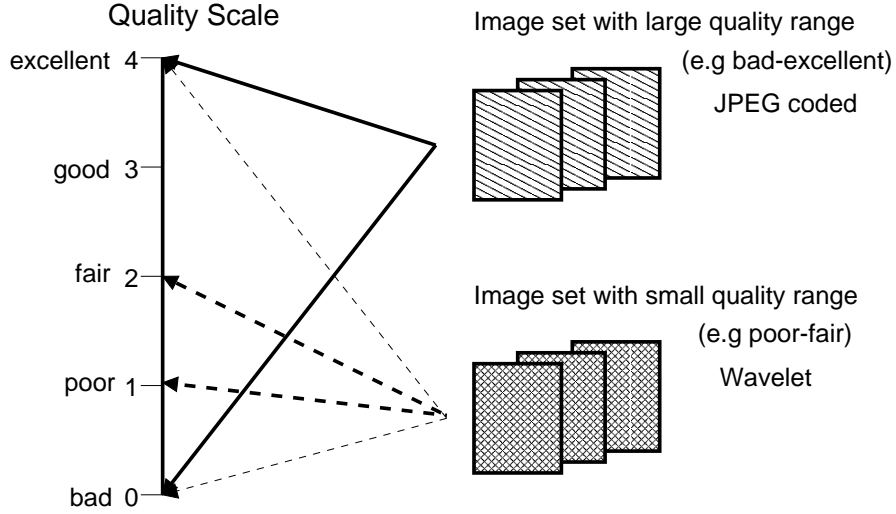


Figure 3.1: Two identifiable sets of images are used in an experiment, e.g. one set of JPEG coded images and one set of wavelet coded images. If separate quality rating scales are used by the observers, the images of both coders could be mapped to the ends of the quality scale. For the JPEG images this is indicated by the thick solid lines and for the wavelet images by the thin dashed lines. This in contrast to the actual quality of the images. The JPEG images indeed cover the whole quality range while the wavelet images should be mapped to the categories indicated by the thick dashed lines.

The motivation for the experiments described in this chapter was to test this general statement by means of specific experiments. We investigated if observers indeed used separate rating scales when identifiable classes, e.g. scenes or distortion types, were not compared explicitly. In such a case observers might assign the ends of the image quality scale to the extreme instances of a class, e.g. the original and the most impaired image. Obviously, this alters the meaning of the quality scores for each class and therefore quality scores assigned to images containing different distortion types or scene content can not be related to each other in a one-to-one mapping. Linear or non-linear transforms are necessary to obtain comparable quality scores across distortion types or scenes.

As an example we consider an image set which is produced by two coders, namely a JPEG and a wavelet coder. Each coder introduces a specific type of distortion which results in a different quality range for the images of each coder, one set of images may have a large quality range and the other may have a small quality range (see figure 3.1). Let us now consider an experiment in which the images of both coders are randomly

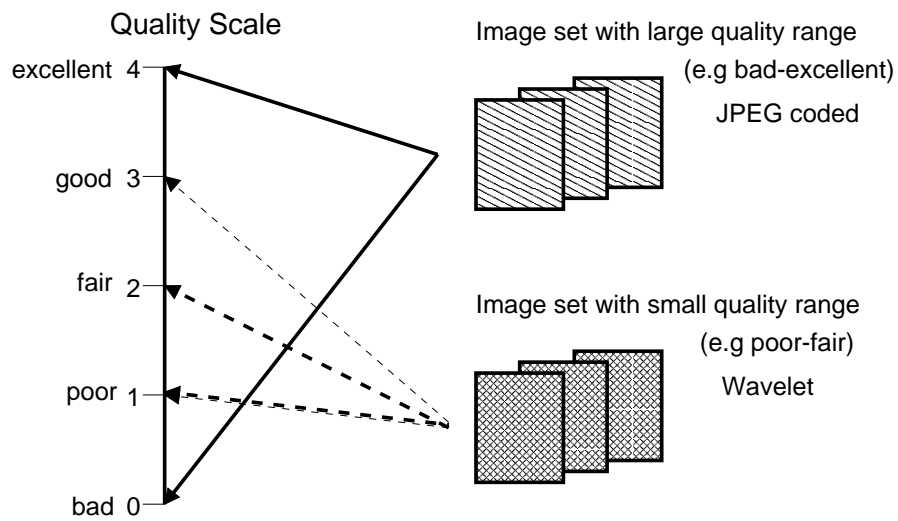


Figure 3.2: Two identifiable sets of images are used in an experiment, e.g. one set of JPEG coded images and one set of wavelet coded images. The observers do not map the images of both coders to the ends of the quality scale. The JPEG images (thick solid lines) are judged in a different quality range than the wavelet images (thin dashed lines). However the quality judgements are not comparable across codecs. The JPEG images indeed cover the whole quality range while the wavelet images should be mapped to the categories indicated by the thick dashed lines.

presented to the observers. They are instructed to judge the quality of each image on a scale from 0 to 4, indicating bad and excellent quality, respectively. If images created by the individual coders are judged on separate quality scales, both in the same range from 0 to 4 (see figure 3.1 a thick line for JPEG and a thin dashed line for wavelet), then a quality of 4 in the first set of images is not comparable to a judged quality of 4 in the second set. The meaning assigned to the categories is different for both quality scales: a 4 in the first set represents a higher quality than a 4 in the other set. In order to link these quality scores we need to know the mapping of the categories from one scale to the other. As suggested by van Dijk and Martens (1996), this can be accomplished if images from both sets are compared explicitly. Therefore, the following hypothesis is tested in the experiments of section 3.3 for different processing methods and in section 3.4 for different scenes:

Hypothesis I: The extreme images of each identifiable class are mapped to the ends of a quality scale when no explicit comparisons are incorporated in the experiment.

When we actually find results supporting this hypothesis there are two possible situations, 1) the quality range of images in different classes are perceptually similar, or 2) it is an artifact of the scaling method and the quality range of images in different classes are perceptually not similar. In the former case, the possibility of a single rating scale can not be rejected since the quality ranges are similar. In the latter case the quality judgements are not comparable between the different classes and we can reject the assumption that observers use a single rating scale. To test which of these situations is actually present the following hypothesis was also tested in the experiments of sections 3.3 and 3.4:

Hypothesis II: Quality judgments obtained with and without explicit comparisons across processing methods or scenes are similar.

Let us assume that the extreme images of a particular coder are not mapped to the ends of the quality scale and hypothesis I would be rejected. This is illustrated in figure 3.2. The JPEG images are judged between 0 and 4 (solid lines) and the wavelet images are judged between 1 and 3 (thin dashed lines). The question remains whether the quality judgements of both coders are similar. This is not the case. The actual quality range, from 1 up to 2, of the wavelet coded images is indicated by the thick dashed lines. In the example, assuming that subjects divide the quality scale in equal steps, a JPEG image judged as 3 is then not comparable to a wavelet image judged 3. After all, in this case the image quality of a JPEG image judged to be of quality 3 is always higher than the quality of any wavelet image. Therefore, in sections 3.3 and 3.4 hypothesis II was also tested when the extreme images were not mapped to the ends of the quality scale.

In the next section quality difference scaling is discussed to investigate these issues of quality scaling across identifiable classes.

3.2.2 Experimental procedure

The single-stimulus scaling technique seems not adequate for an image set containing different characteristics of impairment (van Dijk and Martens, 1996). However, whether this

is due to the scaling technique or to the lack of explicit comparisons can not be deduced from this study. Therefore, in the experiments discussed in sections 3.3 and 3.4, quality difference scaling was used to obtain quality judgements across processing methods and scene contents. In quality difference scaling subjects indicate by a scalar value the difference in perceived quality between two images. This experimental procedure was chosen for the following reason; the same scaling method could be used to obtain quality judgements with and without explicit comparisons across processing methods or scenes. Moreover, comparison scaling is known to have high sensitivity and accuracy.

For the experiments described in sections 3.3 and 3.4 quality difference scaling was applied in the following way. All experiments were divided into two sessions. In both sessions the instructions, images and scaling method were identical; only the presented image pairs were different. In the first session, the image pairs contained two images of identical processing method or scene content. Only the degree of one particular distortion varied. Consequently, the viewers were not forced to compare the image quality of different processing methods or scenes. In order to understand whether the quality variation across processing methods or scene content were incorporated in these quality judgements, in the second session also image pairs containing two images of different processing methods or scene content were compared. To make a comparison for these image pairs, the observers were forced to use a joint quality scale for all processing methods or scenes. Comparing the results of both sessions should illustrate if there is any difference in the use of quality scales.

3.3 Experiments: processing methods

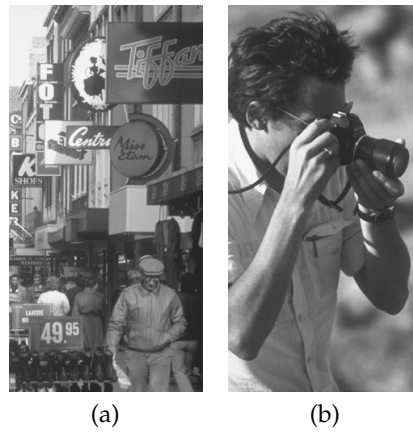


Figure 3.3: Original scenes: (a) *shopping-street* and (b) *photographer*.

Processing methods introduce typical distortions depending on the applied algorithm. We

investigated in two experiments whether subjects used a separate image quality rating scale for each processing method or whether they judged the image quality of all processing methods on one single scale. In the former case, we assume that subjects recognize the different types of distortion, categorize them, and implicitly apply a different rating scale for each category.

Two experiments were designed to investigate which of the above rating scales was used by the observers. Each experiment consisted of two sessions. In the first session, image pairs processed by the same method were judged. In the second session, the same set of images were combined such that also image pairs containing different distortion types were formed. Comparing the subjective data of both sessions should illustrate if there is any difference in the quality scores.

Two scenes were investigated: *shopping-street* and *photographer*. The gray-scale originals with a size of 240x480 pixels are shown in figure 3.3 (a) and (b). They were used in two separate experiments, to avoid an effect of scene content on the quality judgements. The location and degree of distortion is scene dependent and may therefore lead to different quality judgements per scene.

3.3.1 Stimulus sets

Four processing methods were applied to each scene:

1. JPEG coding with Q-parameter of 15, 20, 25, 30 and 60.
2. DCTune coding, which is a JPEG coding with optimized quantization table, with a perceptual error of 4, 3.5, 3, 2 and 1.5 (Watson, 1993).
3. wavelet coding at bit-rates of 0.15, 0.2, 0.3, 0.4 and 0.6 bits per pixel.
4. low-pass binomial filtering with blur kernel lengths of 7, 6, 5, 4, and 2.

Two distinct distortions are introduced by these methods. JPEG and DCTune introduce mainly blockiness. The characteristic impairment in wavelet coded images and in low-pass filtered images is blur. In wavelet coding the blur occurs at specific locations in the image. Some parts are blurred whereas other parts remain sharp. The low pass filter blurs the complete image.

For each scene two stimulus sets were created.

The first stimulus set was constructed as follows. For each processing method, the five processed versions and the original image were combined into pairs such that both images contained a different strength of distortion. This gave 15 image pairs per processing method. In total $4 \times 15 = 60$ image pairs were presented in random order to the observers. The processing technique, and thus the type of introduced impairment, was the same for each image pair, only the degree of distortion varied.

In the second stimulus set, the pairs of images were formed from three levels of each processing method and the original. The selected levels were, for JPEG Q-parameters of 15, 25

3.3. Experiments: processing methods

and 60, for DCTune perceptual errors of 4, 2 and 1.5, for wavelet bit-rates of 0.15, 0.3 and 0.6, and for low-pass filtering kernel lengths of 7, 5 and 2. These levels were chosen such that the quality range for each method was the same as in the first stimulus set. In both sessions the highest and lowest processing levels were incorporated. The processed versions and the original, in total 13 images, were combined into 78 image pairs. Each image pair was unique and consisted of two images with either the same processing method but different degrees of distortion, or different processing methods and equal or different degrees of distortion.

3.3.2 Method

For each scene, *shopping-street* and *photographer*, the different stimulus sets were presented in two separate sessions. In both sessions the observers were seated at a distance of 0.80 m from the monitor. The two images in a stimulus were displayed simultaneously on a calibrated BARCO monitor placed in a dimly lit room, one image on the right hand side and one image on the left hand side of the screen. Care was taken that each degree of distortion and processing method was displayed an equal number of times on both sides of the screen. To avoid order effects, the pairs were presented to each subject in a unique random order.

The instructions were the same for both sessions. Observers had to rate the quality difference between the two images that were simultaneously shown on a scale from -4 to +4. If no difference was perceived, they had to judge 0 and the largest perceivable image difference had to be judged 4. With a "+" or "-" subjects indicated which image had the best quality; a "+" indicated that the quality of the right hand sided image was judged better and a "-" indicates that the left hand sided image was judged better. Before each session started, the observers participated in a trial containing a set of 12 image pairs. These image pairs were chosen such that the observers could get acquainted with the distortion types and quality range in the experiment. The subjects were urged to calibrate their quality difference scale on the image pairs of the trial. For each experiment six subjects participated in both sessions. Subjects took part in only one of the experiments.

3.3.3 Results

The subjective quality difference data of the scenes *shopping-street* and *photographer* are analyzed in the same way.

For the analyses we will assume for the moment that in the first experimental session the rating scale used by an observer is fixed across distortion types. This implies that the quality differences between distortion types are meaningful. Hence, the judged quality differences can be compared across processing method. The question whether this assumption is reasonable will be tested by comparison with the quality judgements from the second session. If the quality judgements of both sessions are similar, the observers probably used the same judgement strategy, which may indicate that a single rating scale was used. If on the other hand the data for the two sessions differ, this may indicate that separate rating

scales were used.

For both sessions a stimulus configuration is deduced by means of DifScal (Boschman, 2001). This analysis tool is based on Thurstone's judgement theory and transforms for each image the quality difference judgements into a quality scale value on an interval scale. First we will investigate if in session 1 the extreme images, the original and the most impaired image of each processing method, are mapped to the ends of the quality scale. After that, the effect of quality judgements across processing methods is studied by comparing the stimulus configurations of sessions 1 and 2.

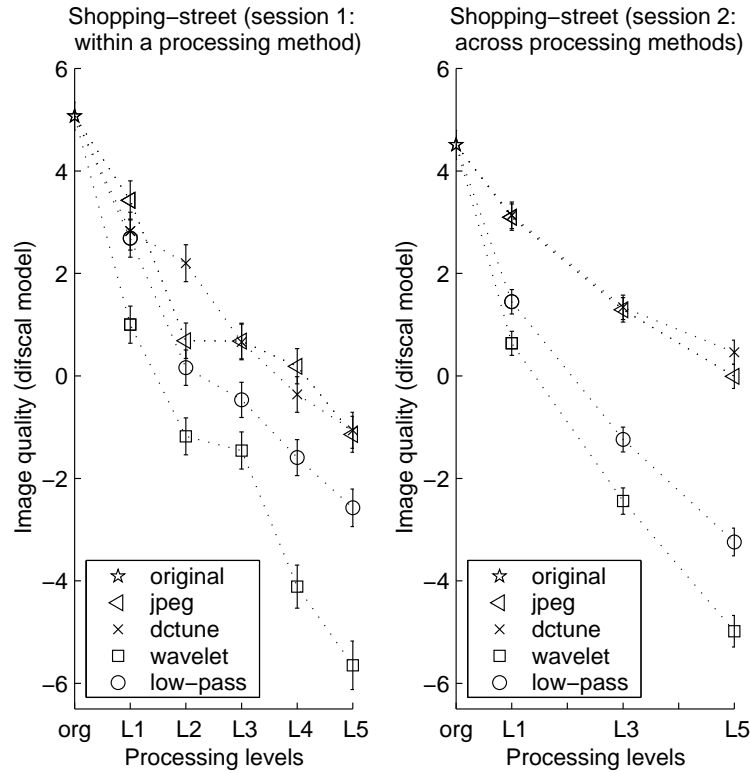


Figure 3.4: The results of DifScal obtained within a processing method (session 1) of the scene *shopping-street* are shown in the left panel. The DifScal results obtained across processing methods (session 2) are shown in the right panel. The x-axis shows the processing levels, L1 up to L5. The original is indicated by org. On the y-axis DifScal's scale values of the images are given. The different ranges of scale values for the processing methods indicate that in session 1 the extreme images are not mapped to the ends of the quality scale.

DifScal stimulus configurations

For each scene a frequency file is generated from the quality difference judgements of session 1 which is used as input for the DifScal analysis. As stated above we assume that the quality scale is fixed across distortion types and that therefore, quality differences of the 21 stimuli (the original and 5 versions of each of the four processing methods) are judged on the same scale. The results of the quality comparisons between these 21 stimuli are collected in *one* single frequency file. The observers rated the quality differences of the 21 pairwise combined images in 9 categories. These categories are represented in the frequency file as nine 21x21 lower triangle matrices. In each 21x21 matrix the original image is represented by the first column and the four 5x5 lower triangles at the diagonal contain the frequencies for each processing method. Each cell in a matrix indicates the frequency with which an image pair (i,j) is judged in that category, summed over all six subjects. For this procedure to be valid we assume that the observers used the categories on the quality difference scale in the same way ¹. In the first session only pairwise comparisons between equal processing methods are rated, which implies that the frequencies in cells between different processing methods are zero. However DifScal can deal with incomplete data as shown in Boschman (2001).

The estimated stimulus scale values and the S-estimate are shown in the left panels of figures 3.4 and 3.5 for the scenes *shopping-street* and *photographer*, respectively. The estimated scale values are shown on the y-axis. The x-axis shows the original image, org, and the five processing levels as L1 up to L5. For each processing method the least impaired image is indicated by L1 and the most impaired image by L5. The error bar at each data-point is the S-estimate of the estimated scale value. This S-estimate is an estimation of the error of the corresponding scale value.

In the same way a frequency file is generated for the comparison data of the second session. The frequency file consists of nine 13x13 lower triangle matrices. To enable comparison between the scale values of both sessions, a linear regression is applied between the scale values of session 1 and those of session 2. This is possible because the stimulus configurations are determined up to a linear transform. The S-estimates of session 2 are scaled by the same factor as in the linear transform applied to the scale values. The linearly transformed stimulus scale values and S-estimates are shown in the right panels of figures 3.4 and 3.5 for the scenes *shopping-street* and *photographer*, respectively.

Quality range

The first hypothesis states that the extreme images of each identifiable class of distortions are mapped to the ends of the quality scale when no explicit comparisons are incorporated in the experiment.

The left panels of figures 3.4 and 3.5 show that the processing methods are judged in a dif-

¹With the help of the program package *xgms* (Martens, 1999) the same analysis can be performed by allowing the use of different quality scales for different observers. Such an analysis shows that the conclusions reached in our main analysis do not critically depend on this assumption.

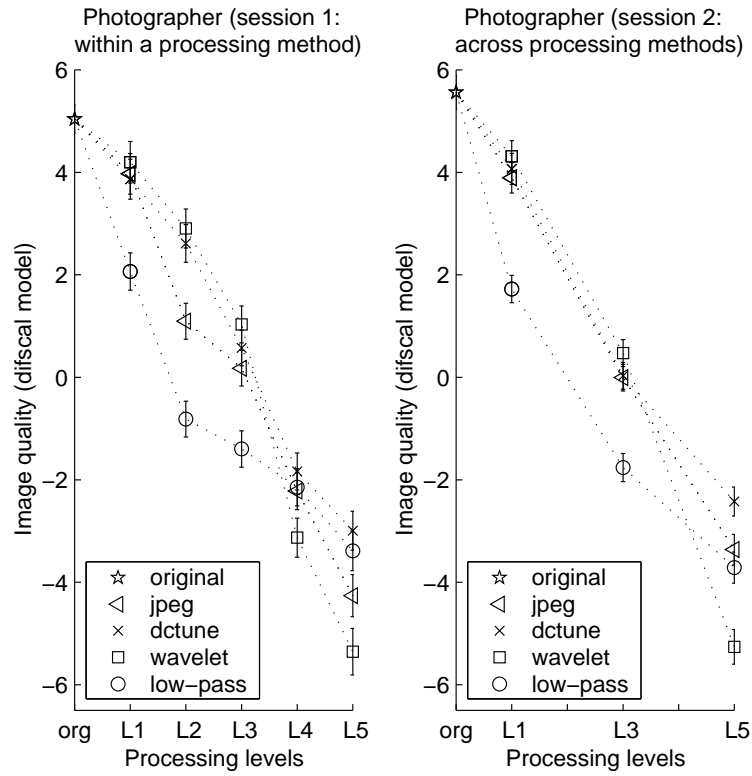


Figure 3.5: The results of DifScal obtained within a processing method (session 1) of the scene *photographer* are shown in the left panel. The DifScal results obtained across processing methods (session 2) are shown in the right panel. The x-axis shows the processing levels, L1 up to L5. The original is indicated by org. On the y-axis DifScal's scale values of the images are given. The different ranges of scale values for the processing methods indicate that in session 1 the extreme images are not mapped to the ends of the quality scale.

3.3. Experiments: processing methods

ferent quality range, especially for the scene *shopping-street*. The original (org) is considered as an image with impairment strength zero for all processing methods and therefore, in each panel, mapped to the same scale value. On the other hand, the most impaired images (L5) show differences in scale values between processing methods. For example, the quality range of the wavelet images is larger than the quality range of the JPEG and DC-Tune coded images. This indicates that in session 1 subjects do not map the extreme images of each identifiable class of distortions to the ends of the quality scale.

Equivalent quality range differences between the processing methods can be observed in the right panels of figures 3.4 and 3.5 for session 2.

Concluding, the results demonstrate that subjects do not tend to identify the distortion types and calibrate their quality scales for a particular distortion (by mapping the extreme images of each processing method to the ends of the quality rating scales). Hence, hypothesis I can be rejected for the parameter processing methods. The question remains whether the scale values of both sessions are comparable.

Linear relationship

The second hypothesis states that quality judgements obtained with and without explicit comparisons across processing methods are similar up to a linear transformation. In order to test this hypothesis the following analyses are performed.

Figures 3.6 and 3.7 show the linear relation for the scenes *shopping-street* and *photographer*, respectively. In these figures the scale values of the original, and those at processing levels L1, L3 and L5 are shown. On the x-axis the scale values of session 2 are given and on the y-axis the scale values of session 1. The horizontal error-bars are the S-estimates of session 2 and the vertical error-bars those of session 1. These S-estimates will be considered as confidence intervals, a range of values that are tenable for the scale values.

The scale values of sessions 1 and 2 show a highly linear relationship with a Pearson correlation coefficient of $r = 0.95$ for *shopping-street* and $r = 0.99$ for *photographer*. However, figure 3.6 illustrates that stimuli of different processing methods which obtain the same scale value in session 1 are discriminated in session 2, or the reverse, discriminated stimuli in session 1 are similar in session 2. Nevertheless, scale values of a particular processing method decrease monotonically with increasing degree of distortion in both sessions.

To test if the observed differences across processing methods are statistically significant, a two tailed t-test at the $p = 0.05$ level of significance with a normal distribution, is performed on the scale values obtained in the two sessions. The scale values of a particular stimulus are considered similar if

$$\frac{S1_i - S2_i}{\sqrt{\sigma_{S1_i}^2 + \sigma_{S2_i}^2}} < 1.96, \quad (3.1)$$

with $S1_i$ and $S2_i$ the scale values of stimulus i in sessions 1 and 2, and σ_{S1_i} and σ_{S2_i} the corresponding S-estimates.

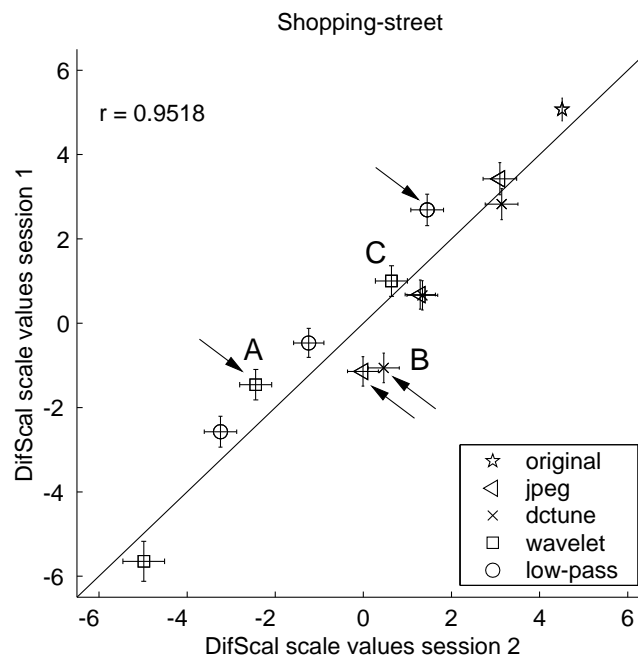


Figure 3.6: The scale values of the scene *shopping-street*. The scale values obtained across processing methods (session 2) are given on the x-axis and the scale values obtained within a processing method (session 1) on the y-axis. The arrows indicate those scale values that differ significantly between both sessions.

In figure 3.6, *shopping-street*, the scale values that differ significantly are indicated by arrows. For instance, in session 1 the JPEG image with Q-parameter 15 at approximately 0.5 bits per pixel (label "B" in figure 3.6) and the wavelet image at 0.3 bits per pixel (label "A" in figure 3.6) obtain a similar quality scale value. However, in session 2 the same JPEG image is comparable in quality with a wavelet coded image at 0.6 bpp (label "C" in figure 3.6). This shows that evaluating coders by means of the subjective data of sessions 1 or 2 can lead to different results. For this reason, the differences between both sessions are large enough to reject hypothesis II.

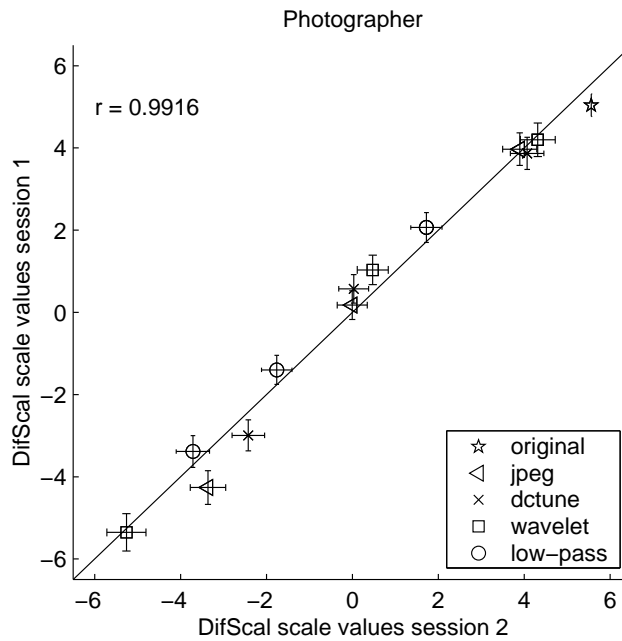


Figure 3.7: The scale values of the scene *photographer*. The scale values obtained across processing methods (session 2) are given on the x-axis and the scale values obtained within a processing method (session 1) on the y-axis. This figure illustrates that the scale values of both sessions are comparable.

No statistically significant difference is found between the scale values of sessions 1 and 2 in the scene *photographer*. The subjective data of both sessions is interchangeable and hypothesis II can not be rejected.

3.3.4 Discussion

Subjective comparison data across distortion types are needed in order to evaluate the performance of codecs which introduce different types of distortions. After all, the extent to which the bit-rate affects the perceived image quality is used to evaluate the performance

of a coder. This means a particular codec performs better if it guarantees similar image quality for an image at a lower bit-rate than other coders. Therefore, the observed differences between sessions 1 and 2 for the scene *shopping-street* can lead to different judgements about codec performances. For instance based on the data of session 1 the wavelet coder performs better than the JPEG coder, (see label "A" and "B" in figure 3.6). In this case the wavelet image coded at 0.3 bpp is similar in quality to a JPEG coded image at approximately 0.5 bpp. On the other hand, if the data of the second session are used JPEG seems to be slightly better than wavelet. The wavelet image coded at 0.6 bpp has similar quality as the JPEG image coded at approximately 0.5 bpp, (see label "A" and "C" in figure 3.6). The method of subjective testing can thus influence the evaluation of codecs.

However, whether the quality judgements of sessions 1 and 2 are the same seems to depend on the quality range of the various processing methods. In the *shopping-street* the quality range of the processing methods is not comparable. It is therefore possibly difficult for observers to judge the image quality between different types of distortions when the differently distorted images are not compared explicitly. In that case observers tend to stretch the quality range of the JPEG and DCTune coded images. As a result, the observed differences between quality judgements of session 1 and 2 indicate that subjects tend to use separate quality rating scales if no explicit comparisons are incorporated. However, no differences are observed between the quality judgements of both sessions for the scene *photographer*. In this case the quality range between the processing methods is less dissimilar. It follows that it is sufficient to use image-pairs with two images of identical processing method to obtain reliable quality judgements across distortion types, on the condition that the quality range of the distortion types is not too dissimilar. The advantage of this is that the experimental labor can be reduced. On the other hand if the quality range is too different explicit comparisons across processing methods are needed. Often it is hard to judge beforehand whether the quality range of differently distorted images is not too dissimilar. Therefore, also in this case it is better to compare differently processed images explicitly.

3.3.5 Conclusions

On the basis of the introduced distortions in the scene *shopping-street* it seems that observers tend to rate image quality on separate quality scales if no explicit comparisons across processing methods are incorporated in the stimulus set. However, the results gathered on the basis of the scene *photographer* seem to indicate that quality judgements can be linked across distortion types even though differently impaired versions of a scene were not compared explicitly. In contrast to the large quality range differences of the distortions in the scene *shopping-street* the quality ranges in the scene *photographer* are not too dissimilar. Thus it seems that the quality ranges of the various distortions determine whether explicit comparisons are needed.

3.4 Experiments: scene content

In the previous section we studied the effect of processing method on the judged image quality by combining differently processed images. In this section, we similarly study the effect of scene content on the judged image quality. The same type of impairment can manifest itself in different ways depending on the scene content. We assume that if different scenes can be identified and classified, observers may use separate rating scales for each scene.

The effect of stimulus presentation on subjective quality judgements was evaluated for two coding methods that introduce distinct distortions: wavelet coding and sequential baseline JPEG coding. They were evaluated in separate experiments. This was done to avoid any effect of coding method on the subjective quality ratings. Hence image pairs within one experiment were only varied in distortion strength and scene content, but not in type of distortion. The stimulus sets, experimental method and the obtained results are discussed next.

3.4.1 Stimulus sets

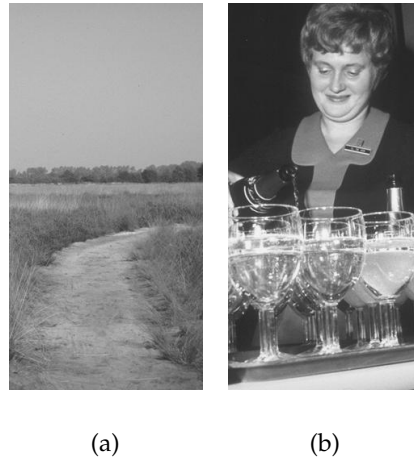


Figure 3.8: Original scenes: (a) *country-road* and (b) *woman*.

Four scenes were used in both experiments. All scenes were gray-scale images with a size of 240x480 pixels: two portraits referred to as *photographer* and *woman*, and two outdoor scenes referred to as *country-road* and *shopping-street*. The scenes *shopping-street* and *photographer* (see figure 3.3) were also used in the previous experiment. The originals of the additional scenes are shown in figure 3.8.

For the first experiment, all four scenes were compressed with a wavelet coder with an embedded zerotree coder (Said and Pearlman, 1996). Each scene was compressed at five bit-

rates, namely 0.15, 0.2, 0.3, 0.4, and 0.6 bits per pixel. The resulting set of test images contained both low quality images, distorted mainly by blur, and high quality images, with little noticeable impairment. No perceptual parameter is incorporated in this algorithm. This means that there may be perceptual quality variations between the various scenes coded at the same bit rate.

Two stimulus sets were constructed. In the first set the five wavelet coded versions and the original of a particular scene were pair-wise combined into 15 image pairs. Each pair was unique and consisted of two different instances of a scene. Since four scenes were included, the first stimulus set contained 60 image pairs.

The second stimulus set contained the original and three wavelet coded versions of each scene. To guarantee the same range of image quality as in the previous stimulus set, wavelet images compressed at 0.15, 0.3 and 0.6 bits per pixel were used. The coded versions and the originals of the four scenes, in total 16 images, were combined into 120 image pairs. Each pair was unique and consisted of two images with either identical scene content but different degrees of distortion, or different scene content and equal or different degrees of distortion.

For the second experiment similar stimulus sets were constructed with JPEG coded images. All four scenes were compressed by the sequential baseline JPEG compression algorithm (Pennebaker and Mitchell, 1993) of the Independent JPEG Group. Five images per scene were obtained by varying the Q-parameter. The Q-parameter settings were 15, 20, 25, 30 and 60. The resulting set of test images contained both low quality images, distorted mainly by blockiness, and high quality images, with little noticeable impairment. Contrary to wavelet coding, in JPEG coding the Q-parameter is a global quality indicator. This implies that in the case of equal quality of the originals, different scenes compressed with the same Q-parameter should have approximately the same perceived image quality.

As in the first experiment two stimulus sets were constructed. To obtain the first stimulus set of JPEG images the five coded JPEG versions and the original of a particular scene were combined into 15 image pairs. Each pair was unique and contained two images of the same scene coded at different Q-parameters. Since four scenes were included the first stimulus set contained 60 image pairs.

For the second stimulus set, three versions with Q-parameter 15, 25 and 60 and the original of each scene were combined into 120 image pairs, in a similar way as for the wavelet images. This second JPEG stimulus set contained unique pairs of images differing in scene content and/or degree of distortions as well as pairs of images solely differing in the degree of distortion.

3.4.2 Method

Viewing conditions, display and instructions were the same as in the experiments described in section 3.3.2. The images of each compression method, wavelet and JPEG, were presented in a separate experiment to six viewers. In the first session of both experiments, the image pairs of the first stimulus set were presented. In the second session, the second

3.4. Experiments: scene content

stimulus set was shown. The 120 image pairs in the second session were judged in two subsessions with a small break in between. The same viewers took part in both sessions of an experiment, but no viewer participated in both experiments. The subjects had to rate image quality difference between -4 and +4 as described in section 3.3.2

3.4.3 Results

The same analyses are applied as in section 3.3.3. Each coding method, wavelet and JPEG, is analyzed separately.

For each session, the quality difference data are transformed into quality scale values on an interval scale by means of DifScal Boschman (2001). For each session, this results in a stimulus configuration, with a quality scale value for each image. First we will investigate if in session 1 the extreme images, the original and the most impaired image of each processing method, are mapped to the ends of the quality scale. After that, the effect of quality judgments across scene content is studied by comparing the stimulus configurations of sessions 1 and 2.

DifScal stimulus configurations

For the analysis of the data of session 1 we will assume that the rating scale used by an observer is fixed across scenes and that therefore quality differences of the 24 stimuli (for each scene the original and 5 coded versions of it) are judged on the same scale. The results of the quality comparison between these 24 stimuli are collected in *one* frequency file. The observers rated the quality difference of 24 pairwise combined images in 9 categories. From this a frequency file is generated with nine 24x24 lower triangle matrices. In each 24x24 matrix the four 6x6 lower triangles at the diagonal contain the frequencies for each scene. The resulting scale values are shown in the left panels of figures 3.9 and 3.10 for the wavelet and JPEG images, respectively. The x-axis shows the bit-rates or Q-parameters of the images. The scale values are shown on the y-axis. In each figure the S-estimates are indicated by the error-bars. In section 3.3.1, the original is the same for all processing methods because only one scene was used per experiment. However, in this case the judged quality of the original of each scene is different due to the use of different scenes.

For session 2 the 16 pairwise compared images are ordered in a frequency file of nine 16x16 lower triangle matrices. In a similar way as in section 3.3.3 the scale values and S-estimates of the second session are linearly transformed. The results are shown in the right panels of figures 3.9 and 3.10.

Quality range

The first hypothesis states that the extreme images of each identifiable class of distortions are mapped to the ends of the quality scale when no explicit comparisons are incorporated in the experiment.

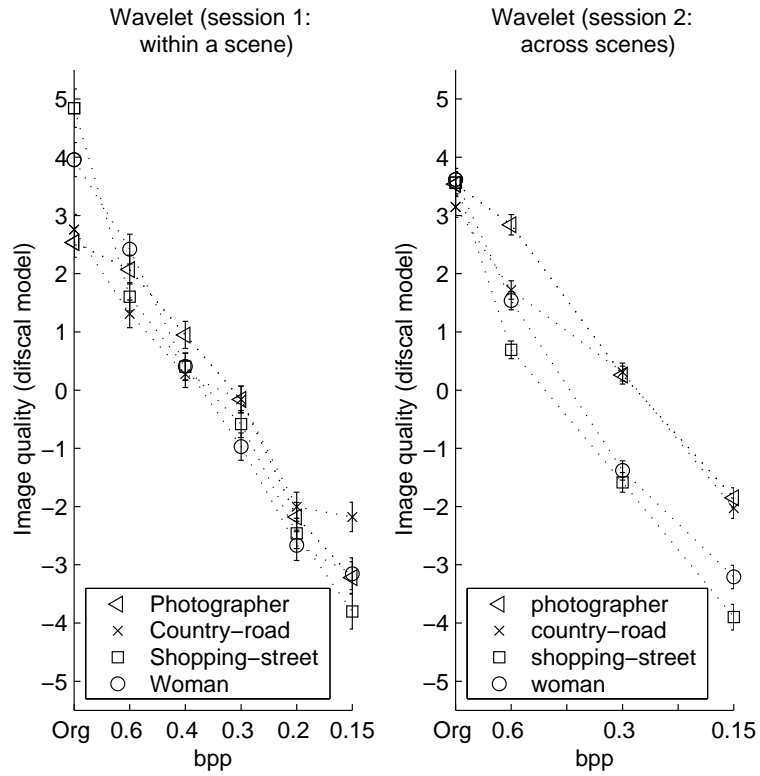


Figure 3.9: The results of DifScal obtained within a scene (session 1) are shown in the left panel for the wavelet coded images. The results of DifScal across scenes (session 2) are shown in the right panel. The x-axis shows the bits per pixel (bpp) from 0.6 to 0.15. The original is indicated by org. On the y-axis DifScal's scale values of the images are given. The different ranges of scale values for the scenes indicate that in session 1 the extreme images are not mapped to the ends of the quality scale.

3.4. Experiments: scene content

The left panel of figure 3.9 illustrates that in session 1 the original and the most impaired image of each scene are not all mapped to the ends of the quality rating scale. The originals of the scenes *shopping-street* and *woman* are judged to be of better quality than those of the scenes *photographer* and *country-road*. For the most impaired wavelet images at 0.15 bpp this is less obvious. The scenes seem to be of similar quality except for the scene *country-road*. Because of this the scenes are judged on different parts of the quality scale. For instance, the wavelet coded images of the scene *shopping street* cover a wider range in judged quality than the wavelet images of the scene *country-road*.

In the left panel of figure 3.10, showing the JPEG coded images, the judged image quality of the various scenes is less diverse. But also in this figure it can be seen that the originals are not judged to have similar quality. For example, the original of the scene *country-road* is judged to be of less quality than the original of the scene *woman*. For the most impaired JPEG images with a Q-parameter of 15 also small quality differences can be observed between the various scenes. In this case, the scene *country-road* is judged to be of better quality than the scene *photographer*, both coded with the same Q-parameter 15. This implies that also for the JPEG images the most extreme images of each scene are not mapped to the ends of the quality scale.

The above analysis shows that observers do not tend to calibrate the image quality rating scale for each scene separately to the ends of the rating scale. For this, hypothesis I can be rejected for images that vary in scene content. In both coding methods the image quality of the most extreme images is not mapped to the ends of the quality scale.

As can be seen most clearly in figure 3.9, the quality judgements of sessions 1 and 2 seem to lead to different results. In the first session, without explicit comparisons across scenes, the image quality of the originals are judged to be different while in the second session, with explicit comparisons across scenes, the image quality of the originals is judged to be similar. In the case of the JPEG images (figure 3.10) this is less obvious. In the following it will be tested whether the quality judgements of sessions 1 and 2 are comparable.

Linear relationship

The second hypothesis states that quality judgements obtained with and without explicit comparisons across processing methods are similar up to a linear transformation. In order to test this hypothesis the following analysis was performed.

The linear relations between the scale values of sessions 1 and 2 are shown in figures 3.11 and 3.12 for the wavelet and JPEG coded images, respectively. The data points represent the scale values of the originals and those at bit-rates 0.6, 0.3 and 0.15 or Q-parameters 60, 25 and 15. On the x-axis the scale values of session 2 are shown and on the y-axis the corresponding scale values of session 1.

For both coding methods, the scale values of sessions 1 and 2 show a highly linear relationship with a Pearson correlation of $r = 0.96$ for wavelet coding and $r = 0.98$ for JPEG coding. A t-test is performed on the scale values obtained in sessions 1 and 2. In figures 3.11 and 3.12 the scale values that differ significantly between sessions are indicated by arrows.

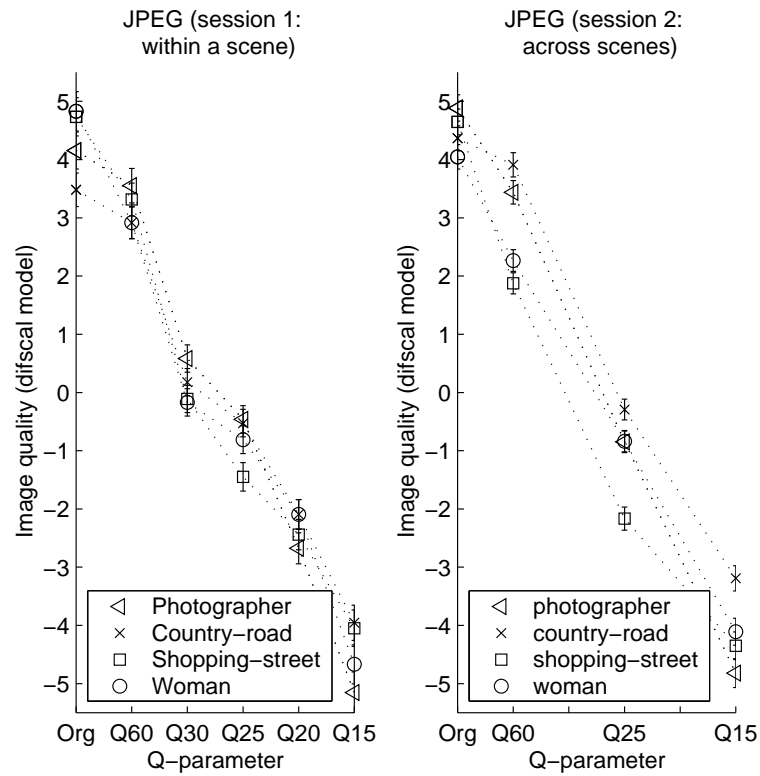


Figure 3.10: The results of DIFScal obtained within a scene (session 1) are shown in the left panel for the JPEG coded images. The results of DIFScal across scenes (session 2) are shown in the right panel. The x-axis shows the Q-parameter, Q60 to Q15. The original is indicated by org. On the y-axis DIFScal's scale values of the images are given. The different ranges of scale values for the scenes indicate that in session 1 the extreme images are not mapped to the ends of the quality scale.

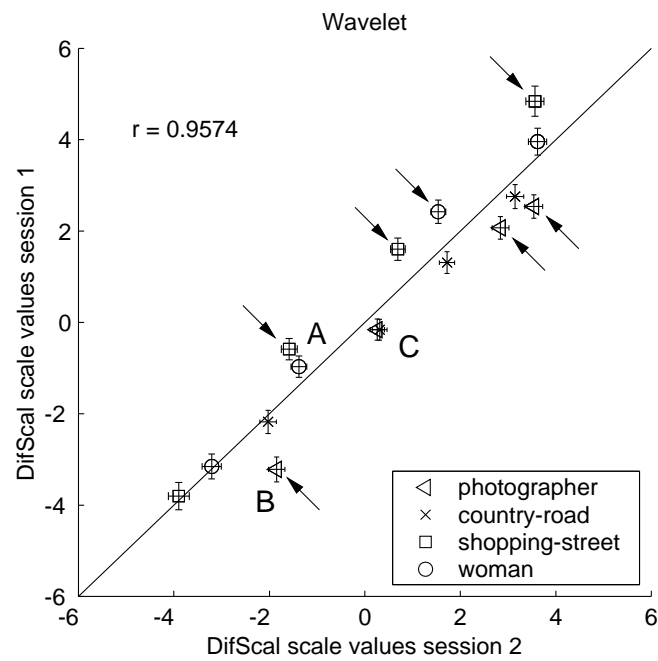


Figure 3.11: The scale values of the wavelet coded scenes. The scale values obtained across scenes (session 2) are given on the x-axis and the scale values obtained within a scene (session 1) on the y-axis. The arrows indicate those scale values that differ significantly between both sessions.

The effect of the significantly different scale values of sessions 1 and 2 can be interpreted in the same way as for the scene *shopping-street* in the previous section 3.3.3. In both sessions, the scale values of a particular scene decrease monotonically with decreasing bit-rate or Q-parameter. On the other hand, the results of both sessions are not comparable across scenes. The wavelet coded and JPEG coded images that differ in scale value are in session 1 overestimated for the scenes *shopping-street* and *woman* and underestimated for the scene *photographer* or *country-road*, see figures 3.11 and 3.12.

For instance, in session 2, the quality of the wavelet coded scenes *shopping-street* and *woman* at 0.3 bpp (label "A" in figure 3.11) is comparable to the scene *photographer* at a lower bit-rate, namely 0.15 bpp (label "B" in figure 3.11) while in session 1 these images are comparable in quality with the scene *photographer* at 0.3 bpp (label "C" in figure 3.11). The same can be observed for the Q-parameter. The scenes *shopping-street* and *woman* coded at Q-parameter 60 (label "A" in figure 3.12) are of comparable quality in session 1 with the scenes *photographer* and *country-road* at the same Q-parameter of 60 (label "B" in figure 3.12). This is in contrast to session 2 where an apparent difference is found between the quality of these scenes. In this case the quality of the scenes *photographer* and *country-road* is higher than the quality of the scenes *shopping-street* and *woman*. For this reason, hypothesis II can be rejected in both cases.

3.4.4 Discussion

We demonstrated that the method of subjective testing can influence the evaluation of a codec. The observed differences between sessions 1 and 2 for wavelet as well as JPEG coded scenes lead to different results. It was shown that in session 1 observers discriminate less between the image quality of scenes coded at the same bit-rate (as in wavelet coding) or at the same JPEG Q-parameter than in session 2. This indicates that subjects tend to use a separate rating scale for each scene. Therefore, subjective comparison data across scene content is needed to evaluate the performance of a particular codec.

The results of session 2 showed that not all JPEG coded images with equal Q-parameter but different scene content are judged the same. This indicates that JPEG's Q-parameter is not scene-independent. However, compared to the effect of coding levels, the scene effect is minimal, except for high quality images, namely the original of the scene *country-road* and its JPEG coded version at Q-parameter 60.

On the other hand, wavelet coded scenes of the same bit-rates are rated to have dissimilar quality. In particular, the scenes *shopping-street* and *woman* are qualitatively worse than the scenes *photographer* and *country-road*. This indicates that these scenes suffer more from the introduced distortions than the other scenes. This has been compensated for in JPEG coding. Here the quality differences between the scenes at identical Q-parameter have been minimized by compressing the scenes *shopping-street* and *woman* at higher bit-rates than the scenes *photographer* and *country-road*.

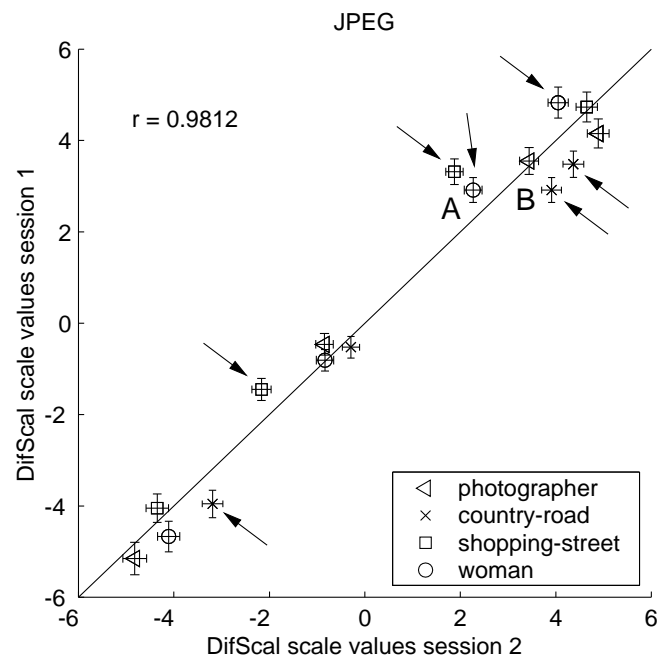


Figure 3.12: The scale values of the JPEG coded scenes. The scale values obtained across scenes (session 2) are given on the x-axis and the scale values obtained within a scene (session 1) on the y-axis. The arrows indicate those scale values that differ significantly between both sessions.

3.4.5 Conclusions

For two different coding methods, JPEG and wavelet coding, we demonstrated that subjective tests, with image pairs containing two images of identical scene content only, are not sufficient to measure the perceived image quality across scenes. It seems that reliable subjective data can only be obtained by comparing the scenes explicitly.

3.5 Concluding remarks

In this chapter we discussed two experimental methods to obtain quality judgements across distortion types or scene contents. The question is whether this can be achieved using only image pairs containing two images of identical distortion type or scene content. One view is that in such a case subjects tend to map the extreme images, e.g. the original and the most impaired image, to the ends of a quality scale and thus use separate rating scales for each distortion type or scene. If the observers actually calibrate the quality rating scale for each identifiable class the experimental procedure is not suited to obtain accurate quality judgements across these classes. However, this is not the case as was substantiated in sections 3.3 and 3.4 for processing methods and scene contents, respectively. In these sections we showed that observers discriminate between the various distortion types or scene contents even though no explicit comparisons were made to obtain quality judgements.

However, this does not necessarily imply that quality judgements obtained from two images containing identical processing method or scene content are comparable to those obtained from explicit comparisons. Even when the images of each processing method or scene are scaled on different parts of the quality scale, the quality judgements are not always linked across processing method or scene content.

Comparable results between the two experimental procedures are only obtained in one case, namely if the quality range of the various processing methods is not too dissimilar. In this case the observers seem to be able to link the quality rating scales without explicit comparisons. In this case the experimental time can be reduced since quality judgements without explicit comparisons are sufficient and adequate. On the other hand, if the quality ranges between processing method differ the results are not comparable, even though without explicit comparisons observers discriminate between the processing methods.

As for processing methods, the quality range determines whether explicit comparisons are needed or not. The experiments in section 3.4 showed that, especially for the JPEG coded images, the results of both sessions are not comparable, although the quality range of the various scenes differ not so much. This means that explicit comparisons are necessary to determine the image quality across scene contents.

In general we may conclude that observers use separate quality scales for identifiable classes of stimuli if these are not compared explicitly. This may have consequences if the performance, bit-rate versus image quality, of different coders are compared. As shown in section 3.3, the conclusions drawn from quality judgements without explicit comparisons can be misleading.

Chapter 4

A Single-Ended Blockiness Measure for JPEG-coded Images

Abstract

In three subjective experiments, dissimilarity data and numerical category scaling data were obtained to determine the underlying attributes of image quality in sequential baseline coded JPEG images. Although several distortions were perceived, namely blockiness, ringing and blur, the subjective data for all attributes were highly correlated, so that image quality could approximately be described in one dimension. We therefore proceeded by developing an instrumental measure for one of the distortions, namely, blockiness. In this chapter a single-ended blockiness measure is proposed, i.e., one that uses only the coded image. Our approach is therefore fundamentally different from most (double-ended) image quality models that need both the original and the degraded image. The proposed measure is based on detecting the low-amplitude edges that result from blocking and estimating the edge amplitudes. Because of the approximate one-dimensionality of the underlying psychological space, the proposed blockiness measure also predicts the image quality of sequential baseline coded JPEG images.

4.1 Introduction

The present demand for data storage and transmission of large quantities of images necessitates the use of data compression. The international JPEG (Joint Picture Expert Group) standard for image compression emerged after an extensive comparison of existing algorithms. Four modes of operation are supported by JPEG, one lossless mode: (1) the sequential lossless mode, and three lossy coding techniques: (2) the sequential DCT (Discrete Cosine Transform) based mode, (3) the hierarchical mode, and (4) the progressive sequential mode. The baseline system is a particular restricted form of the sequential DCT-based mode and is supported by all JPEG decoders (Pennebaker and Mitchell, 1993).

JPEG is a block-based coding algorithm. The original image is first subdivided into blocks of 8x8 pixels and the blocks are subsequently coded independently using a DCT transform. The bit-rate and image quality of JPEG-coded images are mainly determined by the degree of quantization of the DCT coefficients. Quantization implies that the spatial frequency components in each 8x8 block can only be reconstructed approximately. As a consequence, degradations such as blockiness, ringing and blur are introduced in the reconstructed image. Blockiness artifacts are visible as discontinuities at adjacent 8x8 pixel block boundaries. The sudden intensity changes are most conspicuous in uniform regions and are caused by a coarse quantization of the DC coefficients on the one hand and the absence of low-frequency AC coefficients on the other hand. Ringing artifacts appear as ringing patterns around sharp edges in the image, due to coarse quantization of the high frequency AC coefficients. Quantization of the low frequencies as well as of the high frequencies causes image blur and loss of detail. In the experiments described below we aim at obtaining a better understanding of the perceived strengths of these attributes and their relation to overall image quality.

As described by Eskicioglu and Fisher (1995) and Ahumada and Beard (1998), most frequently used instrumental image quality measures, such as the mean square error (mse), are based on the statistical distribution of pixel value differences between the original and the degraded image or between processed versions of both. Such measures often do not provide reliable predictions for perceived image quality, especially not when quality is determined by multiple attributes (Martens and Meesters, 1998). This occurs for instance when different coding techniques are compared. These measures represent image quality as a single scalar value, and hence do not reveal the effect of separate impairments on image quality. A different approach is to recognize that perceived image quality is multidimensional (Ahumada and Null, 1993), i.e., several distinct degradations can be perceived in an image, and they can all contribute to the overall image quality impression. Kayargadde and Martens (1996c,d) used this approach to model the image quality of images degraded by blur and noise. A similar approach is used in this chapter to obtain a better understanding of how different attributes influence the perceived quality of sequential baseline coded JPEG images.

As stated before, most existing instrumental measures for image quality are based on some distance between the coded/degraded and original image. In such cases, all detected differences are interpreted as degradations. No a priori assumptions are needed about the kind of distortions introduced. A disadvantage of this approach is of course that the orig-

4.2. Experiments

inal image must be available, and in perfect registration with the coded image. Especially in the case of coded images these requirements can most often not be met (ITU-WP-2/12, 1995). Human observers on the contrary are perfectly able to judge image quality without an explicit reference image being available. We similarly propose a single-ended technique to estimate blockiness that is based solely on the coded image. Although the proposed measure may potentially be applied to estimating blockiness in MPEG-2 video, it has only been tested on JPEG-coded images at this point.

The chapter is organized as follows. Section 4.2 describes the experiments carried out to investigate the underlying attributes of image quality in sequential baseline coded JPEG images. The relationship between the distortions blockiness, blur and ringing, and the overall perceived image quality are determined. Since the attributes turn out to be highly correlated, it is proposed that a good prediction for image quality can be based on an estimate for any of these attributes. Blockiness seemed the easiest attribute to estimate by means of an instrumental measure. In section 4.3 we therefore introduce a single-ended instrumental measure for blockiness. The resulting measure is compared with subjective blockiness data in section 4.4.

4.2 Experiments

In this section we will show experimentally that perceived image quality of sequential baseline coded JPEG images can be described in one dimension. In subsection 4.2.1, image quality is studied for two sets of JPEG-coded images by means of dissimilarity scaling and numerical category scaling. Dissimilarity scaling is used to measure the subjective difference between two images without a-priori knowledge of the impairment types. Numerical categorical scaling is then used to identify the perceived strength of particular predefined attributes. Although the measured properties are different in these scaling methods, both show that the image quality of JPEG coded images can be described in one dimension. In the first experiment the distortion in the images varied over a wide range, while in the second experiment a subset of these images with less conspicuous blockiness was presented. A third experiment in subsection 4.2.2 was carried out to collect data for a larger group of subjects using a double stimulus method that is known to have high sensitivity and accuracy (de Ridder, 1996; Parducci and Wedell, 1986). Table 4.1 summarizes the stimulus sets and the scaled attributes in all experiments.

4.2.1 Image quality and its underlying attributes

Observers Six and five subjects took part in the first and second experiment, respectively. All subjects were familiar with scaling experiments, image quality and typical coding artifacts such as blockiness, blur and ringing. They all had normal or corrected-to-normal visual acuity.

Display and viewing conditions The images were displayed for 5 seconds on a BARCO-CCID-7351B high-resolution, non-interlaced 50-Hz monitor, placed in a dark room. The

Table 4.1: Overview of experiments.

	experiment 1	experiment 2	experiment 3
stimulus set	$Q = 20, 25, 30, 40, 50, 60, 70, 80, 90$	$Q = 30, 40, 50, 60, 70, 80, 90$	$Q = 20, 25, 30, 40, 60, \text{original}$
session 1	dissimilarity	dissimilarity	-
session 2	quality blockiness blur	quality blockiness ringing	quality blockiness -

monitor was calibrated to have a grey-value-to-luminance characteristic equal to

$$L = \max[L_{min}, L_{max} (g/g_{max})^\gamma], \quad (4.1)$$

for $0 \leq g \leq g_{max}$, with $g_{max} = 255$, $L_{min} = 0.2 \text{ cd/m}^2$, $L_{max} = 60 \text{ cd/m}^2$ and $\gamma = 2.5$. The adaptation time between 2 successive stimulus presentations was determined by the time that subjects took to input their response (on a keyboard), but with a minimum of 2 seconds. The adaptation field was uniform and had a luminance of 13 cd/m^2 . This value was approximately equal to the average of the mean luminances of the images. The observers were seated at 0.80 m from the monitor, resulting in a ratio of viewing distance to display height of 3.2. The corresponding visual angle between successive pixels was 2.03 minutes of arc. A short viewing distance¹ was preferred because in many applications using JPEG images, the images are viewed on a computer monitor screen from a short distance.

Stimulus set Four different natural test scenes, referred to as *boat*, *child*, *girls* and *light-house* (see figure 4.1), were acquired from a Kodak PhotoCD demonstration disc. In order to enable display of two images simultaneously on the screen the size of the original 8-bit grey-scale images was cropped to 240x480 pixels. The Baseline Sequential JPEG compression software package of the Independent JPEG Software Group² with default quantization table was used to generate for each scene different versions at various compression rates. The compression rate and visual fidelity were determined by the ‘Q-parameter’. Images with a high compression ratio were obtained by low ‘Q-values’ (e.g. 20 and 25), and therefore contained the most conspicuous distortions. Compressed images with less distortion were generated by selecting high ‘Q-values’ (e.g. 60). For the first experiment the ‘Q-parameter’ was varied within the set $Q_1 = \{20, 25, 30, 40, 50, 60, 70, 80, 90\}$, while for the second experiment it was taken from the set $Q_2 = \{30, 40, 50, 60, 70, 80, 90\}$. The images in stimulus set Q_1 contained distortions that varied over a wide range, whereas in stimulus set Q_2 , the distortions varied over a more limited range. The original images were not included in the experiments, because they were hardly distinguishable from the images with $Q = 90$.

¹For broadcast applications, a viewing distance of six times the image height, corresponding to a visual angle between pixels of 1 arcmin, is standard.

²<http://www.iijg.org/>

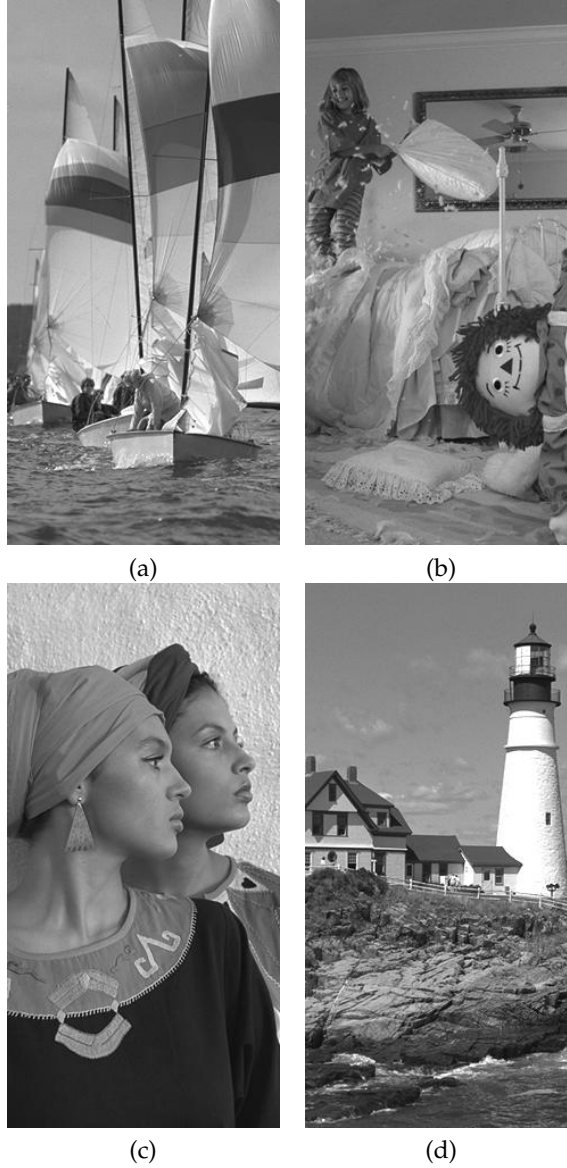


Figure 4.1: Original gray-scale images of the scenes (a) *boat*, (b) *child*, (c) *girls* and (d) *lighthouse*.

Method I: dissimilarity ratings During the first session of both experiments, dissimilarity scores were obtained for each scene for stimulus pairs (i, j) . In the first experiment $i, j \in Q_1$ with $j < i$, resulting in 36 distinct stimulus pairs. The 21 distinct stimulus pairs in the second experiment were obtained by $i, j \in Q_2$ with $j < i$. The stimulus pairs were presented in random order and displayed for a fixed time of 5 seconds. Each pair contained two JPEG-coded images of the same scene, simultaneously displayed on the right-hand side and left-hand side of the screen, respectively. The subjects had to rate the dissimilarity between the two images on an 11-point numerical category scale ranging from 0 to 10. A score of 0 indicated no perceived dissimilarity and a score of 10 indicated the largest perceived dissimilarity. Before the actual experiment started the subjects took part in a training session.

Method II: numerical category scaling The second session of both experiments consisted of three subsessions in which the subjects were asked to judge the perceived strength of three attributes. In experiment 1 these attributes were blockiness, blur and image quality. In experiment 2, they were blockiness, ringing and image quality. Each attribute was judged in a separate subsession on an 11-point numerical category scale ranging from 0 to 10. Absence of any perceptual occurrence of an attribute (blockiness, ringing or blur) in the image should be rated by 0. The image quality should be judged 10 for the best and 0 for the worst quality perceived. In each subsession, the different realizations of the scenes were presented three times in random order. Before each subsession a training session was performed by the subjects.

In the second session of both experiments, subjects had to judge the most prominent attributes. They were different in both experiments, due to the different stimulus range used. Stimulus set Q_1 consisted of images with both highly visible distortions and almost no visible distortions, and the most conspicuous distortions were blockiness and blur. The range for the ringing impairment was small in comparison to the range for blockiness and blur. The images in stimulus set Q_2 were less distorted. In these images, the range over which blur varied was small in comparison with the range over which blockiness and ringing varied. Therefore, in this second set the most obvious impairments were blockiness and ringing.

Results The dissimilarity data resulting from the first sessions can be summarized in a 9×9 and 7×7 dissimilarity matrix for each subject, respectively. These matrices were input into the multidimensional scaling program MULTISCALE (Ramsay, 1991) and transformed into distances (Kayargadde and Martens, 1996c,d). According to the specified dimension, Euclidean coordinates are estimated for the stimuli such that the distances between the stimulus coordinates correspond to transformed dissimilarities. In the resulting two-dimensional stimulus configurations, the first dimension is most dominant.

The numerical category scaling data for the attributes blockiness, ringing, blur and image quality were analyzed as described below. For each attribute, the mean of the three judgments was calculated. Next, the mean scores were transformed to z-scores per subject and scene to compensate for a potentially different use of the numerical scale by different subjects. The z-score transformation translates the overall mean to the origin and normalizes the overall variance to unity. The effect of this z-score processing is that all subjects are

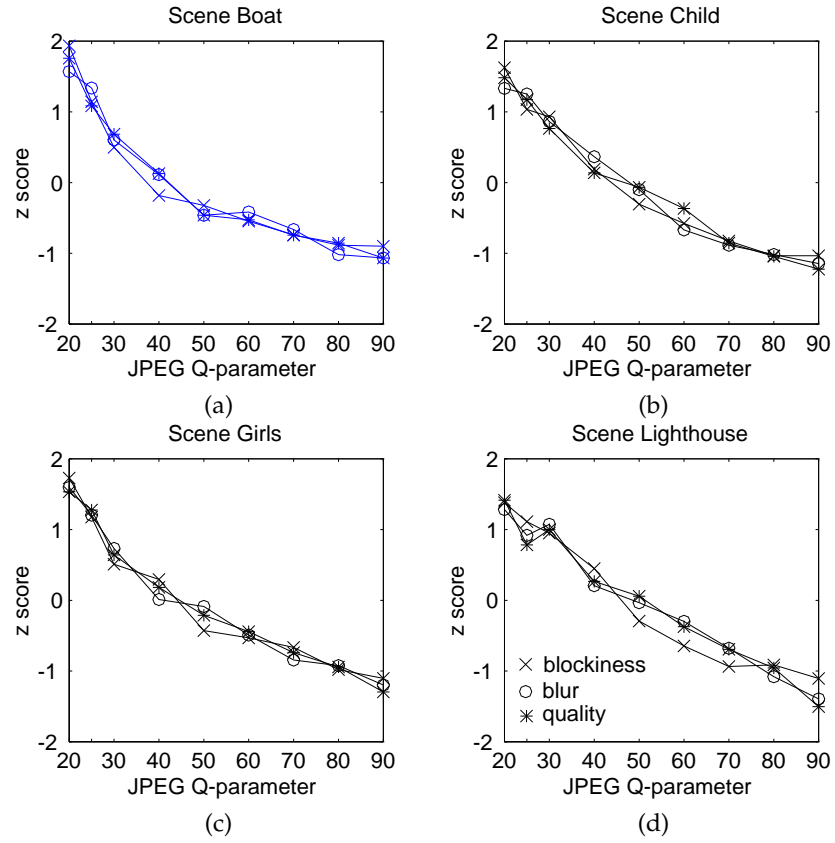


Figure 4.2: Numerical category scaling results of stimulus set Q_1 . The graphs present the z-scored average taken over all subjects of the scaled image quality and strength of the impairments blockiness and blur of the four scenes (a) *boat*, (b) *child*, (c) *girls* and (d) *lighthouse*. The distortions in this set vary over a wide range, they are most striking in the images $Q = 20$ and $Q = 25$.

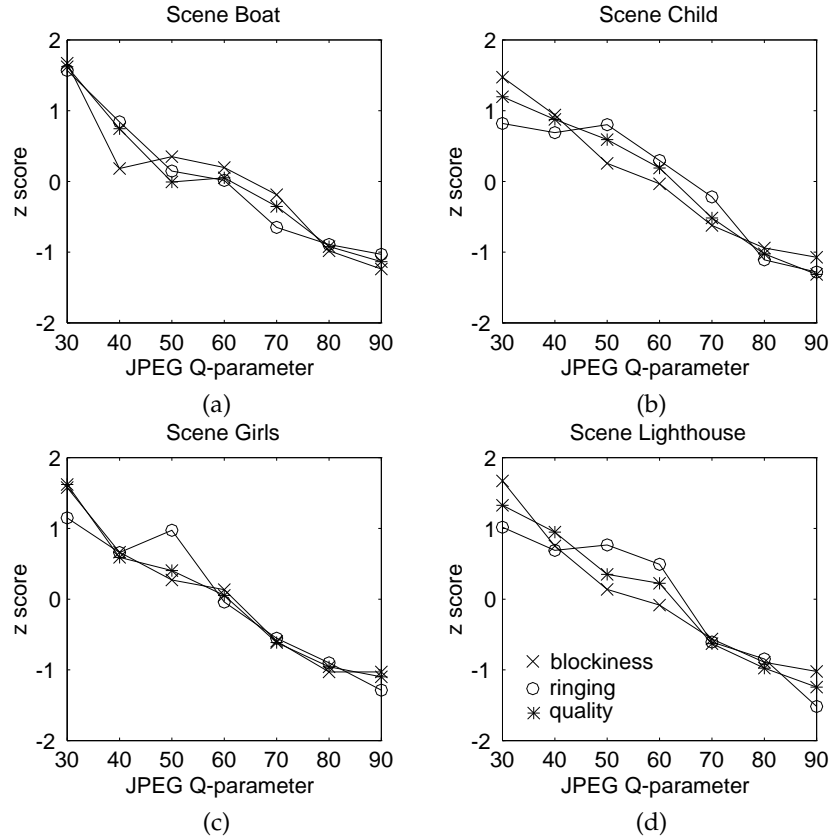


Figure 4.3: Numerical category scaling results of stimulus set Q_2 . The graphs present the averaged z-scores taken over all subjects of the scaled image quality and strength of the impairments blockiness and ringing of the four scenes (a) *boat*, (b) *child*, (c) *girls* and (d) *lighthouse*. Q_2 is a subset of stimulus set Q_1 , the low quality images of Q_1 , with the most striking distortions ($Q = 20$ and $Q = 25$), are omitted in this stimulus set.

4.2. Experiments

attributed the same weight in the overall averaging. Figures 4.2 and 4.3 show the numerical category scaling results for stimulus set Q_1 and stimulus set Q_2 , respectively. In these figures, the signs of the quality scores were reversed to facilitate the comparison with the attributes. Within the stimulus set, the attributes vary in a highly correlated way (see Table 4.2). The impairments are particularly conspicuous in JPEG images with a high compression ratio, and their visibility jointly decreases when the compression ratio decreases.

Table 4.2: The linear correlations between the scaled attributes in all experiments. The correlations are calculated separately for each of the four scenes: *boat*, *child*, *girls* and *lighthouse*.

Scene	Experiment 1		Experiment 2		Experiment 3
	blockiness versus blur	blockiness versus quality	blockiness versus ringing	blockiness versus quality	blockiness versus quality
<i>boat</i>	0.98	-0.99	0.93	-0.95	-0.99
<i>child</i>	0.99	-0.99	0.89	-0.97	-0.99
<i>girls</i>	0.98	-0.99	0.93	-1.00	-0.99
<i>lighthouse</i>	0.97	-0.96	0.88	-0.97	-0.99

The scaled attributes, blockiness, blur, ringing and quality were fitted into the stimulus configurations, obtained from dissimilarity scaling, by means of linear regression. The direction of the regressed attributes is approximately opposite to the direction of the regressed quality data (for more detail see Meesters and Martens (1999)). Hence, this gives additional evidence for a 1D stimulus configuration in which all attributes are linearly correlated.

In summary, visual inspection of the JPEG-coded image material reveals three attributes: blockiness, blur and ringing. The conducted experiments, however, imply that image quality can be described by any of these attributes, since they all have high linear correlation with quality. Therefore, in section 4.3 we will introduce an instrumental measure for blockiness that can also be used to predict the image quality of sequential baseline coded JPEG images. It should be noted that the blockiness measure is only expected to also predict quality in case the images are coded with a fixed quantization table that is scaled by varying the Q-parameter. This is the most commonly used practice in JPEG coding. In such case, the measured strengths of the distortions increase jointly as the perceived quality decreases. This probably does not hold any longer if the quantization matrix is not simply scaled, but varied in a more general way.

4.2.2 Blockiness in natural images

We now describe a third experiment in which subjective quality and blockiness ratings were gathered to evaluate the instrumental blockiness measure described in the next section.

Observers Ten subjects took part in this experiment. All subjects had a normal or corrected-to-normal visual acuity.

Display and viewing conditions These conditions are identical to those mentioned in section 4.2.1.

Stimulus set The same four natural test scenes, *boat*, *child*, *girls*, and *lighthouse*, as described in section 4.2.1 were used to generate stimulus set Q_3 . Five JPEG-coded images were derived for each scene with the Q-parameter set to 20, 25, 30, 40 and 60. The original of each scene was included as reference which resulted in the stimulus set $Q_3 = \{20, 25, 30, 40, 60, \text{original}\}$.

Method: Comparison scaling Comparison scores were obtained for 15 stimulus pairs (i, j) per scene, with $i, j \in Q_3$ and $j < i$. The pairs of images were presented in random order. Each pair contained two images of the same scene, simultaneously displayed on the right-hand side and left-hand side of the screen. The subjects were asked to rate the quality difference and the blockiness difference between the two images on a scale from -5 to +5. They were instructed to base their quality judgements on all perceivable distortions and not only on blockiness. The numerical value given should indicate the perceived difference in image quality or blockiness, while the sign should indicate which image is preferred or has the highest degree of blockiness, respectively. No perceived difference should be judged as 0 and the largest perceivable difference between two images should be judged as 5. Before the actual experiment started the subjects took part in a training session.

Results Since the judgements were based on images of the same scene, data analysis was performed for each scene separately. The comparison data are analyzed using the in-house software tool DIFSCAL (Boschman and Roufs, 1997; Boschman, 2001). The model underlying DIFSCAL is Thurstone's law of comparative judgements (Torgerson, 1958). This model assumes that the paired judgements are measured on an psychological internal scale with Gaussian noise distribution. The frequency distribution per category for each stimulus is the input to DIFSCAL. From this frequency distribution, stimulus scale values are calculated. The standard deviation of the Gaussian noise is used as the unit value for these scale values. The resulting subjective quality and blockiness data (see figure 4.4) are again highly correlated, with linear correlation values of -0.99 for all scenes, see Table 4.2.

4.3 Blockiness model

In the previous section it was proposed to base the image quality prediction of sequential baseline coded JPEG images on the prediction for only one of the attributes. Blockiness and ringing are distortions that introduce new edges. Block boundaries are oriented horizontally or vertically which is important a priori information that can be used to simplify the implementation of a blockiness measure. Ringing and blurring on the contrary are related to existing edges in the image, and may hence have arbitrary orientations. Since the refer-

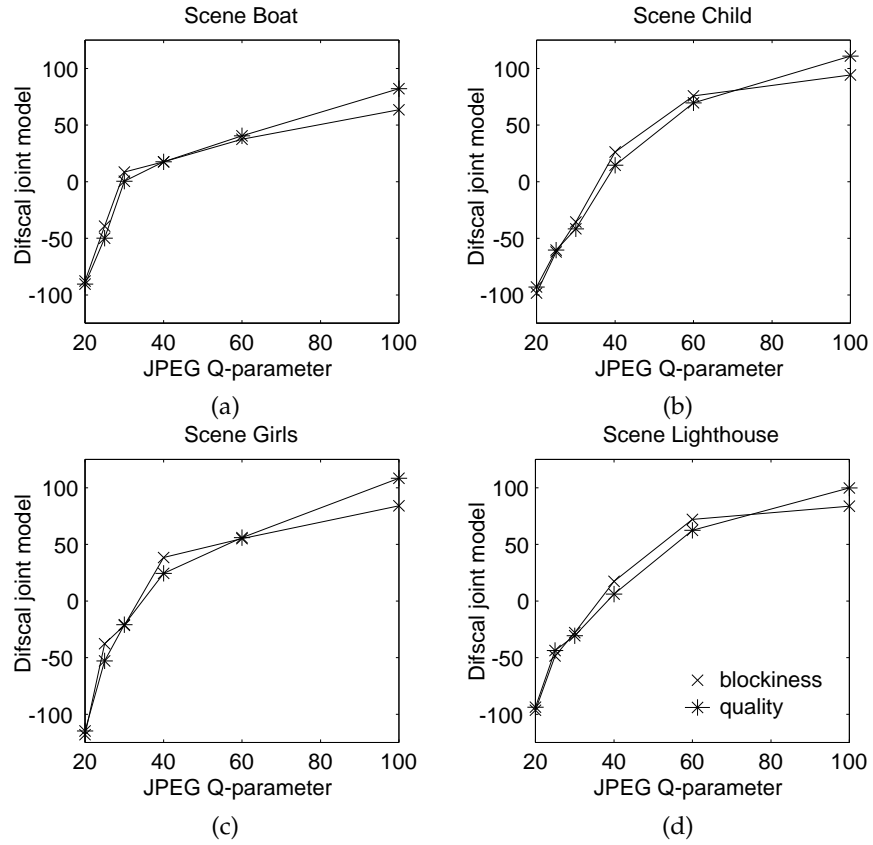


Figure 4.4: Comparison of scaling results for stimulus set Q_3 . Subjective quality and blockiness scores, obtained with DifScal, are plotted against the JPEG Q-parameter for four scenes (a) *boat*, (b) *child*, (c) *girls* and (d) *lighthouse*.

ence image is not available in a single-ended measure, added edges are most likely easier to detect than lost image information. These arguments motivate why we decided to develop a blockiness measure, rather than a measure for ringing or blur.

Blockiness is often expressed in terms of added edge energy, derived from the difference between Sobel-filtered original and coded images. These measures use mostly first order statistics such as average, standard deviation and root mean square error, and express the added edge energy into one single scalar value (Lin and Mersereau, 1995; ITU-WP-2/12, 1995; Webster *et al.*, 1993; Beerends, 1997). A more sophisticated model, developed by Karunasekera and Kingsbury (1995), is based on the sensitivity of the human visual system to vertical and horizontal edges. The difference image between the original and the degraded image is weighted based on the contrast sensitivity of the visual system to spatial frequencies (at different mean luminances). In these double-ended models, the blockiness measure uses both the degraded and the original image. We will show in this section that edge information needed for blockiness estimation can also be recovered without explicit use of the original image.

The *first computational assumption* underlying our blockiness measure can be phrased as follows: ‘Horizontal and vertical low-amplitude step edges present in a (coded) image arise from the coding process and are not present in the original image. These structures are therefore to be interpreted as artifacts’. We describe how these low-amplitude step edges can be detected and how the edge parameters (most noticeably the edge amplitude and edge blur) can be estimated using the Hermite transform as a tool for analysis (Martens, 1990b).

The *second computational assumption* underlying our blockiness measure is that blockiness is related to some first order statistic of the estimated edge amplitudes (i.e., that it can be derived from the histogram of these edge amplitudes). It is a simplification which is needed in order to combine the estimated edge parameters into a single number indicating the amount of ‘blockiness’. A similar assumption is used in most existing instrumental quality measures. This second assumption still needs further experimental evidence to substantiate it. It should be realized, however, that it is an independent and separate assumption from the first computational assumption. Existing problems with the currently proposed blockiness measure (as well as with many existing measures) are in our view mostly related to this second integration step, and the computational assumption underlying it.

The image analysis method that we use to implement our blockiness measure is based on the Hermite transform, a signal decomposition technique in which signals are locally approximated by polynomials within a Gaussian window (Martens, 1990a). The approach was also applied by Kayargadde and Martens (1996b,a) for noise and blur estimation. The three stages of the proposed blockiness model are described below (see figure 4.5).

4.3.1 Front end processing

The display characteristic and the luminance adaptation of the visual system are modeled in the front-end processing. First, the gray-value image is transformed into a luminance image L , using the gray-value-to-luminance characteristic of the monitor see equation(4.1),

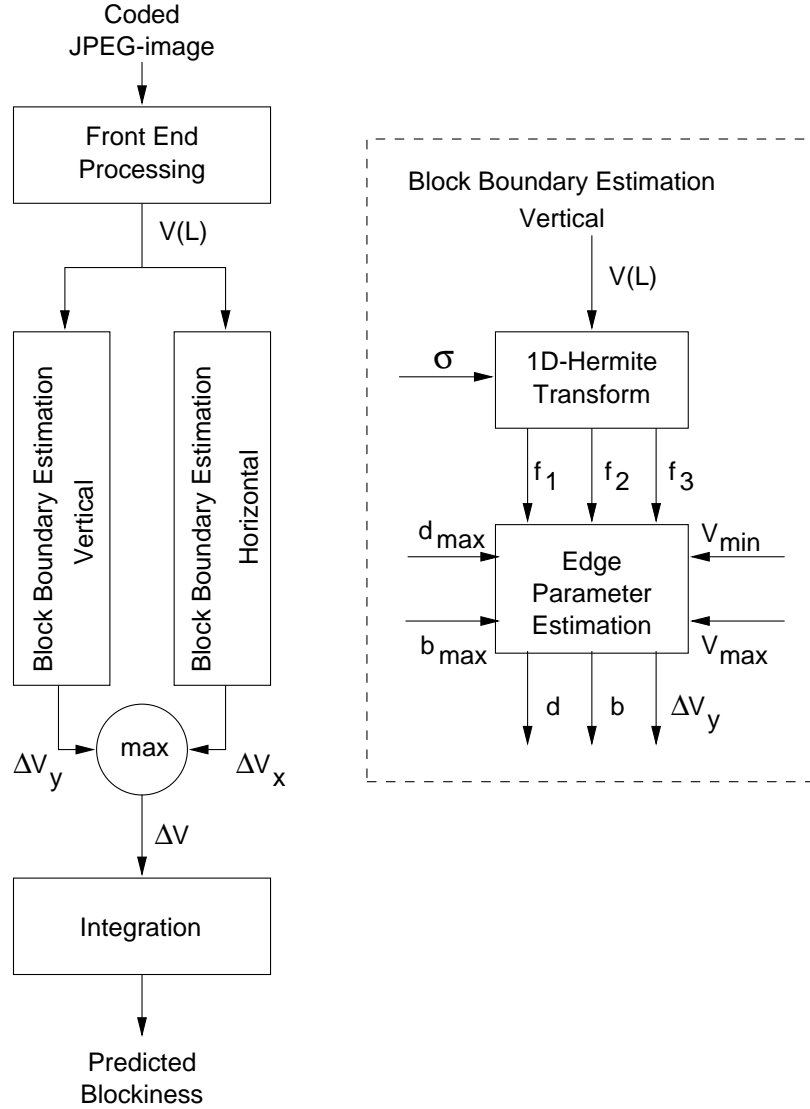


Figure 4.5: The three stages of the blockiness estimation model: (1) Front End Processing: transformation of the grey-scale image into a threshold step-edge visibility image, $V(L)$. (2) Block Boundary Estimation: estimation of the edge parameters, performed in horizontal and vertical direction separately. The Block Boundary Estimation stage is presented in more detail on the right. (3) Integration: collapsing the estimated amplitudes into a single scalar value.

and then into a step-edge-visibility image $V(L)$. In this image, the just perceivable increments or decrements, ΔV , needed to detect a step-edge are constant and independent of V . This means that $(V, V + \Delta V)$ are equally significant for all base values V , which simplifies the processing in the blockiness model. This does not hold for the luminance image in which the just perceivable differences, ΔL , needed to detect step edges are dependent on the base luminance L . Such a relation was experimentally determined in Lubin (1993) and can be expressed as:

$$\Delta L = aL^b \quad (4.2)$$

where $a = 0.01667$ and $b = 0.8502$, if the luminances L and ΔL are expressed in cd/m^2 . More specifically, this increment corresponds to a probability of 84% to detect a vertical step-edge $(L, L + \Delta L)$. Since the human sensitivity is the same for horizontal and vertical edges, this expression can be used for edges in both directions (Karunasekera and Kingsbury, 1995).

Inherent to equation(4.2), the luminance variation needed to detect step edges grows in proportion to the base luminance. In the visibility image $V(L)$, this detection threshold should correspond to a fixed increment $\Delta V = 1$, the just noticeable difference (jnd). Provided V is a differentiable function of the luminance L , ΔV can be expressed approximately as $[dV(L)/dL] \cdot \Delta L$. Substituting $\Delta L = aL^b$ and $\Delta V = 1$ gives $dV(L) = [1/aL^b] \cdot dL$. Finally, integration

$$V(L) = \int \frac{1}{aL^b} dL = \frac{1}{a \cdot (1-b)} \cdot L^{1-b} \quad (4.3)$$

gives the relationship between the luminance image L and the step-edge-visibility image $V(L)$.

4.3.2 Block boundary estimation

The detection and analysis of edges is based on the Gaussian blurred edge model as proposed by Kayargadde and Martens (1994, 1996a). The original model was used to estimate parameters of two-dimensional edges, and hence required the use of a two-dimensional Hermite transform. In the case of blockiness, we are interested in one-dimensional (horizontal and vertical) edges. We will therefore base our analysis on two separate one-dimensional Hermite transforms (Martens, 1990a,b) along the rows and columns of the image, respectively.

Below we describe the edge estimation in horizontal direction giving the following model for a blurred edge

$$\text{edge}(x; V_e, \Delta V, d, b) = V_e + \frac{\Delta V}{2} \operatorname{erf} \left[\frac{x-d}{b} \right], \quad (4.4)$$

with parameters: mean signal value (V_e), step-amplitude (ΔV), blur parameter (b), and distance from the origin (d). The edge parameters are estimated from Hermite coefficients up to order three. These coefficients are obtained by linear filtering of the step-edge-visibility image $V(L)$. The origin in the above notation is the (moving) center position of the Gaussian window, that is used in the Hermite analysis. The standard deviation of the Gaussian window was set to $\sigma = 2$ times the pixel distance.

4.3. Blockiness model

The Hermite coefficients for a blurred edge (Kayargadde and Martens, 1994) are given by

$$\begin{aligned} f_0 &= V_e - \frac{\Delta V}{2} \text{erf}(D) \\ f_1 &= \frac{R_b}{\sqrt{2}} a_0 \\ f_2 &= \frac{R_b^2}{2\sqrt{2}} a_0 2D \\ f_3 &= \frac{R_b^3}{4\sqrt{3}} a_0 (4D^2 - 2) \end{aligned} \quad (4.5)$$

which are defined in terms of the following intermediate variables

$$\begin{aligned} R_b &= \frac{\sigma}{\sqrt{b^2 + \sigma^2}} \\ D &= \frac{d}{\sqrt{b^2 + \sigma^2}} \\ a_0 &= \frac{\Delta V}{\sqrt{\pi}} \exp(-D^2) \end{aligned} \quad (4.6)$$

The estimation process consists of estimating the (unknown) edge parameters from the (given) Hermite coefficients. In this estimation process two things can happen. First, no solution for the edge parameters can be obtained, indicating that the current window is not positioned near an edge (and the Hermite coefficients do hence not correlate well with the Hermite coefficients for an edge). Second, the estimated edge parameters are not accepted as belonging to an edge that can be regarded as a block boundary.

Only positions for which f_1 is nonzero are considered. For these positions, we start by estimating the intermediate (blur) variable

$$R_b^2 = 2 \frac{f_2^2}{f_1^2} - \sqrt{6} \frac{f_3}{f_1}, \quad (4.7)$$

which in turn can be used to find the original blur parameter

$$b = \sigma \cdot \sqrt{\frac{1}{R_b^2} - 1}, \quad (4.8)$$

provided that $R_b^2 \leq 1$. An effect of DCT-based block coding, like in JPEG, is that ‘real image edges’ are blurred. To avoid detecting false edges, the allowed blur parameter for a block boundary edge is limited to $b \leq b_{max}$, or equivalently

$$R_b^2 \geq \frac{\sigma^2}{b_{max}^2 + \sigma^2}. \quad (4.9)$$

Through this restriction blurred edges are rejected.

The intermediate distance parameter is estimated as

$$D = \frac{f_2}{R_b f_1}, \quad (4.10)$$

which in turn can be used to derive the original distance parameter d . We also require that $d \leq d_{max}$, since an analysis some distance away from the edge will most likely be unreliable. Values of d larger than 0.5 times the pixel distance indicate that the edge position is closer to the previous or next pixel.

Without any restriction on the amplitude of the detected edges, all horizontal and vertical edges in an image are marked as block boundaries. Newly introduced edges as well as edges present in the original image are included. However, block boundaries occur at positions where the original image is mostly slowly varying and are therefore typically of small amplitude. Only edges with amplitudes $\Delta V \leq V_{max}$ will be identified as edges due to blockiness. Setting such a threshold may not exclude all ‘real image edges’ but reduces the number considerably. On the other hand, edges with very small amplitudes are most likely invisible, so that we also require $\Delta V \geq V_{min}$ in order to accept an edge as block boundary. The required edge amplitude is estimated by

$$\Delta V = f_1 \frac{\sqrt{2\pi} \exp(D^2)}{R_b}. \quad (4.11)$$

The zero-order Hermite coefficient f_0 can be used to derive the background value V_e . This background is for instance required in case we want to base the visibility of the block boundary on a contrast measure such as

$$c = \frac{\Delta V}{V_e}, \quad (4.12)$$

instead of on the amplitude. Such a contrast measure can serve as an alternative for amplitude in case the block boundary estimation is performed on the luminance image.

Summarizing, the thresholds b_{max} , d_{max} , V_{min} and V_{max} , together with the estimated edge parameters b , d and ΔV control the process of rejecting or accepting a detected edge as being due to blockiness. Typical values for the threshold parameters are $b_{max} = d_{max} = 0.5$ pixels, $V_{min} = 1$ and $V_{max} = 20$ on a step-edge visibility scale. Since the unit of the amplitude scale is 1 jnd, setting the minimum perceived amplitude step to $V_{min} = 1$ jnd seems obvious. The other parameter choices will be discussed further in section 4.4.1.

Block boundaries are estimated independently in the horizontal and vertical direction. At each pixel location, the edge estimates of one of the two directions will be selected and used in the integration stage. Whether these are the estimates in horizontal or vertical direction is determined by their edge amplitude ΔV . To make a distinction between the amplitudes in both directions they are referred to as ΔV_x and ΔV_y , for the horizontal and vertical amplitudes, respectively. If either of the conditions $\alpha|\Delta V_x| > |\Delta V_y|$ or $\alpha|\Delta V_y| > |\Delta V_x|$, with $\alpha = 0.1$, is satisfied, then $\Delta V = \max(\Delta V_x, \Delta V_y)$ is retained as the edge amplitude. By this selection rule, only edges in the horizontal or vertical direction are accepted as block boundaries, and oblique edges are rejected.

False detections occur in any estimation process. The following (heuristic) way of pruning the detected edge positions has therefore been included as a post-processing operation. Block boundaries typically extend over several pixels in the horizontal or vertical direction

4.3. Blockiness model

(8 pixels is a typical block size for JPEG-coded images). Therefore, all detected edge positions that do not belong to a horizontal or vertical edge segment of a minimum length are eliminated. A typical value is 4 pixels.

Information about the coder can also be used to optimize the edge parameter estimation. For instance, block boundaries occur on a regular 8x8 grid if 8x8 DCT-block coding is used. Estimating edge parameters only at those positions would reduce both the number of false detections and the number of total operations required. Such a priori information has however not been used here.

4.3.3 Integration

In the previous section we showed how edges in an image can be estimated by using Hermite coefficients and we described the selection procedure to make a distinction between block boundaries and ‘real image edges’. The obtained outputs are maps of estimated edge parameters: positions, amplitudes and blur. Although these maps visualize blockiness at a glance, the evoked sensation of human observers in relation to blockiness is not expressed explicitly. The maps need therefore to be interpreted in terms of ‘the perceived degree of blockiness’, as the model intends to predict. To meet this requirement the edge estimates have to be collapsed into a single scalar value, indicating the perceived degree of blockiness.

As discussed in section 4.2, subjective blockiness responses can be recorded, using numerical category scaling. Human observers are hence capable to detect block boundaries as well as to express the overall perceived blockiness. The integration process that this implies is, however, not obvious. Detected block boundaries can differ in shape, size, visibility and number. The shape varies from squares and rectangles to more complex structures that are composed of horizontal and vertical edges of different sizes. Besides this diversity in shape, size (large or small), amount (many or a few) and visibility (conspicuous or just noticeable) of the block boundaries all play a role in the perceived degree of blockiness.

The variable appearance of block boundaries is determined by the DCT-coding algorithm (the quantization step and block size), as well as by the scene content. For instance, images with large uniform regions suffer more from perceivable blockiness than highly detailed images. Blockiness is also a local distortion. Within a scene, blockiness is mainly visible in particular (uniform) areas and hence not equally distributed across the entire scene. The viewer’s attention is drawn to these regions and blockiness is experienced as unnatural and disturbing. Keeping this in mind, it is understandable that two images with different blocking appearances can summon the same response of blockiness.

Although we are aware of the complicated processes that play a role in arriving at a subjective blockiness response, we have used a simple straightforward integration procedure in our model. The assumption underlying the integration step is that human observers integrate the blockiness amplitude map, resulting from the estimation stage, into one blockiness response. Large amplitudes are expected to contribute more to the overall blockiness impression than low amplitudes. In section 4.4.2 different amplitude statistics (see Table 4.3) are evaluated: 1. the most frequently occurring amplitude (peak), 2. the average

of all estimated block amplitudes (avg), 3. the variation around this average (stdev), 4. the number of estimated block boundaries (nrp), 5. the amplitude at a particular percentage of estimated block boundaries (level), and 6. a weighted summation taken over all estimated amplitudes (Minkowski). Although the blockiness measures that we present here are based on the estimated amplitudes only, it may be useful to also incorporate other edge estimates, such as d and b , in such measures in a later stage.

Table 4.3: Combination rules used in the integration stage of the blockiness model to collapse the estimated edge amplitudes into a single scalar value representing the predicted blockiness.

Amplitude integration rules	
peak	most frequently occurring amplitude ΔV_i
average (avg)	$\frac{1}{N} \sum_i^N \Delta V_i = \overline{\Delta V}$
standard deviation (stdev)	$\sqrt{\frac{\sum_i^N (\Delta V_i - \overline{\Delta V})^2}{N-1}}$
number of detected points (nrp)	N
Level	Amplitude, ΔV_i , at a particular percentage in the cumulative histogram, i.e. at 50% this is the median
Minkowski summation	$\left(\sum_i^N \Delta V_i ^p \right)^{\frac{1}{p}}$
A-priori known blockiness indicator (extracted from JPEG coding)	
quantization step (q-step)	$q = \begin{cases} \frac{5000}{Q} & Q \text{ parameter} < 50 \\ 200 - 2 * Q & Q \text{ parameter} \geq 50 \end{cases}$

4.4 Model evaluation

The JPEG-coded images, described in the experimental set-up in section 4.2.2, were fed into the single-ended blockiness model of section 4.3. The performance of the model is evaluated on the basis of the estimated edge parameter maps and the comparison of the overall measure with the subjective blockiness data. The control parameters: maximum distance to the origin of a block boundary edge (d_{max}), the restricted amplitude range (V_{max} and V_{min}) and the maximum blur of an edge (b_{max}) are discussed in section 4.4.1. It will be shown that by an adequate choice for these control parameters block boundaries can be distinguished from ‘real image edges’. In section 4.4.2 the model is evaluated by means of the Pearson correlation of the predicted blockiness and the subjective blockiness data of section 4.2.2 for alternative integration rules.

4.4.1 Block boundary estimation on natural images

Initially, four different front ends were evaluated: 1. the original gray scales, 2. gray values converted to luminance values, 3. a lightness image produced from the luminance image according to a CIE standard (Martens and Meesters, 1998), and 4. the gray scales transformed by a threshold step-edge-visibility relation, see equation 4.3. Although the estimated edge amplitudes differ in all cases, the detected block boundary positions were almost identical. The visibility of an edge is only taken into consideration in the integration stage. Therefore, the results presented in this section are based on the threshold step-edge-visibility front-end processing.

As mentioned in section 4.3.2, information about the DCT coder can be used to optimize the block boundary estimation. In view of the fact that block boundaries appear on an 8x8 pixel grid in JPEG-coded images, the processing can be optimized by estimating the edge parameters only at those positions. Although this would reduce the falsely detected edges considerably this information is not used in the results presented below. Some a priori information about the block size is incorporated in the post-processing that eliminates the estimated block boundary edges with a length smaller than 4 pixels. We now discuss the remaining control parameters in the block boundary estimation. The distance parameter d_{max} is set to 0.5 in all cases. This implies that all detected edges are located within an interval, centered on the current window position, of one pixel wide. The significant effects of the blur parameter b_{max} as well as the amplitude restrictions, V_{min} and V_{max} , on the edge estimation will be illustrated below.

Figure 4.6 shows edge estimates resulting from the original *boat* scene and the JPEG-coded image, $Q = 20$. None of the control parameters V_{min} , V_{max} or b_{max} were set in 4.6(a) and 4.6(c). In both images, 'real image edges' were detected. Especially in the original no edges should be detected since blockiness is absent. The estimated block boundaries in the JPEG-coded image are located in the uniform areas (canvas and sky) as well as in the textured regions (water). Compare the block boundaries to the original scene presented in figure 4.1. Figures 4.6(b) and (d) show the edge estimates obtained by setting the blur parameter to $b_{max} = 0.5$. The detected number of edges in the original decreases drastically. The amount of 'real image edges' in the coded image also diminishes. The block boundaries are mainly detected in the uniform regions of the image (canvas and sky).

The corresponding cumulative histograms of the estimated amplitudes are given in figure 4.7. This figure includes all coded JPEG-versions and the original of the scene *boat*. Figure 4.7(a) represents the estimated amplitudes without any control parameter set and 4.7(b) shows similar results with blur parameter $b_{max} = 0.5$. The former, contains a large amount of high amplitudes whereas the latter has mainly low amplitudes. This indicates that block boundaries are low-amplitude step edges. This is in line with the assumed characteristics, described in section 4.3.

The effect of the amplitude control parameters $V_{min} = 1$ and $V_{max} = 20$, alone and in combination with $b_{max} = 0.5$, is shown in figure 4.8. Although both results show detected block boundaries of low amplitude, the edge boundaries in figure 4.8(a) represent both blurred and sharp transitions whereas in figure 4.8(b) only sharp edges were detected. Compared to figure 4.6 the parameters V_{min} and V_{max} reduce the number of false block boundaries

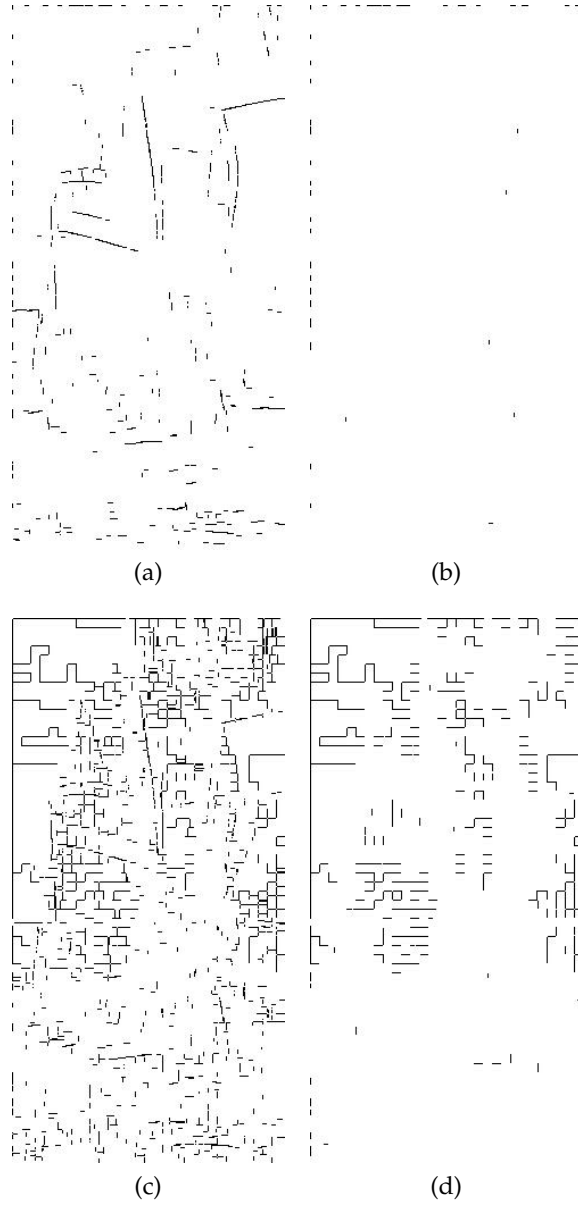


Figure 4.6: Block boundary estimates from the original *boat* scene, (a) and (b), and the same scene coded with Q -parameter set to 20, (c) and (d). The estimates in (a) and (c) were obtained without any of the control parameters V_{min} , V_{max} or b_{max} set. The results in (b) and (d) are obtained with b_{max} set to 0.5.

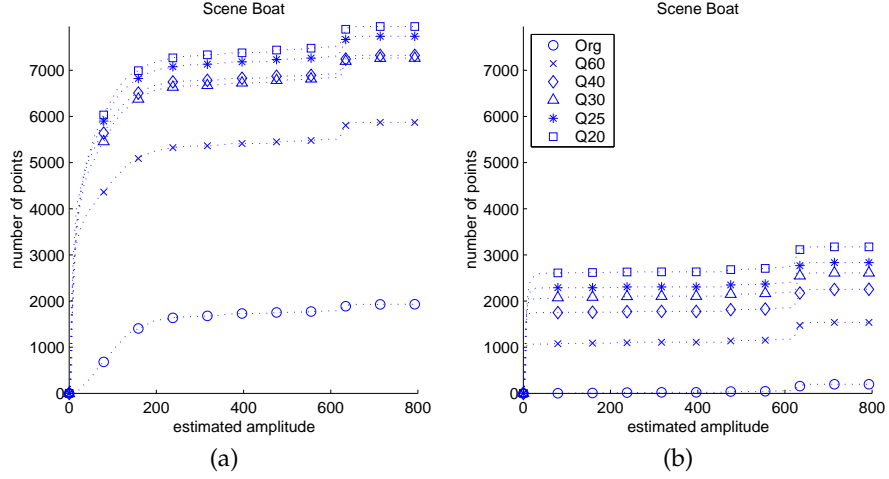


Figure 4.7: The estimated amplitudes corresponding to the edge positions as shown in figure 4.6 are presented as cumulative histograms. The amplitudes are shown on the x-axis and the number of detected amplitudes on the y-axis. All histograms are derived from the scene *boat*: (a) estimated amplitudes without control parameters, (b) with the blur parameter b_{max} set to 0.5,

considerably if b_{max} is not used. However, in combination with the blur parameter b_{max} , they have a very limited effect.

As expected, the amount of detected block boundaries decreases if the compression ratio decreases (higher Q -values). Block boundaries are most easily distinguished from ‘real image edges’ by making a distinction between blurred and non-blurred edges. Therefore b_{max} is the most relevant threshold parameter in the algorithm. Similar results as the ones presented here hold for the other scenes *child*, *girls* and *lighthouse*.

4.4.2 Integration of the estimated block boundaries

Although several block boundary parameters can be estimated, such as position, blur and amplitude, the integration of these estimated parameters into a single scalar value is currently confined to the amplitude estimates. The combination rules as listed in section 4.3.3 and Table 4.3 are evaluated in this section. Figure 4.9 shows that the versions ($Q20$, $Q25$, $Q30$, $Q40$, $Q60$, and the original) of a particular scene can be distinguished by the cumulative histogram of the amplitude. We also investigated if the JPEG Q -parameter, used to regulate the image quality, is a suitable measure to predict blockiness. The a-priori known quantization step q , which is uniquely determined by Q , will therefore also be compared to the subjective blockiness data.

As stated in section 4.3.3, we try to model the blockiness response of observers in this in-

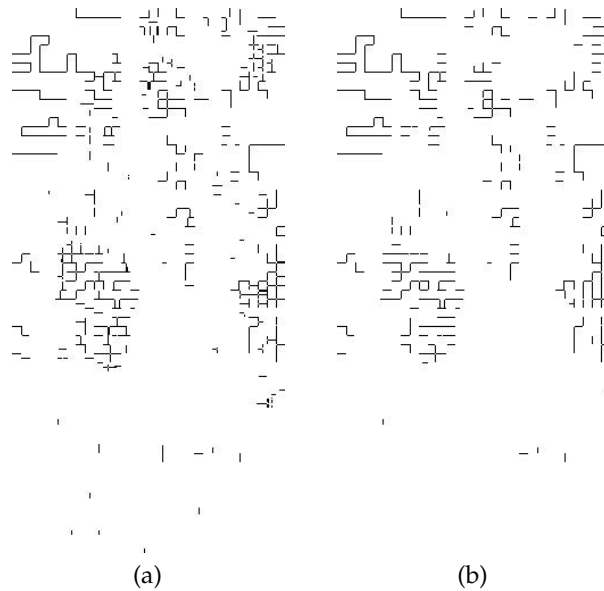


Figure 4.8: In both (a) and (b) the block boundaries of scene *boat* coded with $Q = 20$ were detected with $V_{min} = 1$ and $V_{max} = 20$. No blur parameter b_{max} was set in (a), whereas in (b) $b_{max} = 0.5$.

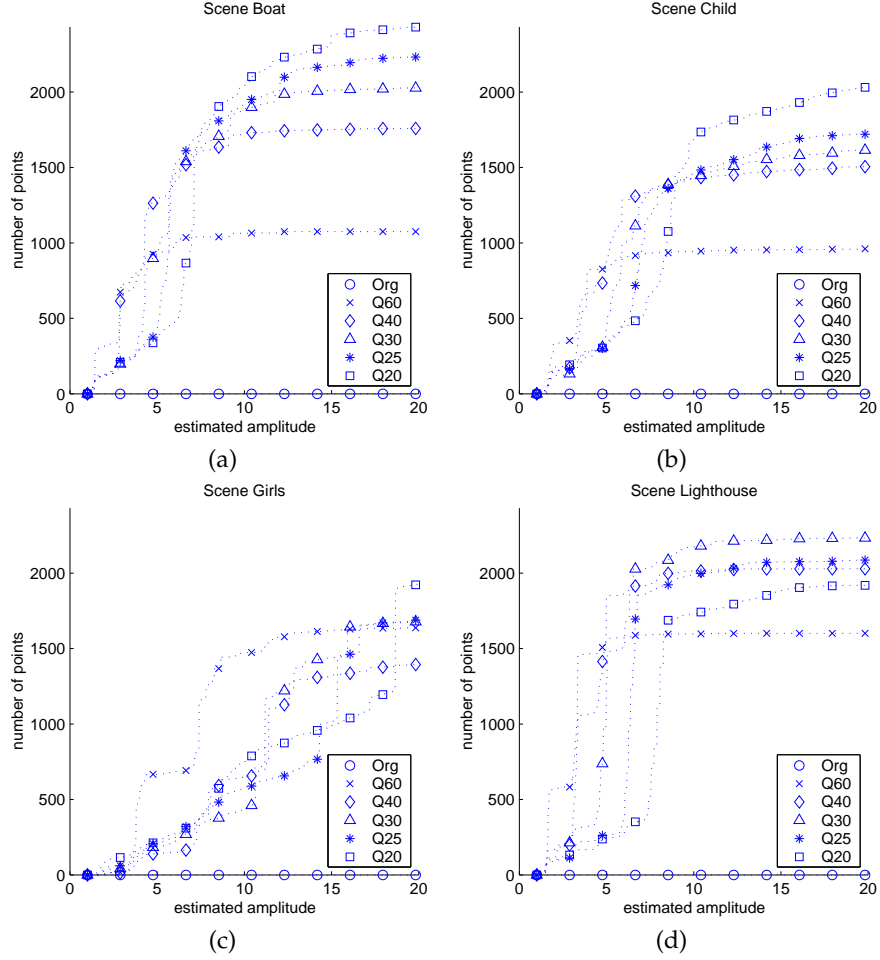


Figure 4.9: Cumulative histograms of all coded versions of the scenes (a) *boat*, (b) *child*, (c) *girls* and (d) *lighthouse*. The histograms show the estimated amplitudes between 1 and 20. The block boundary estimation is performed with $V_{min} = 1$, $V_{max} = 20$ and the blur parameter b_{max} set to 0.5.

tegration step. Although the mathematical expressions are easy to construct, one should be aware that different assumptions underly the various combination rules. By comparing different integration rules, we aim to get a better understanding of the combination rule used by the observers. Predictions that correlate poorly with the subjective data are considered to indicate an inadequate integration rule. However, high correlation between predictions and subjective data does not automatically imply that the corresponding combination rule reflects the human mechanism. More experimental evidence is necessary to substantiate this.

We first show that a reliable selection of block boundaries is essential before integrating the amplitudes. The instrumental measures perform poorly if real image edges are erroneously interpreted as block boundaries, irrespective of the applied combination rule. This is demonstrated in figure 4.10 showing the Pearson correlation (y-axis) of the subjective and the predicted blockiness for all combination rules (x-axis). The Pearson correlation was calculated between the predicted blockiness for all coded versions and the averaged subjective blockiness data, as presented in section 4.2.2.

Figure 4.10(a) depicts the case when no parameters are set in the block boundary estimation stage. This corresponds to the block boundaries of for instance the original and JPEG-coded version of the scene *boat* with $Q = 20$ as shown in figure 4.6(a) and 4.6(c), respectively. Real image edges are falsely identified as block boundaries and wrongly taken into consideration in the integration step. Consequently, the subjective blockiness data correlate poorly with most predicted blockiness. Figure 4.10(c) shows the results when only $V_{min} = 1$ and $V_{max} = 20$ are set in the block boundary estimation stage. The corresponding block boundaries of e.g. the JPEG-coded image with $Q = 20$ of the scene *boat* is shown in figure 4.8(a). In this case, only low amplitude edges contribute. Most blockiness predictions based on both sharp and blurred edges, as in figure 4.10(c), correlate poorly with the subjective blockiness data. Multiple factors may contribute to this. Blurred edges are less perceivable than sharp edges and should therefore be weighted accordingly in the overall predicted blockiness. They are also more likely to be image edges, another reason not to include them in the measure. In the previous section it was shown that V_{min} and V_{max} , in combination with the blur parameter b_{max} , have limited effect on the estimated positions of the block boundaries. However, the effect of the high amplitudes on the blockiness prediction is clearly seen by comparing figure 4.10(b) and figure 4.11. If image edges are erroneously interpreted as blockiness edges most blockiness summation rules give a poor performance, while a weighted Minkowski summation of the detected edge amplitudes performs well. In all cases an exponent p can be determined such that the predicted blockiness correlates highly with the perceived blockiness.

Next we discuss the results, with the estimation parameters fixed to $V_{min} = 1$, $V_{max} = 20$ and $b_{max} = 0.5$. The Pearson correlation coefficients, r , between the perceived blockiness and predicted blockiness for alternative integration rules are shown in figure 4.11. The integration rules are given on the x-axis and the Pearson correlation coefficients on the y-axis. It can be observed that the blockiness prediction without incorporating the visibility of the edges performs relatively poorly (see *nrp* in figure 4.11). In this case the correlation coefficient varies most across scenes. On the other hand, blockiness predictions using estimated edge amplitudes correlate highly with the perceived blockiness.

4.4. Model evaluation

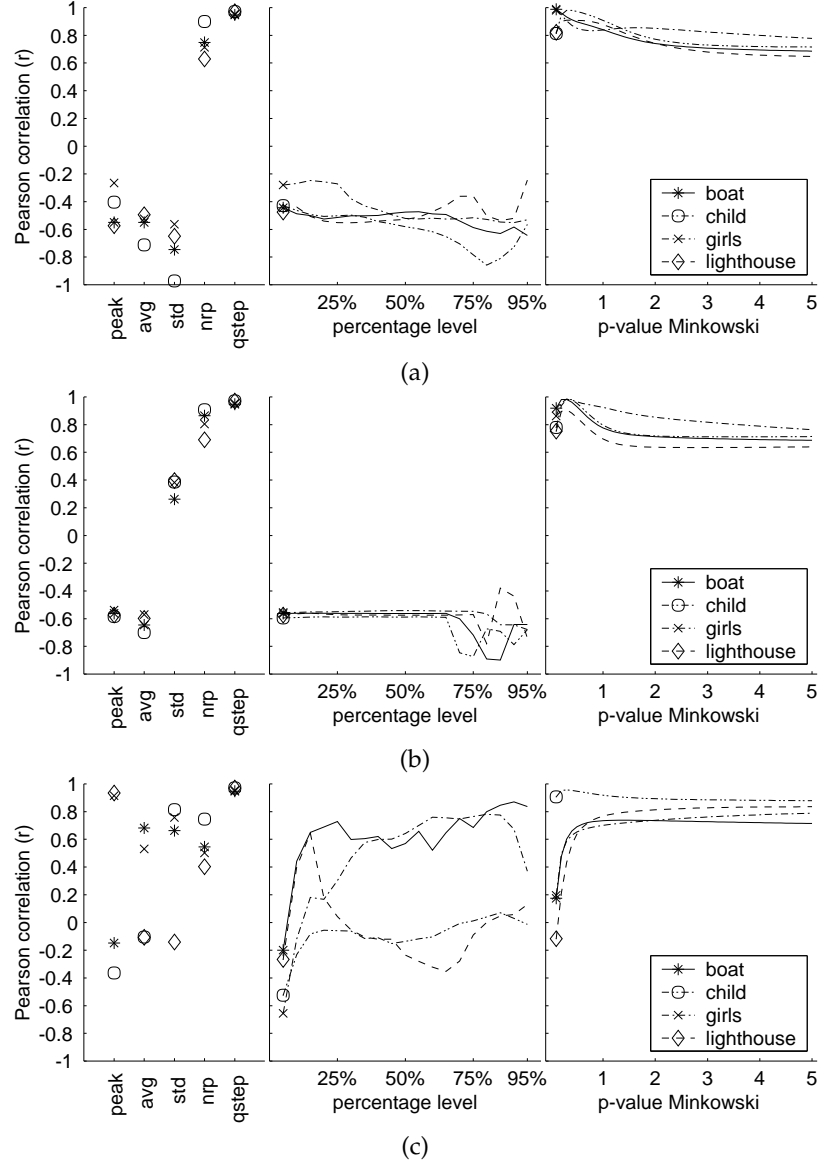


Figure 4.10: Pearson correlation between subjective blockiness data and the predicted blockiness of the four scenes is shown along the y-axis. The various blockiness integration rules are displayed along the x-axis. The parameters in the block boundary estimation were (a) none, (b) solely the blur parameter $b_{max} = 0.5$, and (c) only the amplitude parameters $V_{min} = 1$ and $V_{max} = 20$.

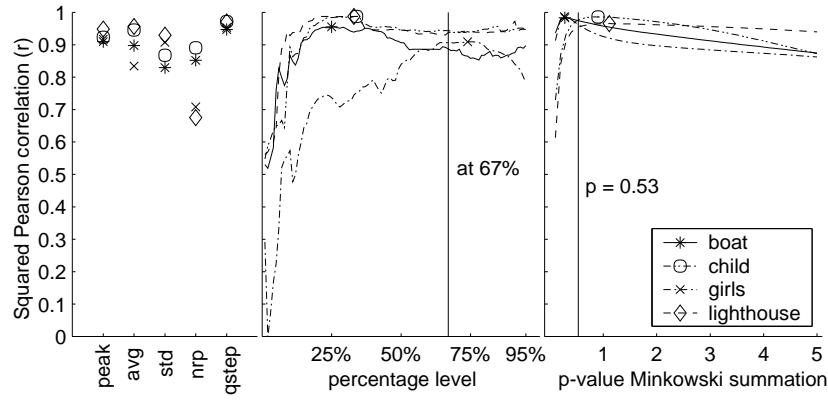


Figure 4.11: The Pearson correlation coefficients, r , obtained between the subjective blockiness judgements and the predicted blockiness for each scene separately. The various blockiness indicators are shown on the x-axis and the Pearson correlation coefficients on the y-axis. The parameters in the block boundary estimation were: $V_{min} = 1$, $V_{max} = 20$ and $b_{max} = 0.5$. The middle panel shows the resulting correlations of an amplitude taken at various levels of the cumulative histogram. The right panel shows the resulting correlations of a weighted Minkowski summation for various values of the exponent p . The symbols indicate the best correlation for each scene. The highest correlation, averaged across scenes, is obtained at a level of 67% in the cumulative histogram and with an exponent of $p=0.53$ for the Minkowski summation.

4.4. Model evaluation

Table 4.4: For each scene the Pearson correlation, r , between the subjective blockiness data and the predicted blockiness. The correlation is obtained for the most optimal parameter per scene as well as for the most optimal parameter for all four scenes.

Summation rule	<i>boat</i>	<i>child</i>	<i>girls</i>	<i>lighthouse</i>
Level, at 25%	0.96			
Level, at 34%		0.98		
Level, at 74%			0.91	
Level, at 33%				0.99
Level, at 67%	0.89	0.94	0.91	0.94
Minkowski, $p=0.28$	0.98			
Minkowski, $p=0.90$		0.99		
Minkowski, $p=0.27$			0.99	
Minkowski, $p=1.11$				0.97
Minkowski, $p=0.53$	0.97,	0.98	0.96	0.95

The middle panel of figure 4.11 shows the resulting correlations of an amplitude taken at various levels of the cumulative histogram. As can be seen in this figure and Table 4.4, the best correlation for each scene is obtained at a different percentage level of the cumulative histogram. Nevertheless, for each scene the Pearson correlation coefficient obtained at its optimal percentage level in the cumulative histogram or at a level of 67% are not significantly different.

The right panel of figure 4.11 shows the resulting correlations of weighted Minkowski summation for various values of the exponent p . Also in this case, the best correlation for each scene is obtained for a different p value. It is approximately one for the scenes *child* and *lighthouse*, implying a linear addition of the estimated amplitudes. A non-linear addition is implied when the optimal p -exponent is smaller than one such as for the scenes *boat* and *girls*. However, for each scene the linear correlation coefficients obtained by its optimal p or $p = 0.53$ are not significantly different (see Table 4.4). Contrary to many existing quality metrics where a Minkowski summation with an exponent close to two is used (de Ridder, 1991, 1992), the optimum p value is much smaller here. It seems that the integration of edge amplitudes into a subjective blockiness response differs from the integration of impairment strengths into an overall quality judgement.

In conclusion, for the scenes *boat*, *child*, *girls*, and *lighthouse*, blockiness predictions based on the estimated edge amplitudes correlate highly with the perceived blockiness. For the integration rules Level and Minkowski a parameter value (67% and $p=0.53$, respectively) is determined for which the correlation averaged across scenes is a maximum.

Up to now the integration rules were judged on the basis of the correlation between the predicted blockiness and the perceived blockiness strengths. In the following we will describe an additional test to compare the blockiness predictors for a large number of scenes.

Considering the experiments in section 4.2 we may assume that the perceived blockiness increases monotonically with decreasing Q-parameter. This leads to the requirement that the predicted blockiness should monotonically increase with decreasing Q-parameter.

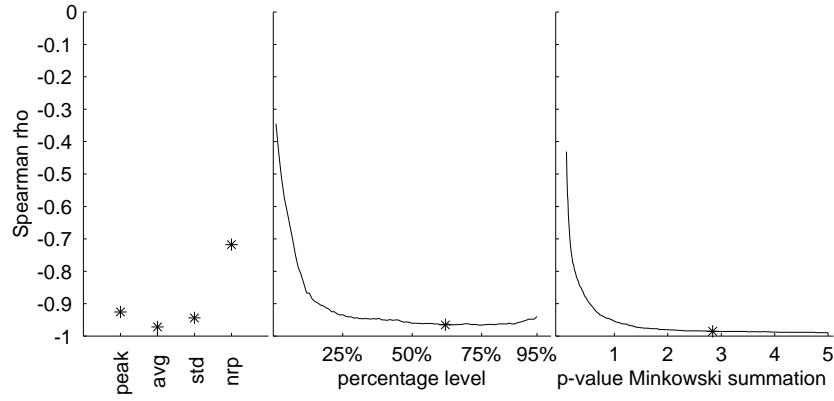


Figure 4.12: Spearman rho correlation coefficients, ρ , between the Q-parameter and the predicted blockiness. The various integration rules are shown on the x-axis. On the y-axis the Spearman rho coefficients are given. The middle panel shows the resulting Spearman rho coefficients of an amplitude taken at various levels of the cumulative histogram. The right panel shows the resulting Spearman rho coefficients of a weighted Minkowski summation for various values of the exponent p . The symbols in the middle and right panel indicate the best Spearman rho coefficient for the particular integration rule. The best Spearman rho coefficient, averaged across scenes, is obtained at a level of 62% in the cumulative histogram and with an exponent of $p=2.84$ for the Minkowski summation.

4.4. Model evaluation

Table 4.5: The Pearson correlation coefficient, r , between the subjective blockiness judgements and the predicted blockiness for each of the scenes *boat*, *child*, *girls* and *lighthouse*. The right column of the table gives the Spearman rho between the predicted blockiness and the Q-parameter if in total 164 scenes are considered.

Summation rule	Pearson r				Spearman rho
	<i>boat</i>	<i>child</i>	<i>girls</i>	<i>lighthouse</i>	164 scenes
Peak	0.91	0.92	0.92	0.95	-0.93
Avg	0.90	0.95	0.83	0.96	-0.97
Std	0.83	0.87	0.91	0.93	-0.94
Nrp	0.85	0.89	0.71	0.68	-0.72
Level, at 67%	0.89	0.94	0.91	0.94	-0.96
Minkowski, $p = 0.53$	0.97	0.98	0.96	0.95	-0.90
q-step	0.95	0.97	0.95	0.97	-1.00

Since no subjective data is needed to test this requirement, it can easily be performed on a large image set. For each scene it is tested whether a specific summation rule predicts a monotonic increase in blockiness with decreasing value of Q . The monotocity is expressed as a rank-order correlation coefficient, Spearman rho (ρ), between the Q -values and the predicted blockiness. A specific combination rule is characterized by the mean value of Spearman's rho averaged across 164 scenes (see Appendix A). For that purpose the 164 JPEG coded images of *Chapter 2* (Q -parameters: 15,20,25,30,40,60 and the original) are used. In figure 4.12 the integration rules are given on the x-axis and the Spearman rho coefficients on the y-axis. A Spearman rho coefficient of -1 indicates that the blockiness is predicted monotonically in all 164 scenes. It can be observed that the Spearman rho coefficient is close to this value for most integration rules. A significant deviation can be observed for the integration rule nrp.

The Spearman rho coefficients obtained for 164 scenes and the Pearson correlation coefficients for each of the scenes *boat*, *child*, *girls*, and *lighthouse* are given in Table 4.5.

Next, the parameter value in the integration rules Level and Minkowski are determined such that the Spearman rho coefficient between the predicted blockiness and the Q -parameter is optimal (see Table 4.6). In the former case this results in a ρ of -0.97 if the amplitude is taken at 62% of the cumulative histogram. In the case of a Minkowski summation, the exponent $p = 2.84$ gives the best results, ρ is -0.99 .

Figure 4.13 shows the monotocity for the 164 scenes in more detail. For each integration rule in Tables 4.5 and 4.6 the Spearman rho coefficient between the predicted blockiness and the JPEG Q -parameters of a particular scene is calculated. In this figure one can see that for no single measure all 164 scenes are predicted monotonically and that a weighted Minkowski summation with exponent $p = 2.84$ gives the best result. In this case a great number of scenes is predicted monotonically, namely 133 scenes out of 164. For the remaining scenes the deviation in monotonicity is small compared to those seen for the other integration rules.

In Tables 4.5 and 4.6 the Spearman rho coefficients indicate the degree to which the pre-

Table 4.6: The Pearson correlation coefficient, r , between the subjective blockiness judgements and the predicted blockiness for each of the scenes *boat*, *child*, *girls* and *lighthouse*. The right column of the table gives the Spearman rho between the predicted blockiness and the Q-parameter if in total 164 scenes are considered.

Summation rule	Pearson r				Spearman rho
	<i>boat</i>	<i>child</i>	<i>girls</i>	<i>lighthouse</i>	164 scenes
Level, at 62%	0.89	0.94	0.85	0.94	-0.97
Minkowski, $p = 2.84$	0.91	0.95	0.88	0.96	-0.99

dicted blockiness increases monotonically with decreasing Q-parameter. Also the Pearson correlation coefficient between the perceived and predicted blockiness for four scenes are reported. Considering only the correlation coefficients most integration rules perform similarly. Nevertheless, the ranking test revealed that not all integration rules predict JPEG versions of a scene monotonically. A weighted Minkowski summation with an exponent of 2.84 predicts the JPEG versions for most scenes monotonically and correlates highly with the perceived blockiness for four scenes. Therefore, this integration rule is proposed in the single-ended blockiness measure.

In figures 4.10, 4.11 and Table 4.5 also the correlation between the a priori known quantization step, q-step, and the subjective blockiness judgements is shown. The high correlation implies that the Q-parameter is suited to indicate the expected degree of blockiness in sequential baseline coded JPEG images.

4.5 Summary

In this chapter, the underlying attributes of the perceived image quality of sequential baseline coded JPEG images were studied. Moreover a single-ended blockiness measure was developed. We showed that blockiness (defined as horizontal and vertical low-amplitude step edges) can be predicted by analyzing the coded image only. Estimated edge parameters, b , d and ΔV , were derived from Hermite coefficients. These estimated edge parameters, together with the thresholds b_{max} , V_{min} and V_{max} , were used to differentiate between real image edges and edges introduced by the coding process. Typical values of the threshold parameters are $b = 0.5$, $V_{min} = 1$ and $V_{max} = 20$ on a step-edge visibility scale. With these values, block boundaries can be detected accurately. The degree of blockiness was solely derived from the estimated edge amplitudes ΔV . For that purpose several integration rules were studied. A weighted Minkowski summation of the estimated edge amplitudes with an exponent of 2.84 appeared to be the most optimal integration rule. Firstly, blockiness predictions derived with this summation rule correlate highly with the perceived blockiness strenghts. Secondly, by analyzing this summation rule for a large image set, consisting of 164 scenes, it could be shown that the predicted blockiness increased monotonically with decreasing JPEG Q-parameter for a maximum number of scenes. Further improvements of this single-ended blockiness measure might be possible by using the

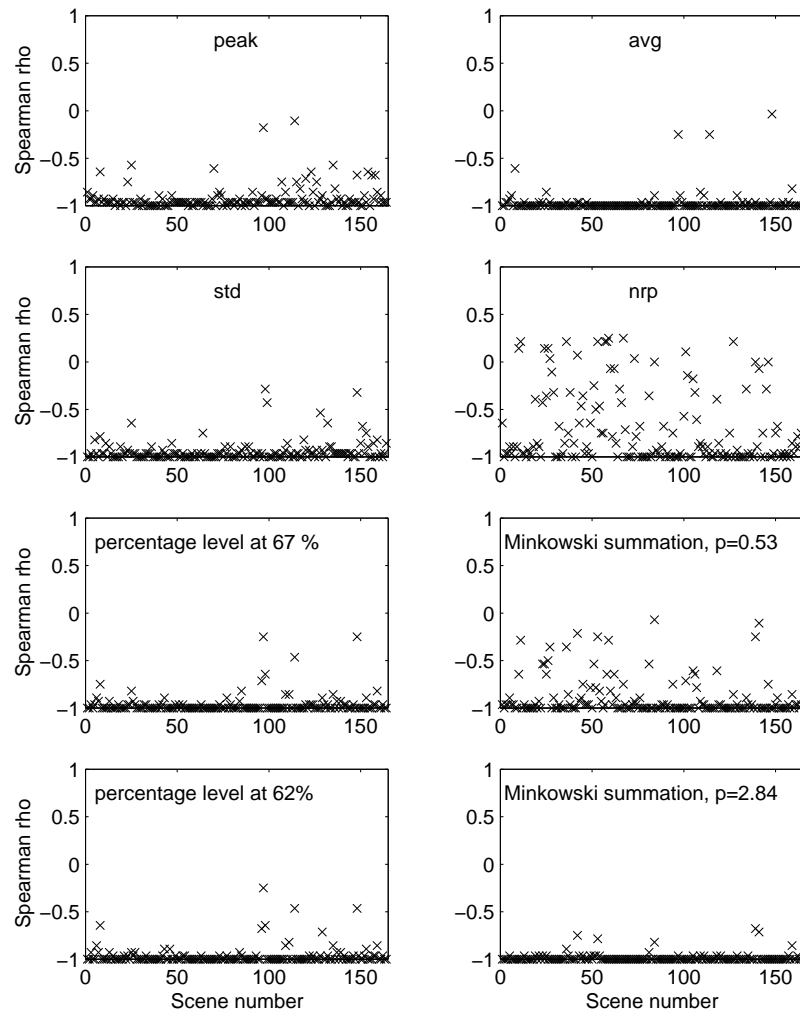


Figure 4.13: Per summation rule the Spearman rho is given for each of the 164 scenes.

estimated edge parameters d , b and c as additional blockiness information in combination with the estimated amplitudes.

We also showed that the measured perceptual strengths of the three attributes blockiness, ringing and blur correlate highly. From this result we can conclude that the image quality of sequential baseline coded JPEG images can be described by a single attribute. This conclusion is also supported by the experimental results of dissimilarity judgements. Therefore the proposed blockiness estimation algorithm can also be used as image quality measure for sequential baseline coded JPEG images.

Chapter 5

Evaluation of instrumental quality measures

Abstract

The predictive power of instrumental quality measures is usually analyzed by comparing the predictions with results from subjective testing. Due to the limited number of images that can be presented in subjective tests the outcome is hard to generalize. In this chapter a large image set is used to compare instrumental quality measures on the basis of their predictions only. In a second step of the analysis a small subset of images is selected which optimally discriminates between the predictions of the instrumental measures. This subset of images is then used in a subjective quality test. Furthermore the scenes were chosen to test whether the quality measures work on an absolute quality scale which would enable them to compare the image quality between different scenes. It is shown that the performance of instrumental quality measures for such a set of well-chosen images is different from the predictions obtained for scenes chosen on the basis of image content. Most measures cannot cope with between-scene comparisons.

5.1 Introduction

As already discussed in the previous chapters a large number of instrumental quality measures has been proposed. The usefulness of such measures is determined by their ability to predict perceived image quality. In order to test whether a specific instrumental quality measure performs well, or to show the difference in performance between different measures they are usually evaluated by means of subjective quality judgements. These quality judgements are usually obtained for an image set consisting of a small number of scenes with particular distortions. Due to the limited number of images used in subjective tests it is hard to generalize the outcome of such an evaluation. This describes a basic problem in comparing instrumental quality measures with subjective data. In terms of generality a large stimulus set has to be analyzed, while a test containing a large number of images is not realistic in subjective testing.

In *Chapter 2* instrumental quality measures were classified on the basis of their quality predictions. In contrast to subjective testing this makes it possible to use an image set with a large number of scenes and quality degradations since only computer resources are needed. However, in such an analysis no reference to the perceived quality is used and therefore no conclusions about the measures' performance can be obtained by such a cluster analysis. The clustering is merely a tool to get a better understanding of the differences between quality measures by using a large image set. To determine whether a measure is suitable to predict the perceived image quality subjective tests are still needed.

Traditionally, the selection criteria for scenes used in subjective tests are based on image features that are, for example, critical for the coding algorithm. In *Chapter 2* and the present chapter, we follow a different approach namely to select scenes for which the examined quality measures differ in their quality predictions. Two objective criteria are used: 1) each quality measure yields different results for the scenes and, 2) scenes yield different results for each of the quality measures. In the first case, scenes are chosen such that per instrumental measure the predicted quality of scenes is different. Scenes chosen on the basis of the second criterion are used to test whether the instrumental quality measures work on an absolute quality scale. For instance, images with different scene content but the same degree of distortion can be of different image quality. Thus instrumental measures should not only predict the image quality within a scene but also across scenes. This implies that the image quality should be measured on a single scale such that image quality predictions of different scenes are comparable.

In conclusion, differences between instrumental measures can be indicated for a large image set without the need of subjective data. Since it is not realistic to use large sets in subjective testing, only those scenes that actually discriminate within and among the measures are recommended to be used in the evaluation. In the present chapter, this is demonstrated by testing the instrumental quality measures introduced in *Chapter 2* for JPEG coded images. In section 5.2.1 the difference between the measures is analyzed by using only their predictions. The selection of scenes that discriminate within and between these measures is described in section 5.2.2.

In *Chapter 4* the conclusion was reached that the perceived image quality of JPEG coded

images is linearly related to its underlying attributes blockiness, blurring and ringing. This suggests that the perceived image quality of JPEG coded images can be modeled in one dimension. This indicates that the one-dimensional instrumental quality measures of *Chapter 2* could correlate highly with the perceived image quality. In order to test the performance of instrumental quality measures two experiments were conducted. The first experiment, described in section 5.3, was carried out to test whether a linear relationship between the perceived image quality and the attribute ringing holds for low-quality JPEG coded images. So far, this relationship has only been investigated for medium and high quality images. The second experiment is described in section 5.4. Here it was investigated whether the relation between perceived image quality and attribute strengths is also linear for across-scene comparisons. In *Chapter 4* it was shown for scenes separately that the perceived image quality and the attribute strengths are linearly related. However, a linear relationship leaves two undetermined variables, a scaling factor and an offset. If the linear relationship between the attributes is different for each scene this would indicate that the image quality can not be predicted in one dimension.

The subjective quality data of both experiments are used to evaluate instrumental quality measures. The most basic factor that determines the image quality is the level of impairment introduced into an image. In the case of JPEG coded images the perceived quality decreases with increasing strength of the attributes blockiness, blurring and ringing. Therefore it is a prerequisite that the measures can predict the level of image quality for a particular scene. This issue will be addressed in section 5.5.1. Another point that is considered is that measures should predict the perceived quality independent of the scene content. Image quality is a subjective qualification that observers can apply across scenes. Human viewers can point out which image they prefer in quality even though the scene content is different. This relation between scene content and the perceived image quality should therefore also be incorporated in an instrumental measure. Whether the measures can deal with this scene independency is described in section 5.5.2. In section 5.5.3 it is shown that the used scenes can make a large difference in the evaluation of the measures. In section 5.6 the single-ended blockiness model is also tested for across-scene blockiness predictions. Finally in section 5.7 the differences between double-ended instrumental quality measures, the single-ended blockiness measure and an average human observer are visualized by means of an MDS stimulus configuration.

5.2 Selection of quality measures and scenes

In *Chapter 2* it has been shown that we can study instrumental quality measures by categorizing these measures into different groups based on their predictability for a large set of images. The categorization was based on an image set formed by 164 scenes processed by 4 methods at 6 processing levels. In the present chapter the JPEG coded versions of these scenes were used to evaluate the instrumental quality measures. In section 5.2.1 the instrumental measures were classified for such a large image set. The predictions of quality measures within a cluster were, on average, similar for this image set. From each of the resulting clusters one quality measure was selected and used to compare the predictions with the perceived image quality.

Obviously, it is not feasible to use such a large image set in subjective testing. Therefore, the criteria as discussed in *Chapter 2* were used to select a reasonable number of images for the evaluation. The scenes used to evaluate the selected instrumental quality measures were chosen such that: (1) each quality measure yields different results for the scenes and (2) the scenes yield different results for the instrumental quality measures. This selection is described in section 5.2.2

5.2.1 Quality measure selection

The same 67 instrumental quality measures and a subset of the images discussed in *Chapter 2* were used. The subset of images contained the 164 scenes (see Appendix A), JPEG coded at 6 levels with Q-parameters of 15, 20, 25, 30, 40 and 60.

In the same way as in *Chapter 2* the 67 instrumental quality measures were classified on the basis of their predictions obtained for this JPEG image set. First the quality predictions were normalized per quality measure. Next, a measure of association between each of the 67 quality measures was obtained by the inner-product correlation resulting in a 67x67 dissimilarity matrix. From this dissimilarity matrix a 2-dimensional stimulus configuration was determined by the multi-dimensional scaling program *xgms*. Finally Ward's hierarchical clustering was used to cluster the quality measures according to their estimated Euclidean distances. The resulting hierarchical cluster tree is shown in figure 5.1.

Figure 5.2 shows the number of instrumental quality measure clusters versus the distance in the hierarchical cluster tree. This distance is proportional to the increase of the within-group error sum of squares. A substantial increase in distance can be observed if the number of clusters is reduced to four or less. This suggests that four main groups of quality measures can be distinguished. The quality measures within a group are considered to predict a similar image quality for the JPEG images.

From each cluster the quality measure with the maximum summed distance to all other cluster centroids was selected, for a more detailed description see *Chapter 2*. The four selected quality measures *gdcor*, *sper95*, *shub95*, and *sarnoff-s* of each cluster are given in figure 5.1.

5.2.2 Scene selection

The four selected quality measures were used to group the 164 scenes (see Appendix A). Scenes were grouped if the image quality was, on average, predicted similarly by the four instrumental quality measures. In the same way as in *Chapter 2*, the city block distance was used as a measure of proximity that indicates the relation from one scene to the other. The resulting 164x164 distance matrix was used in Ward's hierarchical cluster analysis to cluster the scenes. The resulting cluster tree is shown in figure 5.3. Figure 5.4 shows the number of scene clusters versus the distance in the hierarchical cluster tree. A substantial increase in the distance between the clusters occurs if the number of scene clusters is reduced to three or less. Hence, this figure suggests that 3 main groups of scenes can be distinguished.

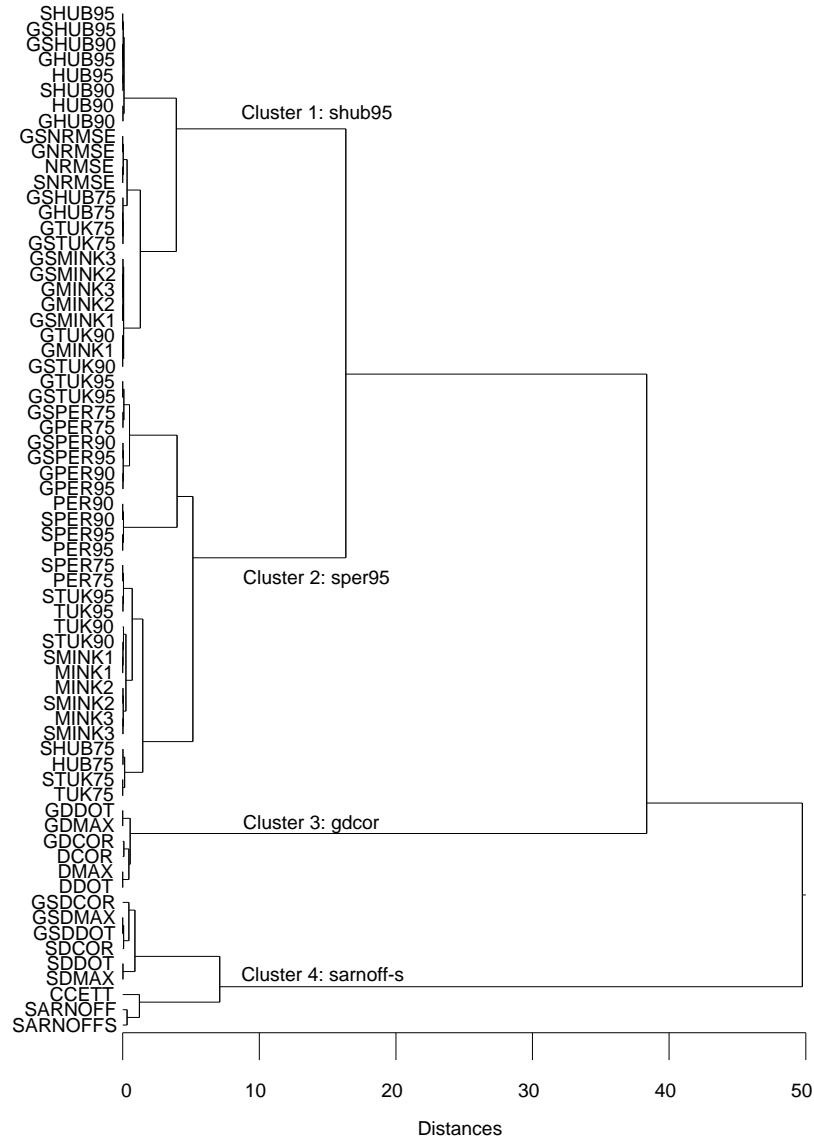


Figure 5.1: Hierarchical cluster tree obtained for 67 quality measures and 164 scenes. All 164 scenes were JPEG coded with six Q-parameters. Four main clusters of quality measures can be identified, and from each cluster one quality measure was selected. These selected measures were shub95, sper95, gdcor and sarnoff-s.

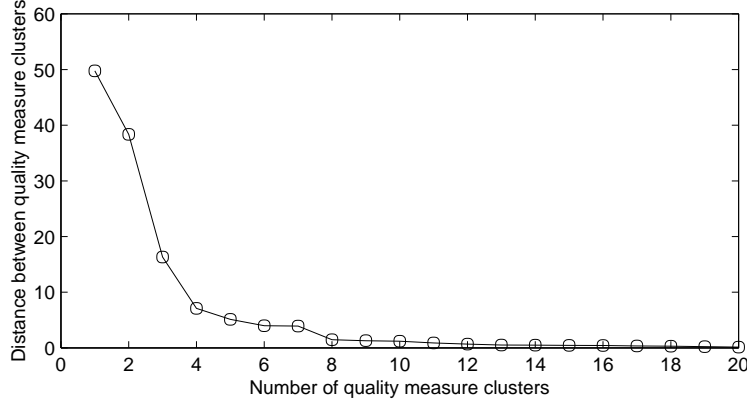


Figure 5.2: The number of instrumental quality measure clusters versus the distance in the hierarchical cluster tree. The distance represents the total within-group error sum of squares. A major increase in the distance occurs if the number of quality measure clusters is reduced to four or less.

In the same way as in *Chapter 2* from each cluster one scene was selected and used in the experiments of section 5.3 and 5.4. The properties of the selected three scenes were:

1. each quality measure yields different results for these three scenes
2. and the three scenes yield different results for each of the quality measures.

Both properties were satisfied by taking from each cluster a scene that discriminates highly between the predictions of the four quality measures. Hence for each scene the city-block distance was used as a measure of proximity that indicates the relationship between the predictions of the four quality measures. This resulted in a 4x4 distance matrix per scene. Finally, for each scene this 4x4 distance matrix was summed and from each cluster the scene with the maximum summed distance was chosen. The three scenes and their position in the hierarchical cluster tree are shown in figure 5.3. The selected scenes will be referred to as *roses*, *boat* and *museum*.

The selected scenes indeed express the difference within and between the quality measures as is shown in figure 5.5. This figure demonstrates for each selected quality measure the predictions for the JPEG coded versions of the three scenes. The predictions were scaled between 0 and 1. A value of zero indicates that the JPEG coded image is of similar quality as the original one. The worst image quality is represented by 1. The quality measures predict the image quality for the three scenes differently. For the measures shub95 and sper95 (top panels in figure 5.5) the scene *boat* has the lowest image quality compared to the scenes *museum* and *roses*. However, this in contrast to the measure sarnoff-s, bottom right, which predicts the image quality in reversed order. This measure does not differentiate in quality between the scenes *boat* and *museum*. The measure gdcor predicts the images quality again differently. The scene *museum* is predicted as the image with the lowest quality and

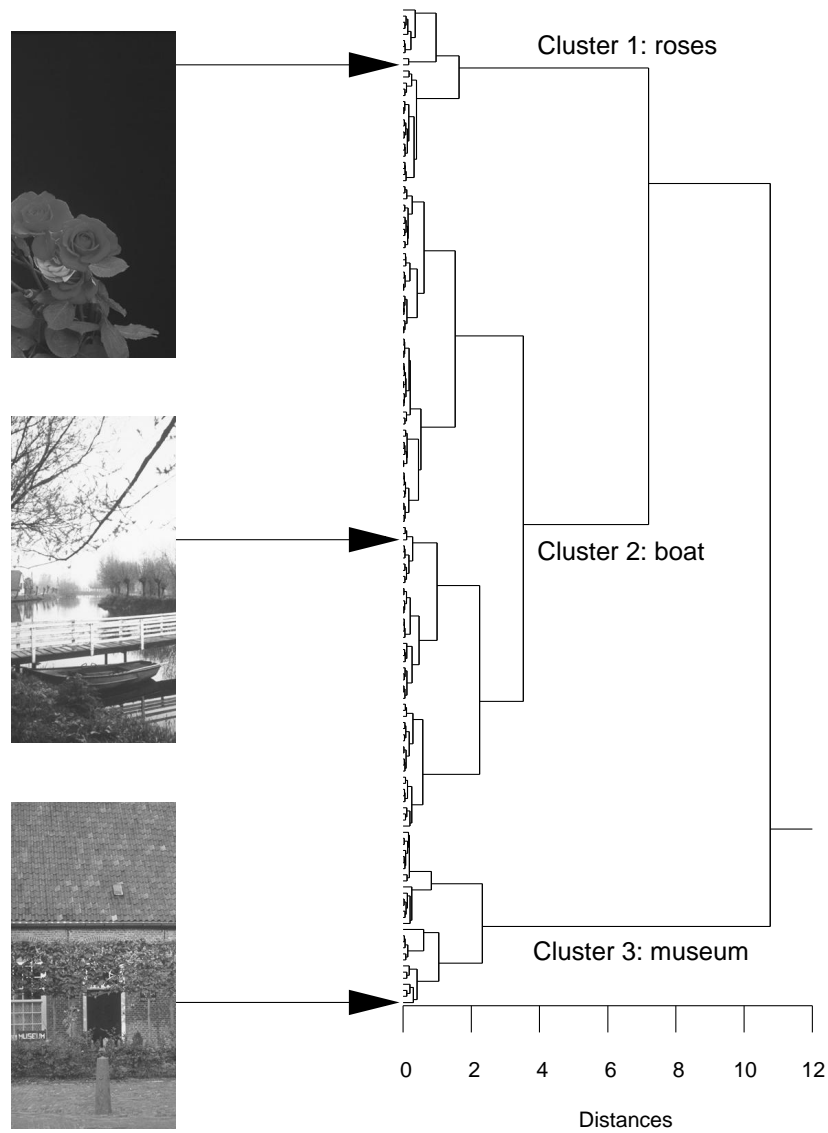


Figure 5.3: Hierarchical cluster tree of scenes obtained for 164 scenes. The cluster tree is derived from the predictions of four quality measure. Each measure was applied on six JPEG-coded versions of the 164 scenes. The three selected scenes and their positions in the tree are shown on the left side.

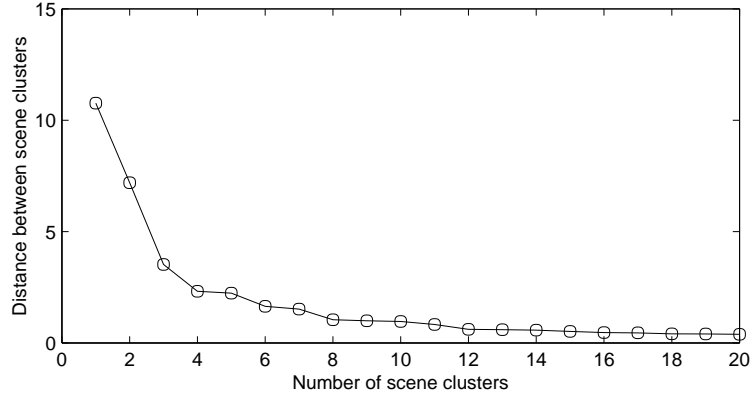


Figure 5.4: The number of scene clusters versus the distance in the hierarchical cluster tree. A major increase in the distance occurs if the number of scene clusters is reduced to three or less.

the scenes *roses* and *boat* are predicted to have a similar higher quality. Since not all measures agreed upon the predicted image quality, they cannot all predict the perceived image quality of JPEG coded images.

5.3 Experiment 1: attribute scaling within a scene

The distortions perceived in JPEG coded images are blockiness, ringing and blur. In *Chapter 4* it was shown that the strength of these distortions are linearly related and that the perceived image quality can be modeled by one of these attributes.

A different conclusion concerning the relation between the three distortions was reached by de Ridder and Willemsen (2000). He used numerical categorical scaling and percentage scaling to investigate the contribution of these three distortions to the overall image quality. This paper indicated that the relation between the attributes depends on the degree of impairment. For high quality images the attributes seem indeed linearly related. For low quality images the relation between the attribute strengths was different. In this case the perceived ringing strength seemed to saturate for low quality images while the blockiness and blur strengths still increased.

In *Chapter 4*, the observers judged the perceived strength of the attribute ringing only for medium to high quality images ($Q30 - Q100$). Therefore in this section an additional experiment was carried out to investigate if the attributes blockiness, blurring and ringing are also linearly related for low quality images. The following hypothesis was tested:

For low quality JPEG images the strength of ringing saturates and is not linearly related to the strength of blockiness and blurring.

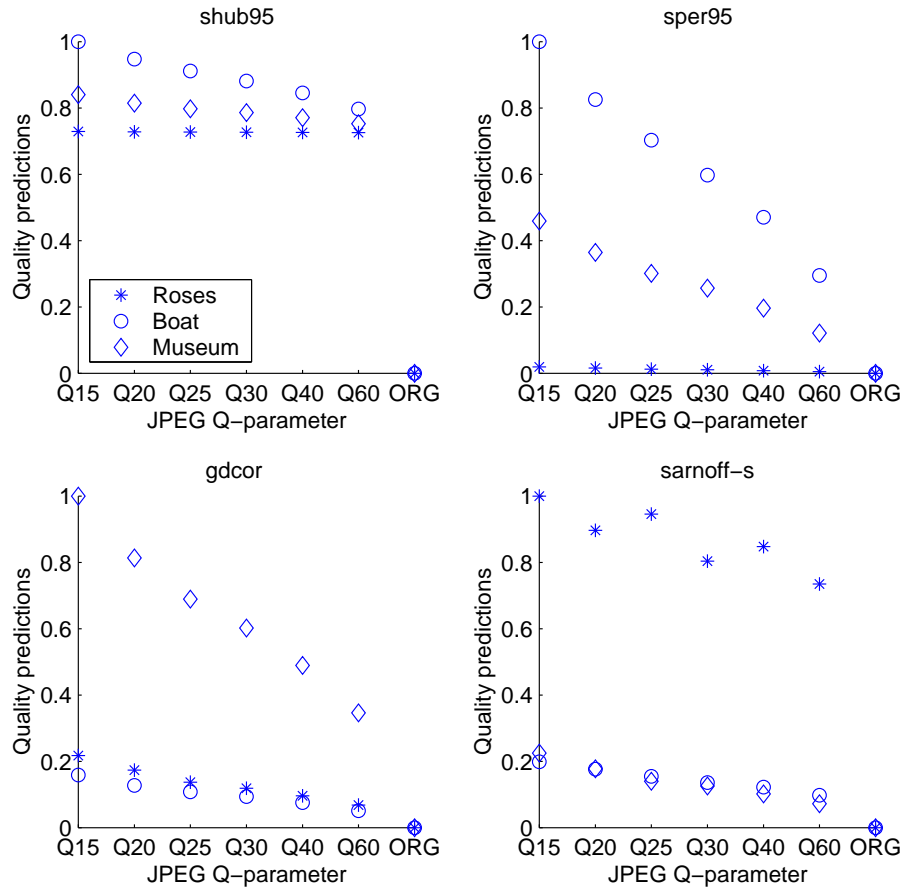


Figure 5.5: The quality predictions of four different measures as a function of the Q-parameter used to code the three images. On the y-axis the quality predictions are a distance between the original image and a JPEG coded version of it. Each graph represents one of the selected quality measures. Top left: shub95, top right: sper95, bottom left: gdcor and bottom right: sarnoff-s. For each quality measure the ranges of predictions for at least two scenes are different. Differences can also be noticed between the predictions of different quality measures. For instance, in the case of the quality measures shub95 and sper95, the scenes *roses* and *museum* are of better image quality than the scene *boat*. This is in contrast to the predictions of the quality measures gdcor and sarnoff-s.

5.3.1 Stimulus set

For each of the three scenes selected in the previous section four different JPEG coded versions with Q-parameters of 15, 25, 40 and 60 were used. For each scene the four coded versions and the original image were combined into 10 unique image pairs. Each image pair contained two images of the same scene. The image pairs were formed such that each JPEG coded version of a scene was compared once with all other JPEG coded versions of the same scene.

5.3.2 Procedure

The experiment consisted of four sessions. In each session 6 subjects participated. In each session the subjects were asked to judge a different attribute. The judged attributes were: quality, blockiness, ringing and blur. The same six observers participated in all sessions. The observers were seated at a distance of 0.80 m from a BARCO monitor in a dimly lit room. The two images of an image pair were shown simultaneously on the screen, one on the right hand side of the screen and one on the left hand side of the screen. In total 30 judgements were obtained.

In each session the subjects were asked to rate the difference in attribute strength between two images on a discrete numerical categorical scale from 0 up to 4. A zero indicated that the observer perceived no difference between the two images. The largest perceived difference should be rated 4. The sign indicated the image with the highest quality or the largest amount of blockiness, blurring or ringing.

The following description of the attributes blockiness, blurring and ringing were given to the observers (Yeun and Wu, 1998):

- **Blockiness** are visible discontinuities between the boundaries of adjacent blocks. This is perceived as horizontal and vertical edges in an image.
- **Blurring** manifests itself as a loss of spatial detail and a reduction in sharpness of edges.
- **Ringing** is most evident along high contrast edges in areas of generally smooth texture. It appears as a shimmering or rippling outwards from the edge up to the encompassing block's boundary.

At the end of each session the subjects were asked to circle, on a printed version of the scenes, those locations they specifically looked at to judge the perceived quality or amount of blockiness, blurring or ringing.

5.3.3 Results and discussion

In *Chapter 3* it has been shown that observers tend to use separate quality rating scales for each scene due to the differences in scene content. If subjects use separate rating scales for

5.3. Experiment 1: attribute scaling within a scene

Table 5.1: Experiment 1: for each scene the squared Pearson correlation coefficients, (r^2), between the quality judgements and the blockiness, blurring and ringing judgements are given.

		blockiness	blurring	ringing
roses	quality	0.97	0.96	0.96
boat	quality	0.98	0.99	0.98
museum	quality	0.94	0.98	0.98

each scene this implies that the quality judgements are not linked across scenes with the result that the perceived image quality of the different scenes is not comparable. For that reason in this section the quality data are analyzed for each scene separately. The judged quality differences are transformed into quality scale values on an interval scale by means of the program DifScal (Boschman, 2001). We assume that the observers use the categories on the quality and impairment scale in the same way. Hence the data is pooled over all six subjects. Per scene this results in a stimulus configuration with a quality scale value for each JPEG coded version and the original. The same analysis is performed for the attributes blockiness, blurring and ringing.

Figure 5.6 shows the DifScal scale values of the three scenes for each attribute separately. The scale values of the scenes *boat* and *museum* are linearly transformed such that the summed squared error between each scene's scale values and those of the scene *roses* is minimized. For each scene the perceived quality decreases with decreasing Q-parameter. The perceived image quality is linearly related across scenes. Also the attribute strengths of blockiness, blurring and ringing are linearly related across scenes. All attribute strengths increase with decreasing Q-parameter. Evidently the perceived ringing does not saturate for low-quality images (Q15 and Q25) but the data show a tendency that for low Q-values perceived ringing increases less steeply than perceived blockiness and blurring.

In figure 5.7 the quality, blockiness, blurring and ringing scale values are shown for each scene separately. The scale values of the attributes: blockiness, blurring and ringing were linearly transformed such that the summed squared distances between each attribute's scale values and the quality scale values is minimized. Due to this the sign of the attribute strengths is reversed. For each scene the attribute strengths are linearly related to the perceived image quality. The squared Pearson correlation coefficients, r^2 , between the perceived image quality and the attribute strengths are given in Table 5.1.

The subjects circled on a printed version of the three scenes the location they looked at to judge the quality, blockiness, blurring and ringing differences. In figure 5.8 the regions subjects looked at to judge the perceived image quality are marked by circles. At each location used for the image quality judgements also the number of subjects that judged blockiness, blurring and ringing are given. For instance, in the scene *Roses* three overall regions are indicated that were used to judge the image quality. One of these locations is the background. Five subjects judged also the blockiness strength at the background. This figure shows that the locations used to judge the image quality were also considered when the subjects judged the three attributes separately.

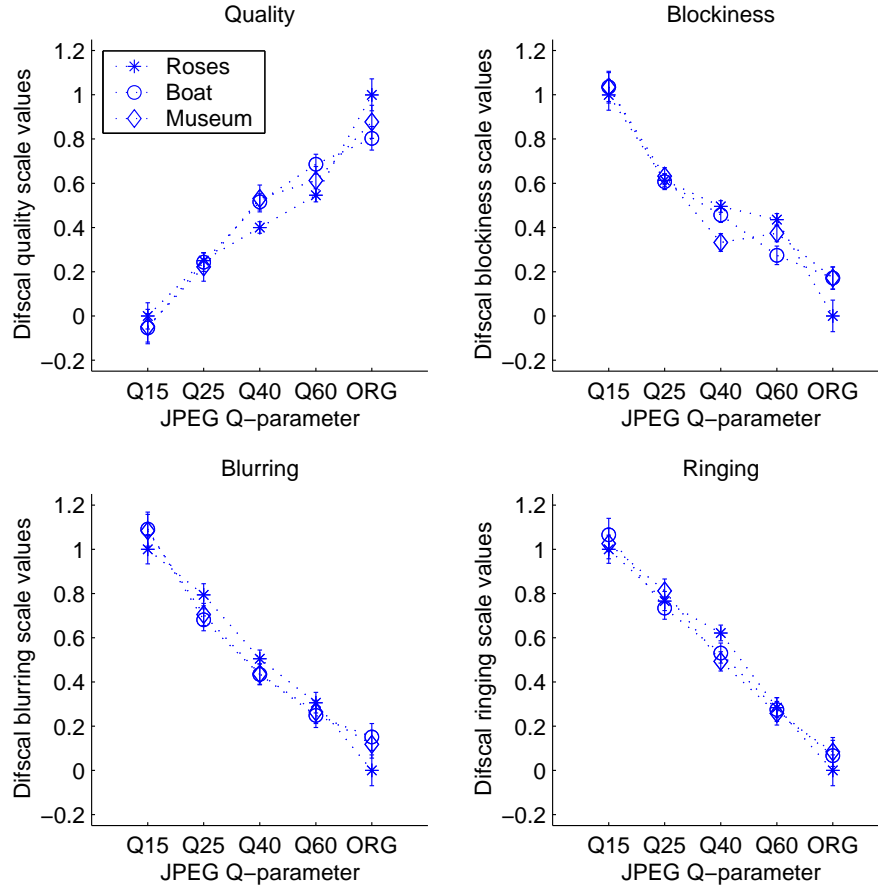


Figure 5.6: The quality (top left), blockiness (top right), blurring (bottom left), and ringing (bottom right) DifScal scale values as a function of the JPEG Q-parameter used to code the three scenes *roses*, *boat* and *museum*. The image quality of all three scenes are linearly related. This holds also for the attributes blockiness, blurring and ringing.

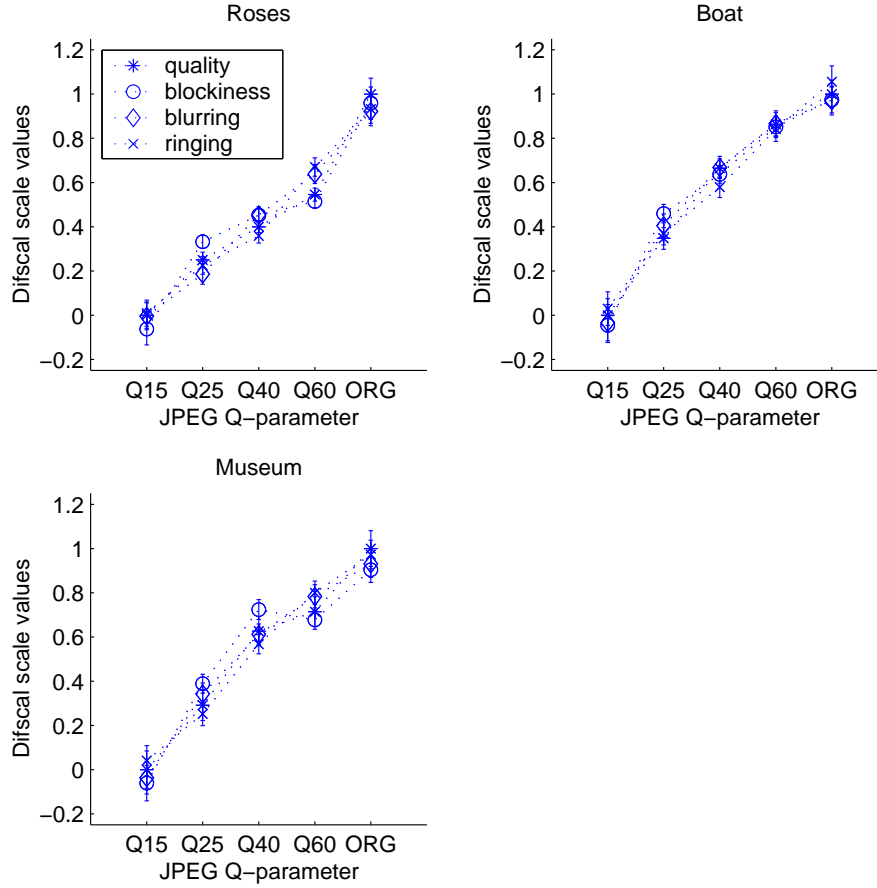


Figure 5.7: The quality, blockiness, blurring, and ringing DifScal scale values as a function of the JPEG Q-parameter used to code the three scenes *roses* (top left), *boat* (top right) and *museum* (bottom left). The distortion strength of the attributes blockiness, blurring and ringing are linearly related to quality for each scene. The strength of the attribute ringing does not saturate for low quality JPEG images (Q25 and Q15). In all panels a linear transform is applied on the perceived attribute strengths. The sign of the judged attributes strengths is reversed.

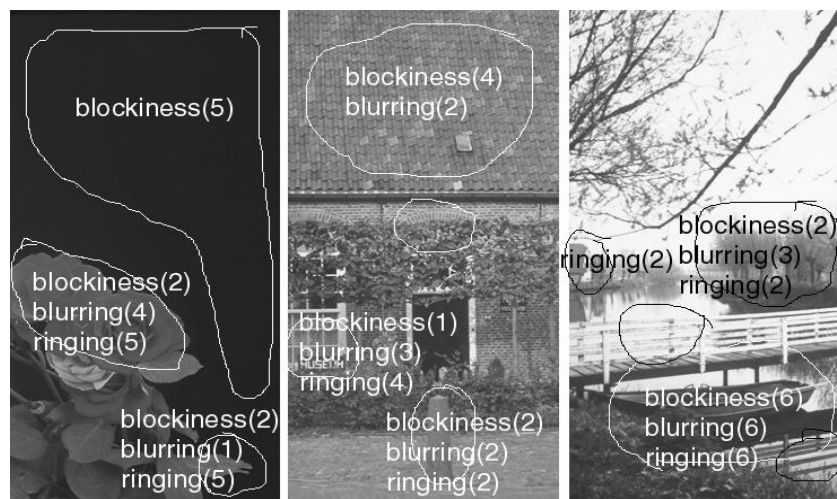


Figure 5.8: The subjects indicated for each scene the locations they looked at to judge the image quality, blockiness, blurring and ringing differences. In each scene the locations of image quality are pointed out by circles. For each region the number of subjects who observed also a particular attribute are given.

In *Chapter 4* it was shown in two separate experiments that the attribute pairs blockiness and blur are linearly related as well as are blockiness and ringing. In the experiment described in the present section the stimulus set contained different scenes and was expanded with low quality images. Also for this stimulus set the results are in line with those of *Chapter 4*. Therefore the image quality of JPEG coded images can be predicted in one dimension even though distinct distortions are perceived.

The JPEG-coded images were created with the default quantization table of JPEG. This implies that the impairment strength is controlled by a single parameter. For within-scene comparisons it is therefore expected that the attribute strengths correlate highly. In the study of Willemsen (1997) an attempt was made to manipulate the three attribute strengths separately. For that purpose the quantization matrix was altered. In such a case the perceived image quality is still mainly determined by the perceived blockiness strength (Willemsen, 1997). Based on these findings a measure that predicts the perceived blockiness strength can also be a good indicator of the perceived image quality for images that are not coded with JPEG's default quantization table.

The conclusions reached for the ringing strength of low-quality images in the present section differ from those obtained by de Ridder and Willemsen (2000) using percentage scaling and single stimulus scaling. In the present section we used comparison scaling and ringing does not saturate for low quality images. Possible reasons for this difference might be the scaling technique, where comparison scaling is considered more sensitive than single stimulus scaling as used by de Ridder and Willemsen (2000). Because of this the saturation of

ringing could be a scaling artifact. On the other hand, the scene effect can not be neglected, since the scenes in this chapter are different from those used by de Ridder and Willemsen (2000). The authors used the same images *child* and *girls* as in *Chapter 4*, though also highly impaired images were incorporated in the stimulus set.

Another consideration is that the attributes blockiness, ringing and blurring are difficult to separate in JPEG coded images. They interfere and an indication of a low quality can bias the ratings of the subjects. For instance if the quality decreases subjects know that the distortions increase. Therefore it could be that the distortion strengths are highly linked to quality. In the case of percentage scaling, subjects judge the overall image impairment as well as the three attributes for each image at the same time. This can facilitate the task of the observers to differentiate between the characteristics of the different distortions and give a more accurate relation between the strength of different distortions.

5.4 Experiment 2: attribute scaling across scenes

In *Chapter 4* as well as in experiment 1 of the present chapter we studied the relation between the distortion strengths and the image quality for each scene separately. This resulted in the conclusions that the distortion strengths are linearly related and that therefore a single attribute can be used to determine the image quality. However the linear relation between the distortion strengths still leaves two undetermined variables, a scaling factor and an offset. In this section we investigated whether these parameters are the same for each scene or whether the relationship between the distinct distortions is scene dependent.

The results of *Chapter 3* showed that subjects use separate quality rating scales if the scenes are not compared explicitly. Therefore in the following experiment the subjects rated the perceived quality differences also between two images containing different scene content. Through this experimental design, the perceived quality was comparable across scenes and the offset and scaling factor between the scenes could be determined. The same holds for the attributes blockiness, blurring and ringing.

5.4.1 Stimulus set

The same JPEG coded versions of the scenes *roses*, *boat* and *museum* as in the first experiment were used. The image pairs were formed such that the stimulus set also contained two images of different scene content. Thus this set contained pairs of images with the same scene content as well as images with different scene content. Each combination of two images with different scene content was included in the image set. However only six image pairs with the same scene content but varying Q-parameter were selected, namely *original* - Q60, *original* - Q25, Q60 - Q40, Q60 - Q15, Q40 - Q25 and, Q25 - Q15. In total the stimulus set contained 93 unique image pairs.

Table 5.2: Experiment 2: for each scene separately as well as across scenes the squared Pearson correlation, r^2 , between the quality judgements and the blockiness, blurring and ringing judgements is given.

		blockiness	blurring	ringing
all scenes	quality	0.95	0.94	0.97
	blockiness		0.93	0.90
	blurring			0.95
roses	quality	0.97	0.92	0.98
boat	quality	0.95	0.99	0.99
museum	quality	0.93	0.98	0.97

5.4.2 Procedure

Again the same attributes were judged as in the first experiment, namely quality, blockiness, blurring and ringing. Each attribute was judged in a separate session by six observers. The two images of an image pair were shown simultaneously on the screen, one on the right hand side of the screen and one on the left hand side of the screen. The observers were asked to judge the difference in attribute strength between the images on a scale from zero to four. If no difference was perceived they had to judge zero. The largest perceived difference should be judged four.

5.4.3 Results

A DifScal analysis was performed for each attribute separately. In figure 5.9 the DifScal scale values for quality, blockiness, blurring and ringing are shown for all three scenes separately. On the x-axis the JPEG Q-parameter is given and on the y-axis the DifScal scale values. The DifScal scale values are scaled for each attribute between the values 0 and 1. The best perceived image quality is mapped to 1 and the worst image quality to 0. In the case of the judged attributes, the largest attribute strength is mapped to 1 and the smallest strength to 0.

In the left panels of figure 5.10 the relationship between the DifScal quality scale values and those of the three attributes blockiness, blurring and ringing is shown. In the right panels of the same figure the three attributes are compared pairwise. For all panels the regression line between the compared DifScal scale values is shown. The corresponding squared Pearson correlation coefficients, r^2 , are given in Table 5.2.

Figures 5.9 and 5.10 show that the attribute strengths correlate highly with the perceived image quality. As in the results of section 5.3 there is a tendency for the attribute ringing to grow less steep at low quality values than the attributes blockiness and blurring. Nevertheless, in a first approximation the perceived image quality depends linearly on its underlying attributes: blockiness, blurring and ringing for each scene. The results substantiate the conclusions of *Chapter 4* and experiment 1, namely that the image quality can be predicted by the strength of a single attribute.

Figure 5.9 is comparable to figure 5.6 in the previous section though more variation in the data can be observed in figure 5.9. This figure reveals that the scale use for the scene *Roses* is slightly different from that for the scenes *Boat* and *Museum*. The perceived quality range of the scene *Roses* is larger and for instance the version compressed at Q60 is judged to have similar quality as the scenes *Boat* and *Museum* compressed at Q40. Scene differences can also be noticed for the judged blockiness strength. Again the perceived blockiness is larger in the scene *Roses*. Less pronounced scene differences can be observed for the attributes blurring and ringing.

5.5 Performance of instrumental quality measures

The experimental quality data of sections 5.3 and 5.4 will be used to evaluate the four selected instrumental quality measures of section 5.2.1 (*shub95*, *sper95*, *gdcor* and *sarnoff-s*). In section 5.5.1 the data of experiment 1 are used to study the performance of each measure within a scene. In section 5.5.2 the quality predictions are assumed to be scene independent and are compared with the perceived image quality obtained in experiment 2. Finally in section 5.5.3 the influence of the scenes in the stimulus set on the evaluation is demonstrated.

5.5.1 Performance within a scene

The model predictions for the four selected quality models are compared to the subjective data of experiment 1. Each quality measure should at least predict the difference in quality between the various JPEG coded versions within each scene. In figure 5.11 the predictions of each quality measure as well as the perceived quality are scaled between 0 and 1 for each scene separately. The best perceived and predicted quality is scaled to 1 and the worst quality to 0. For each scene the linear regression line is shown between these predicted values and the subjective quality data of the first experiment.

Figure 5.11 shows that all instrumental quality measures predict the rank-order of quality degradations within a scene correctly (Spearman $\rho = 1$). In addition three of the measures (*sper95*, *gdcor*, *sarnoff-s*) also show a good linear relation between the predicted and perceived image quality, indicated by the r^2 values in each panel. In the predictions of the measure *shub95*, see top row in figure 5.11, the difference between the predicted image quality of the original and the JPEG coded versions is too large. This results in a poor correlation between the quality predictions and the perceived image quality. This measure is highly sensitive for images that do not differ much from the original but hardly differentiates between the JPEG images in this stimulus set.

So far the quality predictions have been analyzed for each scene separately. Therefore, it remains unclear whether the instrumental quality measures predict the image quality on an absolute scale. This is considered in the following section using the subjective data of experiment 2 which includes explicit comparisons across scenes.

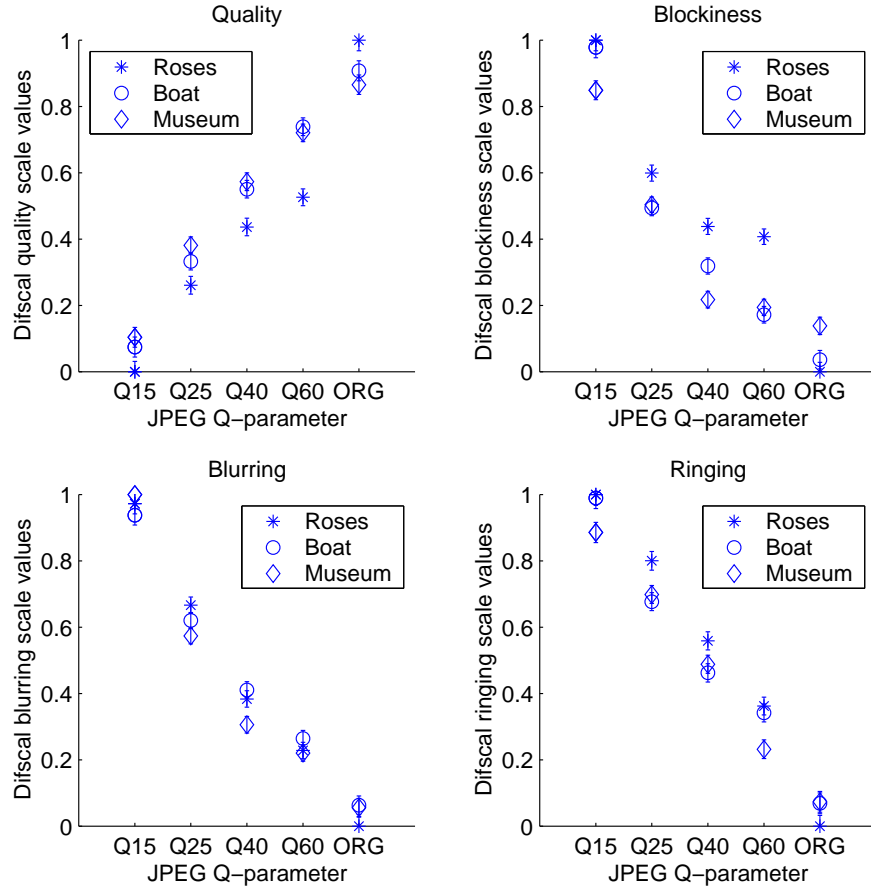


Figure 5.9: The quality (top left), blockiness (top right), blurring (bottom left), and ringing (bottom right) DifsCal scale values as a function of the JPEG Q-parameter used to code the three scenes *roses*, *boat* and *museum*. The perceived image quality and the attribute strength were judged across scenes. The quality and the attribute strength are given in separate graphs. The JPEG compression levels, Q-parameter, are given on the x-axis and the subjective scale values on the y-axis.

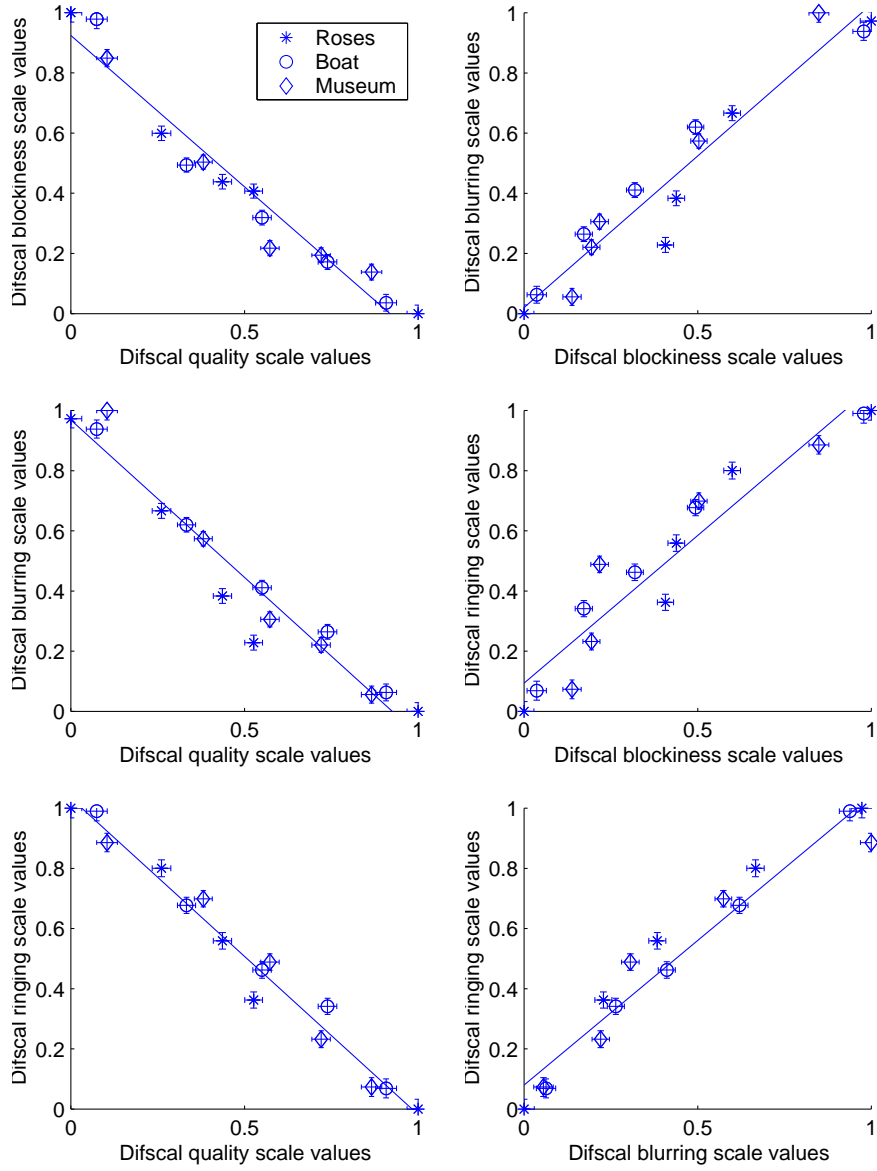


Figure 5.10: The results of the second experiment. The left side panels show the perceived image quality versus the three attributes blockiness, blurring and ringing. The right side panels show the relation between the three attributes.

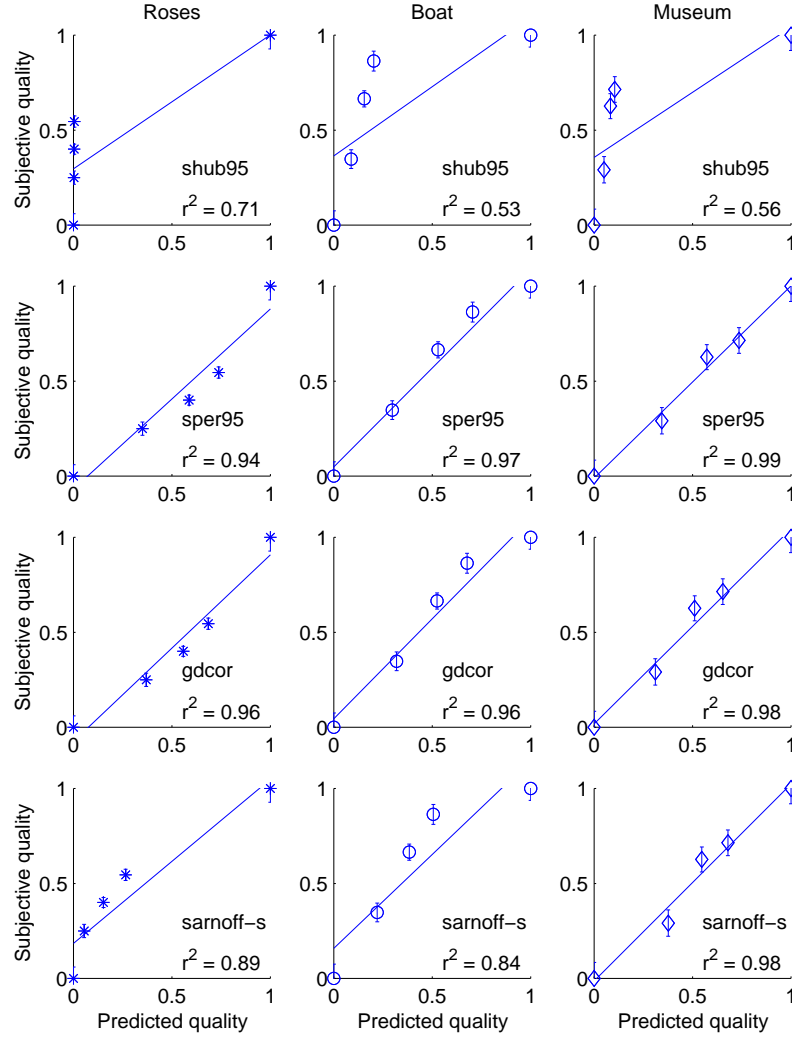


Figure 5.11: The correlation between the quality predictions and the subjective quality data per scene. The predicted quality is given on the x-axis and the perceived quality on the y-axis. Each row shows the results of a particular quality measure, from top to bottom these are *shub95*, *sper95*, *gdcor* and *sarnoff-s*. The scenes are shown in the columns, from left to right *roses*, *boat*, and *museum*. In each graph the regression line between the predicted quality and the perceived quality is drawn.

5.5.2 Performance across selected scenes

The model predictions for the four selected quality models are compared to the subjective data of experiment 2 in figure 5.12. The predictions are normalized across scenes in the range from 0 to 1. The best perceived and predicted quality is scaled to 1 and the worst quality to 0.

An instrumental quality measure is suitable if it can predict the perceived image quality across scenes. In such a case the quality predictions scale is an absolute scale that is scene independent. First of all, the meaning of the value “one” on the quality prediction scales is the same for all scenes and all measures. This is due to the fact that the predicted quality of an original scene is always one (the best quality is transformed to one). Compared to the perceived image quality this assumption of the instrumental image quality measures is reasonable. The perceived image quality of the original of each scene differs slightly but overall their quality difference is small and does not extend to the next level of a JPEG coded version.

Although the predicted image quality is calculated as a dissimilarity with respect to the original it should not be a problem conceptually to obtain quality predictions across scenes. Figure 5.12 shows, however, that the correlation between the predicted and perceived image quality is poor for all measures. The graphs show that for each measure the image quality is predicted incorrectly for at least one scene. Especially the measures shub95 and sper95 show a large deviation between predicted and perceived image quality for all three scenes. For the measures gdcor and sarnoff-s mainly one scene causes the poor correlation.

The measure shub95 shows the same problem as in the previous section, that is the step on the scale from the original to the first level of JPEG coding is too large.

If we consider the predicted quality differences across measures the following can be noticed. The measures shub95 and sper95 have the same fault in their predictions. The scenes are predicted in the same order even though the differences between scenes is smaller in shub95. The quality of the scene *roses* is predicted as the best, then the scene *museum* and as worst quality the scene *boat*. Even though perceptually the quality differences between the scenes are small, the various JPEG coded versions of the scene *roses* were judged of lower quality than those of the scenes *museum* and *boat*. However the quality predictions show the opposite. Only in the case of the HVS measure sarnoff-s the scene *roses* is predicted as the scene that suffers most from the JPEG coding artifacts although the quality differences between the scenes are exaggerated.

5.5.3 Performance across scenes in general

In the previous section it has been shown that the evaluated quality measures do not predict the perceived image quality across scenes. In this section we will show that not each set of scenes can reveal the difference in performance within and between quality measures. In *Chapter 2* we introduced a selection procedure and applied it to obtain the scenes that were used in the experiments of the present chapter. Now let us consider the scenes used in *Chapter 3*, where different selection criteria were used. These scenes, referred to as scenes in

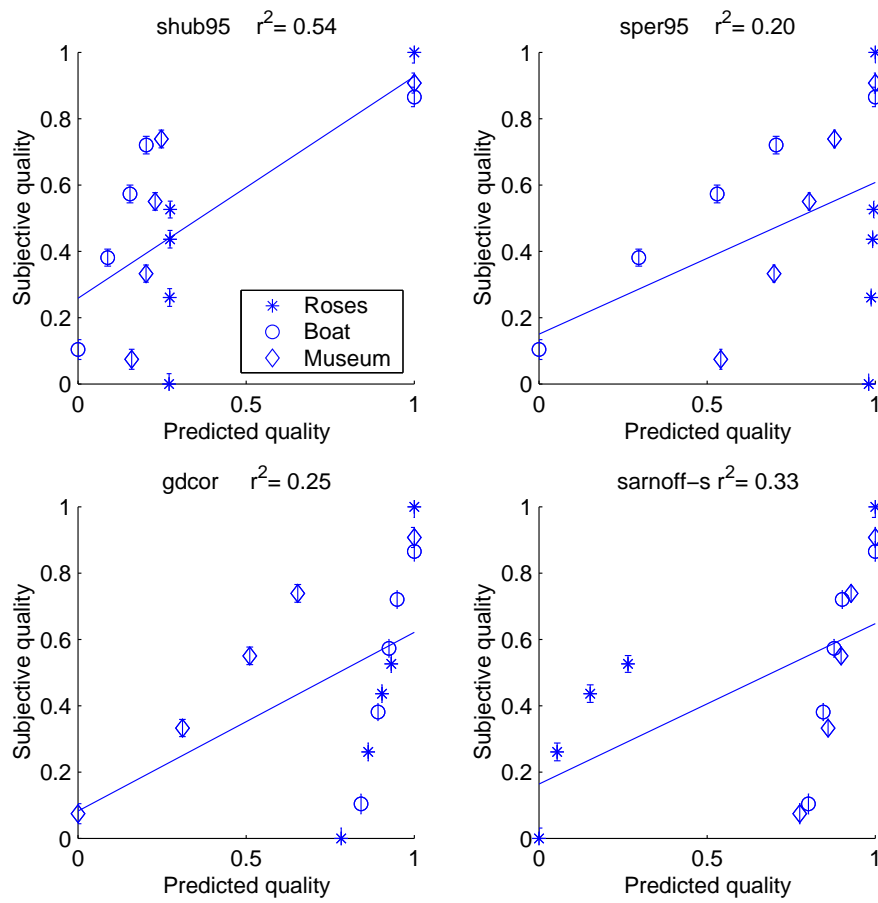


Figure 5.12: The correlation between the quality predictions and the subjective quality data across scenes. The predicted quality is given on the x-axis and the perceived quality on the y-axis. In each graph the regression line between the predicted quality and the perceived quality is drawn.

5.6. Performance of the single-ended blockiness measure

Table 5.3: The squared Pearson correlation coefficients, r^2 , between subjective quality judgements and the predictions for the two image sets.

	shub95	sper95	rmse	gdcor	sarnoff-s	sarnoff
selected scenes: <i>roses,</i> <i>boat,</i> <i>museum</i>	0.54	0.20	0.19	0.25	0.33	0.45
scenes in general: <i>photographer,</i> <i>country-road,</i> <i>shopping-street,</i> <i>woman</i>	0.50	0.73	0.76	0.61	0.39	0.65

general, were selected such that different image characteristics were incorporated. We will investigate whether these scenes chosen on the basis of their scene content and criticality, which is their expected degree of difficulty of coding, can differentiate within and between the quality measures (Yuyama *et al.*, 1998). The positions of these scenes in the cluster tree as obtained in section 5.2.2 are given in figure 5.13. As can be seen the four scenes are divided across two clusters. Three scenes belong to the same cluster and therefore they should not discriminate well within and between the instrumental quality measures.

The image quality predictions of these four scenes together with the selected scenes as used in the previous sections are normalized for each of the four quality measures. Figure 5.14 shows that the predictions of the scenes in general, *country-road*, *woman*, *photographer* and *shopping-street*, indeed differentiate less within and between the measures than the selected scenes *roses*, *boat* and *museum*.

The predictions of the algorithms for the scenes *country-road*, *woman*, *photographer* and *shopping-street* versus the subjective quality judgements of Chapter 3 are given in figure 5.15. For these four scenes the quality measures perform similar or better if we compare the results to those obtained for the selected scenes in the previous section. A summary of the squared Pearson correlation coefficients r^2 is given in Table 5.3. The measures predict the image quality fairly well if scenes from the same cluster are used. Especially the measures sper95 and gdcor perform much better for these scenes. This shows that the performance of instrumental quality measures depends on the used scenes. Therefore, one should be careful in generalizing the results if instrumental measures are evaluated for a small image set.

5.6 Performance of the single-ended blockiness measure

In Chapter 4 we introduced a single-ended measure to predict the perceived blockiness. The performance of this measure was tested for four scenes. Per scene, it was shown that the blockiness predictions correlate highly with the perceived blockiness and that using a large image set for a great number of scenes the blockiness was predicted monotonically for 133

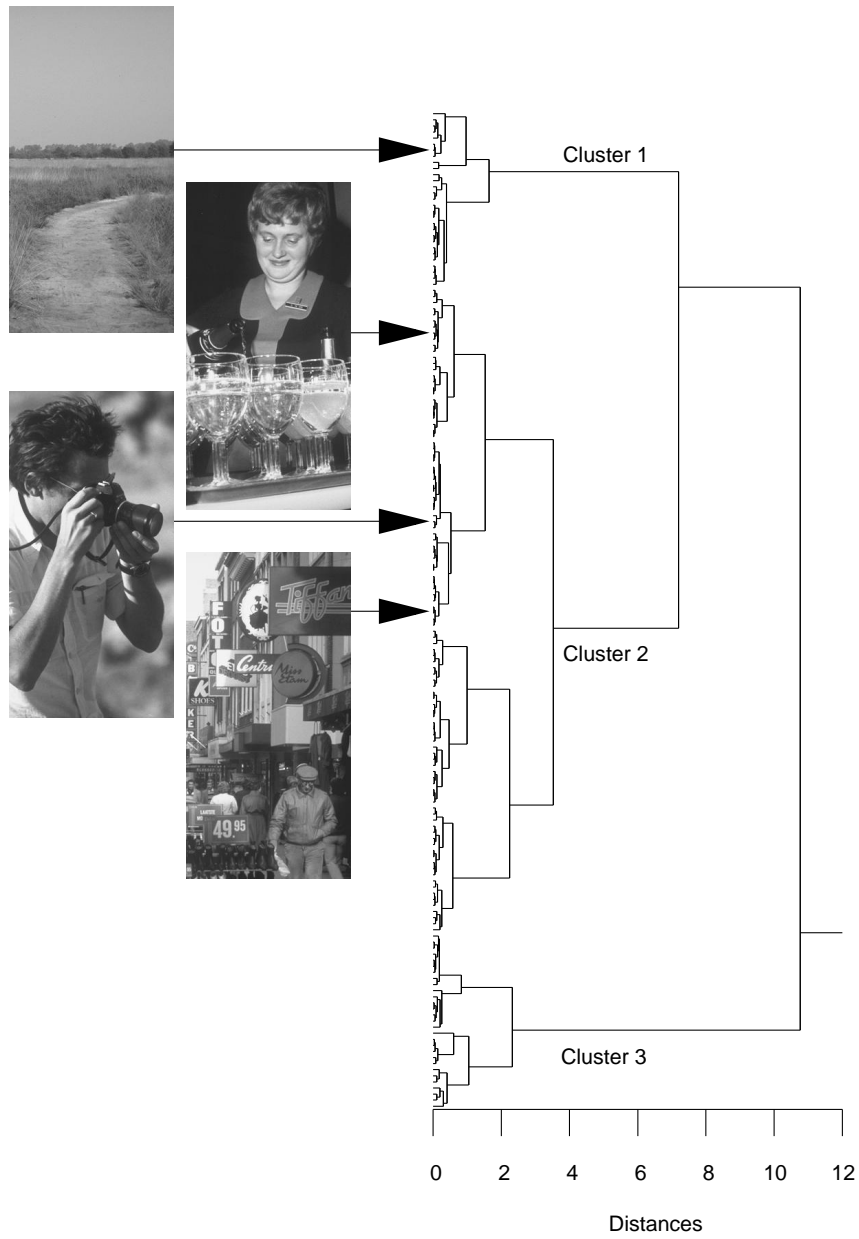


Figure 5.13: The location of the four scenes in general in the hierarchical scene cluster tree. Three scenes are members of cluster 2.

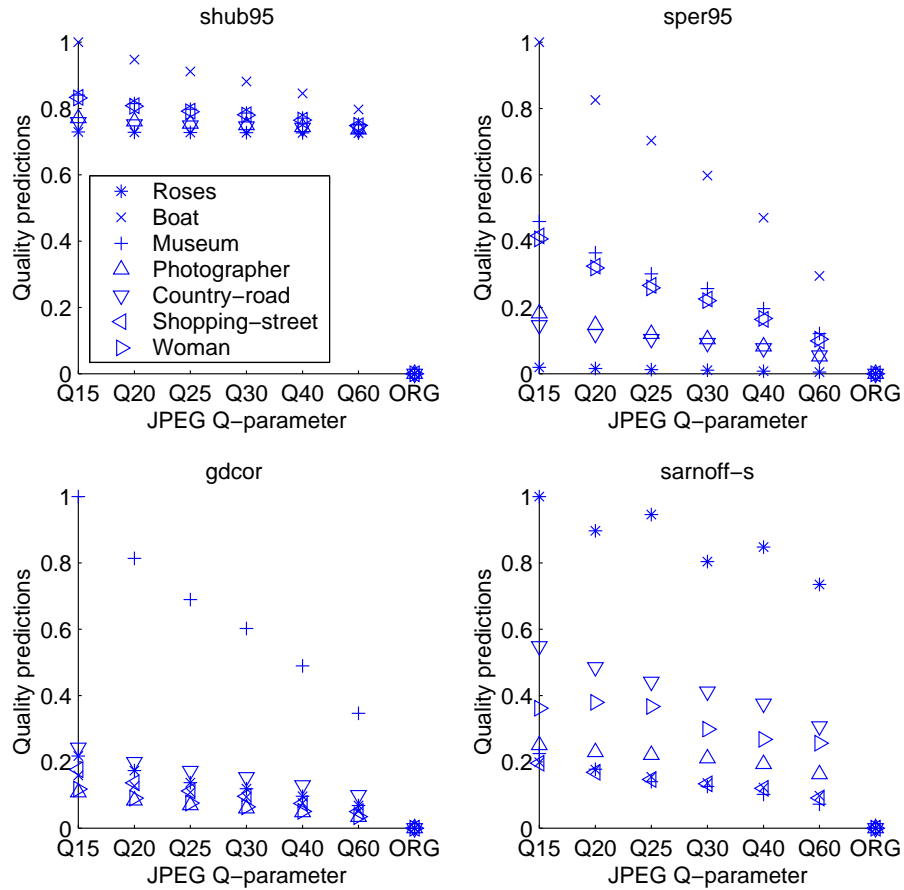


Figure 5.14: Quality predictions of four quality measures for seven scenes coded at various JPEG levels. The scenes *photographer*, *country-road*, *shopping-street* and *woman*, were used in the experiments of Chapter 3. They were chosen on the basis of their scene content and criticality. The scenes *roses*, *boat* and *museum* were selected by objective selection criteria such that they discriminate within and between the instrumental quality measures.

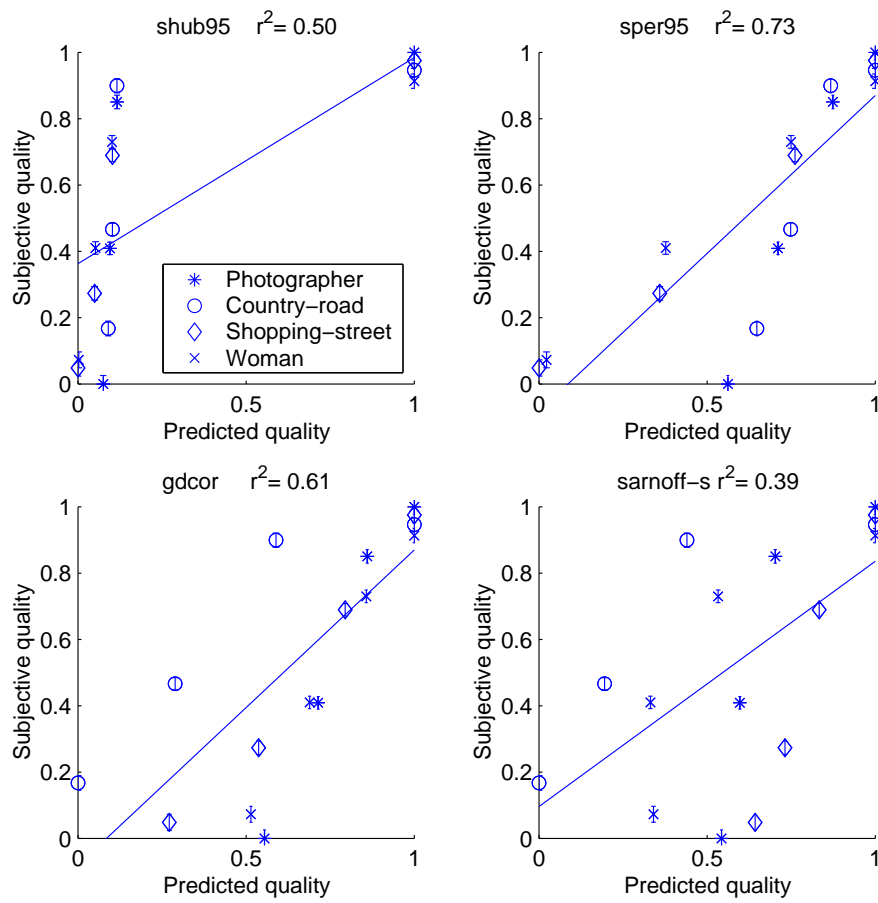


Figure 5.15: The correlation between the quality predictions and the subjective quality data across four scenes in general. The predicted quality is given on the x-axis and the perceived quality on the y-axis. In each graph the regression line between the predicted quality and the perceived quality is drawn.

5.6. Performance of the single-ended blockiness measure

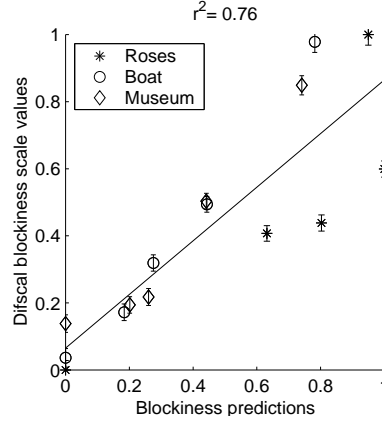


Figure 5.16: The correlation between the predicted blockiness and the perceived blockiness across scenes. The normalized blockiness predictions using the single-ended blockiness measure are given on the x-axis. On the y-axis the DifSca blockiness scale values obtained for across-scene judgements are given.

out of 164 scenes. In this section the performance of the single-ended blockiness measure is tested for a stronger assumption namely that the blockiness scale is an absolute scale and is therefore independent of the scene content. This implies that the blockiness predictions are linked across scenes and should correlate highly with the perceived across-scene blockiness. This assumption is tested by means of the subjective data obtained in section 5.4.

Blockiness predictions were obtained for three scenes: *roses*, *boat* and *museum*. The scene *roses* is one of the 21 scenes that were not predicted monotonically in *Chapter 4*. For each scene the original image and 4 JPEG coded versions (with a Q-parameter of 15, 25, 40 and 60) were used. The blockiness predictions were normalized across scenes in the range from 0 to 1. The highest predicted blockiness was scaled to 1 and the lowest blockiness to 0. These normalized blockiness predictions were compared to the across-scenes blockiness judgements of section 5.4. Figure 5.16 shows that the predicted blockiness correlates quite well with the perceived blockiness ($r^2 = 0.76$).

In section 5.4 it was shown that for across-scene judgements the perceived image quality correlates highly with the perceived blockiness strengths. We can now even go a step further by combining the two observations for across-scene judgements: 1) the perceived and predicted blockiness correlate quite well and 2) perceived blockiness strength and image quality correlate highly. Therefore the single-ended blockiness measure should be able to predict image quality across scenes for sequential baseline coded JPEG images.

Two sets of subjective quality data are used to test this expectation. The first set comprises quality judgements for the scenes: *roses*, *boat*, and *museum* (see section 5.5.2). The second set consisted of the quality judgements as used in section 5.5.3 for the scenes: *country-road*, *photographer*, *shopping-street* and *woman*. The two panels in figure 5.17 show the relation

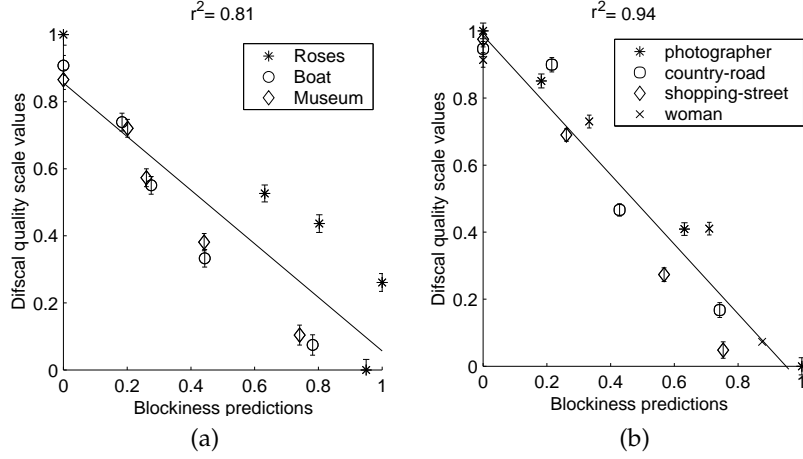


Figure 5.17: Across quality judgements are compared to the predicted blockiness for two different image sets: (a) *roses*, *boat* and *museum* from Chapter 5 and (b) scenes in general: *photographer*, *country-road*, *shopping-street* and *woman* from Chapter 3.

between the predicted blockiness and the judged image quality for these two stimulus sets. In both cases the predicted blockiness correlates highly with the perceived image quality. Evidently, for both image sets the single-ended blockiness measure performs significantly better than the instrumental quality models tested in section 5.5.2.

5.7 Stimulus configuration based on quality predictions and experimental data

In section 5.5.2 it was shown that the double-ended instrumental quality measures: *gdcor*, *sper95*, *shub95* and *sarnoff-s*, do not predict the quality of JPEG coded images on an absolute quality scale. For the three scenes *roses*, *boat* and *museum*, the predictions of these four instrumental measures correlate poorly with the subjective quality judgements, while the single-ended blockiness measure led to better agreement. It seems that this measure predicts the image quality of JPEG coded images on an absolute quality scale.

In this section, the performance of instrumental quality measures on an absolute quality scale is visualized by means of an MDS stimulus configuration. For that purpose the predictions of 67 double-ended instrumental quality measures (section 5.2.1) and the single-ended blockiness measure are compared to the judged image quality of the three scenes *roses*, *boat* and *museum*. Such a stimulus configuration can give a better understanding of the differences between the predicted image quality and the perceived image quality. The configuration will also show whether these three scenes actually differentiate between the instrumental quality measures.

The predictions of the 67 double-ended instrumental quality measures as well as the single-ended blockiness measure were normalized across the three scenes *roses*, *boat* and *museum*. The predictions were standardized by normalizing for each measure the overall RMSE to unity. The DifScal stimulus scale values of section 5.4 were used as a representation of an average human observer. The DifScal scale values were transformed such that the image with the best image quality was scaled to zero. The resulting transformed predictions were normalized by dividing each prediction by the RMSE taken across all DifScal scale values of the three scenes.

In the same way as in *Chapter 2* and section 5.2.1 a distance was calculated between each of the instrumental measures as well as between the measures and the average human observer by means of the inner product correlation. This dissimilarity matrix was fed into the MDS program *xgms* and the resulting two-dimensional stimulus configuration is shown in figure 5.18.

In this configuration the double-ended vision model, CCETT, is most similar to an average human observer. Also the predictions obtained by the single-ended blockiness measure lie close to the perceived image quality. This figure also substantiates the fact that the double-ended instrumental measures: *gdcor*, *sper95*, *shub95* and *sarnoff-s*, correlate poorly with the perceived image quality. The locations of these measures cover the first dimension which indicates that the three selected scenes indeed discriminate between these four instrumental measures.

5.8 Conclusions

Throughout all experiments described in this chapter it is evident that even though different distortions are visible in sequential baseline coded JPEG images, their strengths are linearly related. Moreover, the perceived attribute strengths are linearly related to the perceived image quality. This holds for within-scene as well as for across-scene judgements. Therefore image quality for sequential baseline coded JPEG images can be modeled by a single attribute.

Secondly, the performance of a number of quality measures was evaluated by means of within-scene and across-scene quality judgements. It can be concluded that most measures can predict the image quality of JPEG coded versions of the same scene. A stronger assumption namely that these measures predict image quality on an absolute quality scale is a requirement that most measures cannot cope with. The evaluation revealed that the quality predictions of most measures correlate poorly with across-scene quality judgements. In terms of an absolute quality scale, it can be concluded that the single-ended blockiness measure performs best.

We also compared the results for two image sets that were used to evaluate instrumental quality measures. Each set of scenes was selected by a different criterion. For the first set, scenes were selected which discriminate optimally within and between instrumental quality measures. For the second image set (scenes in general), scenes were selected on the basis of their coding criticality or merely difference in scene content. The former set of images

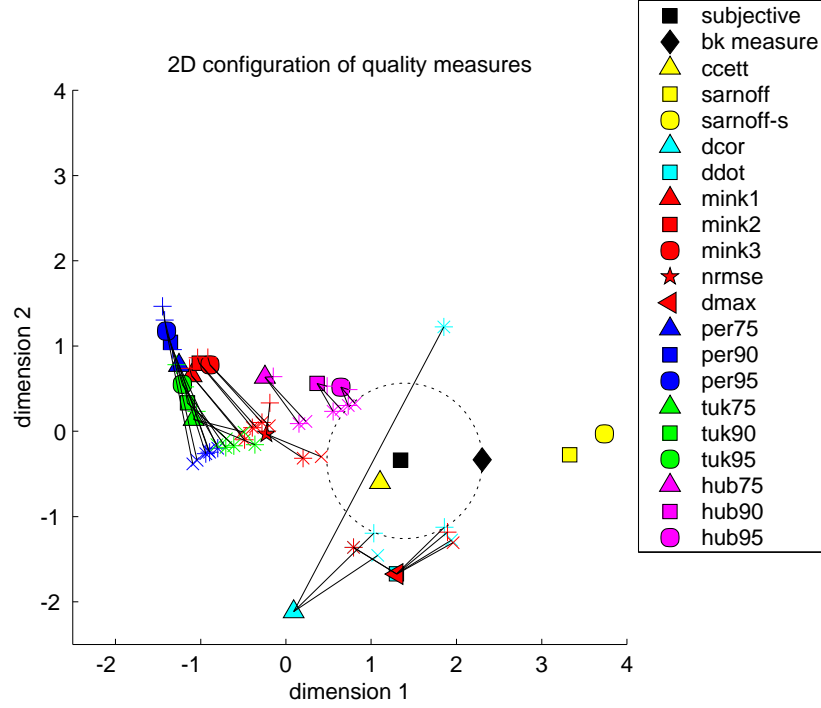


Figure 5.18: Two dimensional stimulus configuration of 67 double-ended instrumental quality measures, the single-ended blockiness measure and the average human observer. This configuration is obtained by *xgms* on the basis of quality predictions and judgements for the 3 scenes *roses*, *boat* and *museum*. The five symbols at the top of the legend show characterize the average human observer, the single-ended blockiness measure and the three vision models. The remaining 16x4 instrumental quality measures are represented by 16 sets of 4 symbols that are connected by lines. The measures applying the 16 different combination rules and: 1) a gray-scale-to-luminance transformation; 2) a gray-scale-to-luminance transformation together with *Sobel* filtering, +; 3) neither gray-scale-to-luminance transformation nor *Sobel* filtering, *; 4) *Sobel* filtering only, x. All instrumental quality measures within the dashed circle have the same or a better performance as the single-ended blockiness measure.

5.8. *Conclusions*

is a more critical test case than the latter. It could be shown that scenes which discriminate within and between instrumental quality measures indeed reveal the differences in performance of the objective quality measures. Scenes in general, which are selected by an intuitive criterion, are less suited to test the performance of all instrumental quality measures.

Chapter 6

Epilogue

During recent years, a number of image quality measures has been proposed (Lubin, 1993; van den Branden Lambrecht, 1996; Kayargadde and Martens, 1996c,d; Daly, 1993; Eskioglu and Fisher, 1995). Though the performance of most measures is tested by means of subjective quality judgements no measure can be pointed out that has a clear advantage (Ahumada, 1993). The main reason for this is that the measures are rarely compared using the same image set and corresponding subjective quality judgements. Joint-collaboratives are formed to define public databases of images and quality judgements such that the instrumental measures can be tested with the same material (Corriveau *et al.*, 2000; Carney *et al.*, 2000). This is a very useful approach to standardize an image quality measure. In relation to this work we attempted to enhance our understanding of how quality judgments and the predicted quality can be used to discriminate between instrumental quality measures.

The work in this thesis mainly focussed on two topics. First of all, if instrumental image quality measures are used as a substitute for human observers, the measures should not only predict the image quality of one particular type of impairment but also across various impairments. By the same reasoning quality measures should also predict the image quality across a wide range of scene contents. This issue was covered in detail in the *Chapters 2, 3 and 5*. Secondly, in every day situations (e.g. TV watching) human observers judge impaired images in the absence of an unimpaired original. Thus, the second topic of this thesis was whether image quality can be deduced from the impaired image only. For sequential baseline coded JPEG images, a single-ended blockiness measure was proposed which outperforms the quality predictions of a particular group of double-ended measures (*Chapters 4 and 5*).

As mentioned above instrumental quality measures are intended as a substitute for human observers. In that sense the performance of such measures is evaluated by means of subjective quality judgements. Therefore it is important to realize that quality judgements depend on the assessment technique used, including presentation of the material, the images and the task imposed on the observers. From the literature it is known that image quality judgements can be affected by the experimental procedure (de Ridder and Willemsen, 2000; van Dijk and Martens, 1996; ITU-R-JWP10-11Q, 1998). In *Chapter 3* we continue the work started by van Dijk and Martens (1996) who showed that single-stimulus scaling leads to

biased results when compression algorithms are tested that introduce different types of impairment. They argued that in single-stimulus scaling separate quality scales are used when observers can identify the different impairment types of each coder. They demonstrated that with stimulus-comparison scaling this problem can be solved since differently compressed images can be compared explicitly and therefore observers are forced to link the quality judgements across impairment types.

In *Chapter 3* we continued along these lines for stimulus-comparison scaling without and with explicitly comparing different impairment types. Two experiments were carried out with stimulus sets containing different impairment types introduced by wavelet-coding, DCTune, JPEG and low-pass filtering. In each experiment a single scene was impaired by these four processing methods. For one scene the observers seem to rate the perceived image quality on a single rating scale whereas for the second scene these findings could not be substantiated. This is probably due to the fact that in the second scene the quality ranges of the differently impaired images are not dissimilar enough. We elaborated on the idea that in stimulus-comparison scaling observers use separate rating scales for identifiable classes, if these classes are not compared explicitly, by using scene content as such classes. In this case subjects seem to use a separate quality scale if images with different scene content are not compared explicitly. This was demonstrated for wavelet- and JPEG-coded images. Further research is needed to substantiate these findings and test whether current quality assessment techniques are suited to obtain quality judgements across scenes and across impairment types. Based on the results described in *Chapter 3* we recommend that instrumental quality measures are best evaluated by means of subjective quality judgements obtained through explicitly comparing different impairment types and different scenes. This method can serve to differentiate better between instrumental quality measures and can form an addition to the currently used evaluation techniques (ITU-R-500-7, 1997).

Instrumental quality measures are usually classified according to the differences in their performance (Fuhrmann *et al.*, 1995). Instead of following this widely accepted method, we investigated in *Chapter 2* whether the measures are essentially different or not. This approach does not require a judgement about the performance of the measures, and therefore the classification is solely based on the measures' predictions. Since the time-limitations of subjective testing are not an issue in this procedure a large collection of images (containing a wide range of impairment types and scene content) can be used. The procedure was demonstrated by classifying a particular group of 67 instrumental quality measures. From this set of instrumental quality measures basically six clusters can be identified that were essentially different in their quality predictions for a large image set. The proposed technique is very helpful in visualizing the difference between quality measures. Moreover, such a classification can contribute to a better understanding of the added value of enhanced or new measures without the necessity of performing subjective tests. We suggest to only spend effort on gathering subjective data if it is known that a measure differs from already existing ones. Otherwise the added value of such a measure is minor.

In *Chapter 2* particular choices were made in relation to the used measure of proximity, indicating the relationship from one instrumental quality measure to another, and the clustering technique. The applied proximity measure (inner-product correlation) substantiates the assumption that the predictions of instrumental quality measures are defined up to a

scaling factor. Ward's hierarchical clustering was used to obtain compact clusters of quality measures. However, the inner-product correlation as well as Ward's hierarchical clustering influence the composition of the instrumental quality measures in a cluster. How the chosen proximity measure and the used clustering technique affect the composition of quality measures clusters should be investigated further.

Although a classification of instrumental quality measures on the basis of their quality predictions only can be used to differentiate between the quality measures, eventually their performance has to be tested by means of subjective quality judgements. It is common practice to use for that purpose a limited image set of, e.g., 3 or 4 scenes. Due to such a limited number of scenes it is hard to generalize the outcome of an evaluation. Thus the conflict exists between the need for a large image set in order to allow generalization, and the restriction to a small set for practical reasons.

In *Chapter 2* we described how a stimulus set can be chosen for subjective tests which reveals the differences between instrumental quality measures. The basic idea is that a stimulus set is chosen on the basis of the following two criteria: 1) each instrumental quality measure yields different results for the selected scenes and, 2) the selected scenes yield different results for instrumental quality measures. These selection criteria differ from those commonly used to select scenes for subjective tests. The criteria described above use the difference in predicted quality whereas usually the coding criticality of an image or the difference in scene content is used to select scenes. Thus, when instrumental quality measures are evaluated by means of an image set that discriminates within and between the measures, their usefulness is tested for a highly critical sample of images.

In *Chapter 2* and *Chapter 5* we used an initial image set of 164 scenes. The aim was to cluster these scenes into groups that discriminate between quality measures and to select from each cluster one scene that can be used in subjective tests. The obtained scene clusters were also compared to a priori defined scene classes, but the relation between the identified clusters and a priori defined scene content classes was weak. Further work is needed to identify the image properties of those images that differentiate between instrumental quality measures. In Wolf and Webster (1997), for instance, a procedure is described to select scenes for testing the quality of compressed video on the basis of their scene criticality. The scene criticality is estimated from spatial and temporal information. It would be of interest to see how well a priori defined criticality classes agree with the scene clusters obtained in our analysis.

In *Chapter 5* instrumental quality measures were evaluated for sequential baseline coded JPEG images. A stimulus set was chosen such that quality judgements were obtained for scenes that discriminate within and between the evaluated quality measures. The evaluation showed that, per scene, most measures' quality predictions correlate highly with the perceived image quality. Furthermore, it was investigated whether the instrumental measures predict the image quality on an absolute scale which would enable them to compare the image quality between different scenes. In this case the evaluation revealed that most instrumental quality measures can not cope with this additional requirement. A second image set (*Chapter 3*), chosen on the basis of different scene content, resulted in the same conclusion. Also in this case, most instrumental quality measures performed poorly for across-scene quality predictions. However, in this case the correlation between the predicted and

perceived image quality was slightly better. The same two image sets were also used to evaluate the performance of the single-ended blockiness measure developed in *Chapter 4*. In terms of an absolute quality scale it can be claimed that the single-ended blockiness measure performs best.

It is remarkable that the single-ended blockiness measure which works solely on an impaired image performed better than double-ended measures, which calculate a difference between the impaired image and its original.

The development of a single-ended blockiness measure was inspired by the work of Kayargadde and Martens (1996c,d). They developed single-ended measures based on the Hermite transform to estimate the perceptual strength of noise and blur. In their study both distortions were controlled independently and a stimulus set was used that was distributed uniformly across the perceptual space. In sequential baseline JPEG coded images the underlying attributes of image quality are blockiness, ringing and blur. Although different distortions are visible the underlying psychological space of image quality is approximately one dimensional. This is reflected in the fact that the perceived attribute strengths are linearly related to the perceived image quality. This was substantiated by various psychophysical experiments. The attribute strengths were measured by means of within-scene judgements as well as for across-scene judgements. Both experiments clearly revealed the linear relationship between the attributes and the perceived image quality. It should be noted that the blockiness measure is expected to predict image quality only in the case that images are coded with the default quantization matrix which is scaled by varying the Q-parameter. This is the most common practice in JPEG coding. Further research is needed to investigate the relationship between the attribute strengths and the perceived image quality if the quantization matrix is not simply scaled but if the quantization coefficients are manipulated separately. Moreover, in future work the predictions of the single ended blockiness measure should be compared to existing measures as for example described by Libert and Fenimore (1999) and Karunasekera and Kingsbury (1995).

Blockiness can be attributed especially to block-based DCT-coded images. However, to generalize the single-ended blockiness measure into an overall quality measure additional impairment types and their effects on the perceived image quality need to be investigated. The results of this thesis show that the attempt to model perceived image quality by a single-ended measure is promising, certainly if we consider the fact that the single-ended blockiness measure is to a lesser degree scene dependent than other proposed measures. Consequently this single-ended measure is robust enough to predict the image quality of JPEG coded images on an absolute scale.

Bibliography

- Ahumada, A., and Beard, B. (1998). "A simple vision model for inhomogeneous image quality assessment," in *Society for Information Display Digest of Technical Papers*, edited by J. Morrela, (Santa Ana CA), vol. 29 Paper 40.1.
- Ahumada, A. J. (1993). "Computational image quality metrics: a review," *SID Digest* **24**, 305–308.
- Ahumada, A. J., and Null, C. H. (1993). "Image quality: A multidimensional problem," in *Digital Images and Human Vision*, edited by A. B. Watson, (The MIT Press), pp. 141–148.
- Allnatt, J. (1983). *Transmitted-Picture Assessment*, (John Wiley and Sons Ltd, New York).
- Anderberg, M. R. (1973). *Cluster analysis for applications*, (Academic Press, New York).
- Beerends, J. G. (1997). "Objective measurement of video quality," KPN Research Netherlands, Internal Report .
- Black, M. J., and Marimont, D. H. (1998). "Robust anisotropic diffusion," *IEEE Transactions on Image Processing* **7**, 421–432.
- Boschman, M. C. (2001). "Difscal: A tool for analyzing difference ratings on an ordinal category scale," *Behavior Research Methods, Instruments, & Computers* **33**, 10–20.
- Boschman, M. C., and Roufs, J. A. (1997). "Text quality metrics for visual display units: An experimental survey," *Displays* **18**, 45–64.
- Carney, T., Klein, S. A., Tyler, C. W., Silverstein, A. D., Beutter, B., Levi, D., Watson, A. B., Reeves, A. J., Norcia, A. M., Chen, C.-C., Makous, W., and Eckstein, M. P. (1999). "The development of an image/threshold database for designing and testing human vision models," in *Human vision and electronic imaging IV*, edited by B. E. Rogowitz, and T. N. Pappas, vol. 3644, pp. 542–551.
- Carney, T., Tyler, C. W., Watson, A. B., Makous, W., Beutter, B., Chen, C. C., Norcia, A., and Klein, S. A. (2000). "Modelfest: year one results and plans for future years," in *Human vision and electronic imaging V*, edited by B. E. Rogowitz, and T. N. Pappas, vol. 3959, pp. 140–151.
- Corriveau, P., Webster, A., Rohaly, A. M., and Libert, J. (2000). "Video quality expert group: the quest for valid objective methods," in *Human vision and electronic imaging V*, edited by B. E. Rogowitz, and T. N. Pappas, vol. 3959, pp. 129–139.

- Cox, T. F., and Cox, M. A. A. (1994). *Multidimensional Scaling*, (Chapman & Hall, London).
- Daly, S. (1993). "The visible differences predictor: an algorithm for the assessment of image fidelity," in *Digital images and human vision*, edited by A. B. Watson, (The MIT Press), pp. 179–206.
- de Ridder, H. (1991). "Comparison of combination rules for digital-image-coding impairments," IPO Annual Progress Report 26 pp. 70–79.
- de Ridder, H. (1992). "Minkowski-metrics as a combination rule for digital-image-coding impairments," in *SPIE Human Vision, Visual Processing, and Digital Display III*, edited by B. E. Rogowitz, and T. N. Pappas, vol. 1666, pp. 16–26.
- de Ridder, H. (1996). "Current issues and new techniques in visual quality assessment." Proceedings of the IEEE International Conference on Image Processing pp. 869–872.
- de Ridder, H. (2001). "Cognitive issues in image quality," JEI **10**, 47–55.
- de Ridder, H., and Willemsen, M. C. (2000). "Percentage scaling: a new method for evaluating multiply impaired images," in *Human vision and Electronic Imaging V*, edited by B. E. Rogowitz, and T. N. Pappas, vol. 3959, pp. 68–77.
- Eskicioglu, A. A., and Fisher, P. S. (1995). "Image quality measures and their performance," IEEE Transactions on Communications **43**, 2959–2965.
- Everitt, B. S. (1993). *Cluster Analysis*, (Halsted Press, New York).
- Falkus, D. (1996). "Digital TV: a testing problem," International Broadcasting pp. 17–19.
- Fuhrmann, D. R., Baro, J. A., and Cox, J. R. (1995). "Experimental evaluation of psychophysical distortion metrics for jpeg-encoded images," Journal of Electronic Imaging **4**, 397–406.
- Gonzales, R. C., and Woods, R. E. (1992). *Digital image processing*, (Addison-Wesley publishing company, Inc).
- ITU-R-500-7 (1997). "Draft revision of recommendation ITU-R BT.500-7," Methodology for the subjective assessment of the quality of television pictures .
- ITU-R-JWP10-11Q (1998). "Investigation of contextual effects," Tech. Rep., International Telecommunication Union, Radiocommunication Study Groups.
- ITU-WP-2/12 (1995). "Selections from the ANSI draft standard: - digital transport of one-way signals - parameters for objective performance assessment," Temporary Document 54-E .
- Janssen, T., and Blommaert, F. (2000). "Visual metrics: discriminative power through flexibility," Perception **29**, 965–980.
- Karunasekera, S. A., and Kingsbury, N. G. (1995). "A distortion measure for blocking artifacts in images based on human visual sensitivity," IEEE Transactions on image processing **4**, 713–724.

- Kayargadde, V., and Martens, J.-B. (1994). "Estimation of edge parameters and image blur using polynomial transforms," *CVGIP: Graphical Models and Image Processing* **56**, 442–461.
- Kayargadde, V., and Martens, J.-B. (1996a). "Estimation of perceived image blur using edge features," *International Journal of Imaging Systems and Technology* **7**, 102–109.
- Kayargadde, V., and Martens, J.-B. (1996b). "An objective measure for perceived noise," *Signal Processing* **49**, 187–206.
- Kayargadde, V., and Martens, J.-B. (1996c). "Perceptual characterization of images degraded by blur and noise: model," *Journal of the Optical Society of America A* **13**, 1178–1188.
- Kayargadde, V., and Martens, J.-B. (1996d). "Perceptual characterization of images degraded by blur and noise: experiments," *Journal of the Optical Society of America A* **13**, 1166–1177.
- Libert, J. M., and Fenimore, C. P. (1999). "Visibility thresholds for compression-induced image blocking: measurement and models," in *Human vision and electronic imaging IV*, edited by B. E. Rogowitz, and T. N. Pappas, vol. 3644, pp. 197–206.
- Lin, F.-H., and Mersereau, R. M. (1995). "A constant subjective quality MPEG encoder," *ICASSP* pp. 2177–2180.
- Lubin, J. (1993). "The use of psychophysical data and models in the analysis of display system performance," in *Digital Images and Human Vision*, edited by A. B. Watson, (New York: The MIT Press), pp. 163–178.
- Lubin, J. (1995). "A visual discrimination model for imaging system design and evaluation," in *Visual Models for Target Detection and Recognition*, (World Scientific Publishers, River Edge N.J.).
- Luce, R. D., and Krumhansl, C. L. (1988). "Measurement, scaling, and psychophysics," in *Stevens' handbook of experimental psychology*, edited by R. Atkinson, R. Herrnstein, G. Lindzey, and D. Luce, (Wiley, New York), vol. I, pp. 3–74.
- Martens, J.-B. (1990a). "Application of scale space to image coding," *IEEE Transactions on Communications* **38**, 1585–1591.
- Martens, J.-B. (1990b). "The Hermite transform - theory," *IEEE Transactions on Acoustics, Speech, and Signal Processing* **38**, 1595–1618.
- Martens, J. B. (1999). "Multidimensional modeling of image quality," Internal IPO report **1297**.
- Martens, J. B., and Meesters, L. (1998). "Image dissimilarity," *Signal Processing* **70**, 155–176.
- Klein Teeselink, I., Blommaert, F., and de Ridder, H. (2000). "Image categorization," *Journal of Imaging Science and Technology* **44**, 552–559.

- Meesters, L., and Martens, J. (1999). "Blockiness in JPEG-coded images," Proceedings of the SPIE **3644**, 245–257.
- Mitchell, J. L., Pennebaker, W. B., Fogg, C. E., and LeGall, D. J. (1997). *MPEG video compression standard*, (Chapman and Hall).
- Parducci, and Wedell, D. (1986). "The category effect with rating scales: Number of categories, number of stimuli, and method of presentation." *Journal of Experimental Psychology: Human Perception and Performance* pp. 496–516.
- Pennebaker, W. B., and Mitchell, J. L. (1993). *JPEG Still Image Compression Standard*, (New York: van Nostrand Reinhold).
- Poynton, C. (1993). "Gamma and its disguises: the nonlinear mappings of intensity in perception, CRTs, film and video," *SMPTE journal* **102**, 1099–1108.
- Ramsay, J. (1991). *MULTISCALE Manual*, (McGill University, Montreal).
- Rohaly, A. M., Corriveau, P., Libert, J., Webster, A., Baroncini, V., Beerends, J., Blin, J.-L., Contin, L., Hamada, T., Harrison, D., Hekstra, A., Lubin, J., Nishida, Y., Nishihara, R., Pearson, J., Pessoa, A. F., Pickford, N., Schertz, A., Visca, M., Watson, A., and Winkler, S. (2000a). "Video quality experts group: Current results and future directions," in *Visual Communication and Image Processing 2000*, edited by K. N. Nqan, T. Sikora, and M.-T. Sun, vol. 4067, pp. 742–753.
- Rohaly, A. M., Libert, J., Corriveau, P., and Webster, A., editors (2000b). *Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment*, download from <ftp://ftp.crc.ca/crc/vqeg>.
- Roufs, J. A. J. (1992). "Perceptual image quality: Concept and measurement," *Philips Journal of Research* **47**, 35–62.
- Said, A., and Pearlman, W. (1996). "A new fast and efficient image codec based on set partitioning in hierarchical trees," *IEEE Transactions on Circuits and Systems for Video Technology* **6**, 243–250.
- Stevens, S. S. (1951). "Mathematics, measurement and psychophysics," in *Handbook of experimental psychology*, (Wiley, New York), pp. 1–49.
- Torgerson, W. S. (1958). *Theory and methods of scaling*, (John Wiley & Sons, New York).
- van den Branden Lambrecht, C. J. (1996). *Perceptual Models and Architectures for Video Coding Applications*, Ph.D. thesis, Ecole Polytechnique Federale de Lausanne.
- van Dijk, A., and Martens, J. (1996). "Subjective quality assessment of compressed images," *Signal Processing* **58**, 235–252.
- Wandell, B. A. (1995). *Foundations of vision*, (Sinauer Associates, Inc, Sunderland, Massachusetts).

Bibliography

- Watson, A. (1993). "DcTune: A technique for visual optimization of dct quantization matrices for individual images," Society for Information Display Digest of Technical Papers XXIV pp. 946–949.
- Watson, A. B. (1987). "Efficiency of a model human image code," Journal of the Optical Society of America A **4**, 2401–2417.
- Webster, A., Jones, C., Pinson, M., Voran, S., and Wolf, S. (1993). "An objective video quality assessment system based on human perception," Proceedings of the SPIE **1913**, 15–26.
- Willemsen, M. C. (1997). "Subjective evaluation of jpeg-coded images: quality versus impairment judgements," Tech. Rep. 1171, IPO, Technische Universiteit Eindhoven.
- Winkler, S. (1999). "Issues in vision modeling for perceptual video quality assesement," Signal processing **78**, 231–252.
- Wolf, S., and Webster, A. (1997). "Subjective and objective measures of scene criticality," Contribution to: Turin ITU Experts Meeting on subjective and objective audiovisual quality assessment .
- Yeun, M., and Wu, H. (1998). "A survey of hybrid mc/dpcm/dct video coding distortions," Signal processing **70**, 247–278.
- Yuyama, I., Nishida, Y., and Nakasu, E. (1998). "Objective measurement of compressed digital television picture quality," SMPTE Journal **107**, 348–352.

Appendix A

The following pages show the 164 scenes used in *Chapter 2* and *Chapter 5* to classify instrumental quality measures on the basis of their quality predictions only. These scenes were also used to select a small image set of, e.g. 3 scenes, for subjective testing. The properties of such a stimulus set are that 1) each instrumental quality measure yields different results for the scenes, and that 2) scenes yield different results for instrumental quality measures. The usefulness of instrumental quality measures can then be ascertained from such a small, well-chosen set of images. The same 164 scenes were used to select an appropriate combination rule for the single-ended blockiness measure of *Chapter 4*.



Appendix A









Appendix A



Summary

The main aim of this thesis was to enhance our understanding of how human observers assess image quality across scenes and impairment types and how such judgements and quality predictions can be used to discriminate between instrumental quality measures. The second aim was to develop a single-ended instrumental blockiness measure for sequential baseline JPEG coded images that is robust enough to predict the image quality across scenes.

In *Chapter 2* we propose a procedure to classify instrumental quality measures on the basis of their quality predictions only. We assumed that quality measures predict image quality on a ratio scale and therefore that for each measure, the distances in predicted quality between different scenes are meaningful. Since time-limitations of subjective testing are not an issue in this procedure a large collection of images could be used. A classification on the basis of quality predictions cannot substitute subjective testing, yet it is a method that can be used to compare newly developed measures with existing ones and that is very helpful in visualizing the differences between measures. The procedure was demonstrated by classifying a particular set of 67 double-ended instrumental quality measures. From this set six main clusters of quality measures could be identified that are essentially different in their predictions for an image set containing 164 scenes processed by four methods. In the same chapter it also demonstrated how scenes that discriminate between instrumental quality measures can be selected for subjective tests. The properties of the image set are that 1) each instrumental quality measure yields different results for the scenes, and that 2) the scenes yield different results for the instrumental quality measures.

In *Chapter 3* we showed that subjective quality judgements can be biased by the imposed experimental procedure. The effect of stimulus presentation on the assessed image quality was investigated for a condition in which different classes (e.g. impairment types or scenes) could be identified in a stimulus set. The question was whether human observers link quality judgements across identifiable classes if they are not forced to compare these classes explicitly. The results showed that in stimulus-comparison scaling subjects seemed to use separate quality rating scales if images with different scene content were not compared explicitly. This was tested for a stimulus set containing wavelet-coded images and a stimulus set consisting of JPEG-coded images. Two experiments were also conducted with stimulus sets containing different impairment types introduced by wavelet-coding, DCTune-coding, JPEG and low-pass filtering. In each experiment a single scene was impaired by each of these four processing methods. For one scene the observers seemed to judge the perceived image quality on a single rating scale whereas for the second scene

these findings could not be substantiated. This is most probably due to the fact that in the second scene, the quality-ranges of the distortions were not dissimilar enough. In general, we may conclude that observers use separate quality scales for identifiable classes of stimuli if these are not compared explicitly.

In *Chapter 4* a single-ended blockiness measure was developed for baseline sequential JPEG-coded images. In this model the blockiness strengths are derived directly from the JPEG-coded image, and thus the original is not required. The proposed measure is based on detecting and estimating low-amplitude edges in the horizontal and the vertical directions that are introduced by the JPEG coder. In spite of the fact that several distortions are visible in JPEG-coded images (blockiness, ringing and blurring) it was shown that the strengths of these distortions are linearly related to the perceived image quality. This was substantiated by various psychophysical experiments. The attribute strengths were measured by means of within-scene judgements as well as across-scene judgements. The experiments revealed a linear relationship between the attributes strengths and the perceived image quality. Therefore it is suggested that the single-ended instrumental blockiness measure can also predict the image quality of sequential baseline coded JPEG images.

Finally, in *Chapter 5* we applied the proposed method of *Chapter 2* to classify a set of 67 instrumental quality measures on the basis of their predictions for sequential baseline coded JPEG images. Four instrumental quality measures that predict the perceived image quality differently were selected and their performance was tested by means of within-scene and across-scenes quality judgements. Subjective quality ratings were obtained for two image sets. One set was chosen on the basis of scene content and the other set consisted of scenes that discriminated between the predictions of the quality measures. For most measures the predicted quality of a single scene correlated highly with the perceived image quality. However, it was shown that most instrumental quality measures perform poorly for across-scene quality predictions. The same subjective image quality judgements were used to evaluate the performance of the single-ended blockiness measure. This single-ended measure appeared to be to a lesser degree scene dependent and even seemed to outperform some double-ended measures.

Samenvatting

Het hoofddoel van dit proefschrift is om beter te begrijpen hoe mensen de kwaliteit van beelden beoordelen, tussen scènes en types van verstoringen, en hoe dergelijke beoordelingen en kwaliteitsvoorspellingen gebruikt kunnen worden om onderscheid te maken tussen instrumentele kwaliteitsmaten. Het tweede doel was het ontwikkelen van een *single-ended* instrumentele maat om *blockiness* te voorspellen voor *sequential baseline JPEG* gecodeerde beelden die robust genoeg is om de kwaliteit van beelden tussen scènes te voorspellen.

In hoofdstuk 2 introduceren we een procedure om instrumentele kwaliteitsmaten te klassificeren op basis van hun kwaliteitsvoorspellingen. We stellen dat kwaliteitsmaten de beeldkwaliteit op een ratio schaal voorspellen. Daarom is voor elke maat de afstand in voorspelde kwaliteit tussen verschillende scènes betekenisvol. Aangezien de tijdsbeperking van een subjectieve test niet van belang is voor instrumentele maten, kon in deze procedure een grote verzameling beelden gebruikt worden. Alhoewel de klassificatie op basis van kwaliteitsvoorspellingen een subjectieve test niet kan vervangen is het een methode die gebruikt kan worden om nieuwe maten met bestaande maten te vergelijken en die zeer nuttig is voor het visualiseren van de verschillen tussen de verscheidene maten. De procedure werd gedemonstreerd aan de hand van een specifieke set van 67 *double-ended* kwaliteitsmaten, dwz. een maat waarvoor zowel het gecodeerde beeld als het originele beeld nodig zijn. Binnen deze set konden zes clusters van kwaliteitsmaten geïdentificeerd worden die verschillen in hun voorspellingen voor een beeldset van 164 scènes bewerkt door vier bewerkingsmethoden. In hetzelfde hoofdstuk demonstreren we hoe scènes, voor een subjectieve test, geselecteerd kunnen worden die tussen kwaliteitsmaten discrimineren. De eigenschappen van de set van beelden zijn zodanig dat 1) elke kwaliteitsmaat geeft verschillende resultaten voor de scènes en 2) de scènes geven verschillende resultaten voor de kwaliteitsmaten.

In hoofdstuk 3 laten we zien dat kwaliteitsbeoordelingen kunnen worden beïnvloed door de gebruikte experimentele procedure. Het gevolg van de stimuluspresentatie op de kwaliteitsbeoordelingen werd onderzocht voor een conditie waarbij verschillende klassen (b.v. type van verstoring of scènes) geïdentificeerd kunnen worden in een stimulus set. De vraag was of mensen kwaliteitsbeoordelingen koppelen tussen identificeerbare klassen als ze niet expliciet gedwongen worden om deze klassen te vergelijken. De resultaten laten zien dat in stimulus vergelijkingsschaling proefpersonen een aparte schaal gebruiken als beelden met verschillende scènes niet expliciet vergeleken worden. Dit werd gevonden voor twee verschillende stimulus sets, één met wavelet gecodeerde beelden en één met JPEG gecodeerde beelden. Ook werden er twee experimenten uitgevoerd met stim-

ulus sets die verschillende type van verstoringen bevatten geïntroduceert door wavelet-codering, DCTune, JPEG en low-pass filtering. In elk experiment werd één enkele scène verstoort door één van deze vier verwerkingsmethoden. Voor één scène lijken de proefpersonen de waargenomen beeldkwaliteit op één aparte schaal te beoordelen terwijl voor de tweede scène deze bevindingen niet gestaafd konden worden. Dit is meest waarschijnlijk het gevolg van het feit dat in de tweede scène de beeldkwaliteit, als gevolg van de verstoringen, niet al te verschillend was. In het algemeen kunnen we concluderen dat de proefpersonen aparte kwaliteitsschalen voor identificeerbare klassen van stimuli gebruiken als deze niet expliciet vergeleken worden.

In hoofdstuk 4 is een *single-ended blockiness* maat ontwikkeld voor *sequential baseline JPEG* gecodeerde beelden. In dit model is de *blockiness* sterkte direct uit het JPEG gecodeerde beeld afgeleid en is een origineel beeld niet noodzakelijk. De voorgestelde maat is gebaseerd op het detecteren en schatten van lage amplituderanden in horizontale en verticale richting die geïntroduceerd worden door de JPEG coder. Ondanks het feit dat verschillende verstoringen zichtbaar zijn in JPEG gecodeerde beelden (*blockiness*, *ringing* en *blurring*) werd aangetoond dat de sterkte van deze verstoringen lineair is met de waargenomen beeldkwaliteit. Dit werd onderbouwd door verschillende psychofysische experimenten. De sterkte van de attributen werd gemeten aan de hand van beoordelingen binnen een scène als ook aan de hand van beoordelingen tussen scènes. De experimenten onthullen een lineair verband tussen de sterkte van de attributen en de waargenomen beeldkwaliteit. Daarom wordt er voorgesteld dat de *single-ended blockiness* maat de beeldkwaliteit kan voorspellen van *sequential baseline JPEG* gecodeerde beelden.

Tot slot hebben we in hoofdstuk 5 de in hoofdstuk 2 voorgestelde methode gebruikt om een set van 67 kwaliteitsmaten te klassificeren op basis van hun voorspellingen voor *sequential baseline JPEG* gecodeerde beelden. Vier instrumentele kwaliteitsmaten die de waargenomen beeldkwaliteit verschillend voorspellen werden geselecteerd en hun prestatie werd getest door middel van kwaliteitsbeoordelingen binnen scènes en tussen scènes. Subjectieve kwaliteitsoordelen werden verkregen voor twee sets van beelden: één set werd gekozen op basis van beeldinhoud en de andere set bestond uit scènes die discrimineren tussen de voorspellingen van kwaliteitsmaten. Voor de meeste maten is er een hoge correlatie tussen de voorspelde en de waargenomen kwaliteit van één enkele scène. Echter, er werd aangetoond dat de meeste instrumentele kwaliteitsmaten slecht presteren voor kwaliteitsvoorspellingen tussen scènes. Dezelfde subjectieve beoordeling van beeldkwaliteit werd gebruikt om de prestatie van de *single-ended blockiness* maat te beoordelen. In feite lijkt deze *single-ended* maat in mindere mate afhankelijk van de scène inhoud en lijkt zelfs beter te presteren dan de meeste *double-ended* maten.

Biography

Lydia Meesters was born the 11th of February 1968 in Lemiers, The Netherlands. She attended Atheneum at "Sophianum" in Gulpen. Subsequently, she studied computer science at the University of Nijmegen (KUN), and completed her M.Sc. thesis on surface reconstruction of medical 3D-voxel data at the University of Aveiro in Portugal. In 1994 she joined Arcobel Graphics BV, in 's Hertogenbosch, The Netherlands and was employed as a software engineer in the field of computer graphics and image processing. In 1996 she started her Ph.D. research at the vision group at the Institute for Perception Research (IPO). The research as described in this thesis was supported by the ACTS AC055 project 'Tapestries'.

