MODELLING PERCEPTUAL QUALITY AND VISUAL SALIENCY FOR IMAGE AND VIDEO COMMUNICATIONS

Ulrich Engelke

Blekinge Institute of Technology
Doctoral Dissertation Series No. 2010:06

School of Engineering

Modelling Perceptual Quality and Visual Saliency for Image and Video Communications

Ulrich Engelke

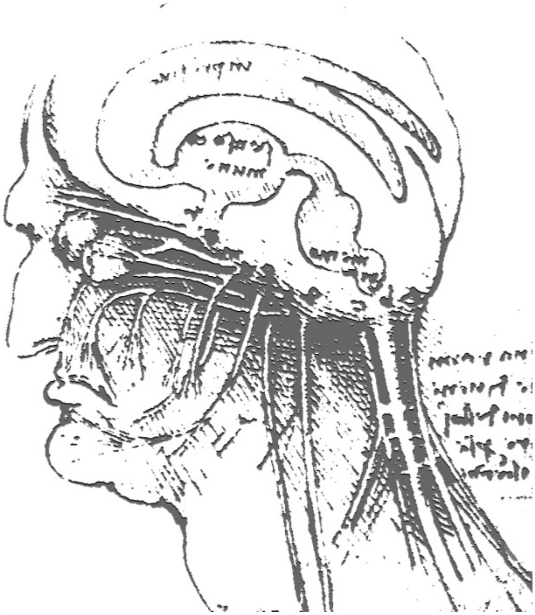
Blekinge Institute of Technology Doctoral Dissertation Series
No 2010:06

Modelling Perceptual Quality and Visual Saliency for Image and Video Communications

Ulrich Engelke

Department of Electrical Engineering
School of Engineering
Blekinge Institute of Technology
SWEDEN

Department of Electrical Engineering
School of Engineering
Publisher: Blekinge Institute of Technology
Printed by Printfabriken, Karlskrona, Sweden 2010
ISBN: 978-91-7295-185-3
Blekinge Institute of Technology Doctoral Dissertation Series
ISSN 1653-2090
urn:nbn:se:bth-00470



Although nature commences with reason and ends in experience it is necessary for us to do the opposite, that is to commence with experience and from this to proceed to investigate the reason.

Leonardo da Vinci

Abstract

The evolution of advanced radio transmission technologies for third and future generation mobile radio systems has paved the way for the delivery of mobile multimedia services. This is further enabled through contemporary video coding standards, such as H.264/AVC, allowing wireless image and video applications to become a reality on modern mobile devices. The extensive amount of data needed to represent the visual content and the scarce channel bandwidth constitute great challenges for network operators to deliver an intended quality of service. Appropriate metrics are thus instrumental for service providers to monitor the quality as experienced by the end user. This thesis focuses on subjective and objective assessment methods of perceived visual quality in image and video communication. The content of the thesis can be broadly divided into four parts.

Firstly, the focus is on the development of image quality metrics that predict perceived quality degradations due to transmission errors. The metrics follow the reduced-reference approach, thus, allowing to measure quality loss during image communication with only little overhead as side information. The metrics are designed and validated using subjective quality ratings from two experiments. The distortion assessment performance is further demonstrated through an application for filter design.

The second part of the thesis then investigates various methodologies to further improve the quality prediction performance of the metrics. In this respect, several properties of the human visual system are investigated and incorporated into the metric design. It is shown that the quality prediction performance can be considerably improved using these methodologies.

The third part is devoted to analysing the impact of the complex distortion patterns on the overall perceived quality, following two goals. Firstly, the confidence of human observers is analysed to identify the difficulties during assessment of the distorted images, showing, that indeed the level of confidence is highly dependent on the level of visual quality. Secondly, the impact of content saliency on the perceived quality is identified using region-of-interest selections and eye tracking data from two independent subjective experiments. It is revealed, that the saliency of the distortion region indeed has an impact on the overall quality perception and also on the viewing behaviour of human observers when rating image quality.

Finally, the quality perception of H.264/AVC coded video containing packet loss is analysed based on the results of a combined subjective video quality and eye tracking experiment. It is shown that the distortion location in relation to the content saliency has a tremendous impact on the overall perceived quality. Based on these findings, a framework for saliency aware video quality assessment is proposed that strongly improves the quality prediction performance of existing video quality metrics.

Preface

This Ph.D. thesis reports about my work within the field of perceptual quality metric design and visual saliency modelling for image and video communications. The research has been conducted at the School of Engineering at the Blekinge Tekniska Högskola (BTH), Karlskrona, Sweden.

Parts of the work have been conducted during two independent research visits at international universities. The first visit of about two months duration took place at the School of Computing and Mathematics at the University of Western Sydney, Sydney, Australia. The second visit of about three months duration was conducted at the Image and Video Communication Department at the University of Nantes, Nantes, France. Full funding for both visits has been awarded by BTH.

The majority of research results that are summarised within this thesis have previously been reported in international journals and conference proceedings. Furthermore, parts of the work have been reported in a Licentiate thesis entitled "Perceptual Quality Metric Design for Wireless Image and Video Communication", also published at BTH.

Acknowledgements

My journey towards the Ph.D. degree would not have been possible without the help of many people. It is my great pleasure to take this opportunity to thank them for the support and advice that I received over the past years.

First of all, I would like to express my deepest gratitude towards Prof. Dr.-Ing. Hans-Jürgen Zepernick for offering me the opportunity to follow him from 'Down Under' into the southern rims of Sweden to pursue my doctoral studies under his supervision. I always admired his ability of having a professional work attitude while perpetually being a considerate and amenable advisor. I am particularly thankful to him for not restricting my research education into predefined paths but instead giving me the freedom to develop my research interests along the way. I would further like to thank my co-supervisor Dr. Markus Fiedler for his encouragement and support over the years and Dr. Maulana Kusuma for the mentoring I received in the early stages of my Ph.D. studies.

My professional development has moreover been considerably influenced by highly rewarding international cooperations. My sincere gratitude goes to Prof. Anthony Maeder for being an outstanding host during my stay at the University of Western Sydney in Australia. His extensive knowledge of human visual perception, that he communicated to me in our long discussions, broadened my mind and has been a great source of inspiration. I would also like to thank Dr. Clinton Fookes from the Queensland University of Technology, Australia, for his 'remote' support with the eye tracking experiments we conducted.

I also had the pleasure to spend some time in the beautiful city of Nantes in France, working with Prof. Patrick Le Callet from the University of Nantes. I am very grateful to him for inviting me to conduct research in a highly competent, inspiring, and welcoming team. He has been an excellent mentor and host, making my stay at the department a highly fruitful and memorable experience. Special thanks also go to Assoc. Prof. Marcus Barkowsky for his unreserved support, both at work and during his spare time. My thanks are further extended to Dr. Fadi Boulos and Romuald Pepion for their help with creating the test sequences and with conducting the subjective experiment.

My great appreciation goes as well to my other collaborators and friends, Dr. Shelley Buchinger from the University of Vienna, Austria, Francesca de Simone from the Ecole Polytechnique Fédérale de Lausanne, Switzerland, Hagen Kaprykowsky from the Heinrich Hertz Institute in Berlin, Germany, Hantao Liu from the Delft University of Technology, The Netherlands, and Andreas Rossholm from ST-Ericsson, Sweden.

Furthermore, I am very honoured to have such highly renowned and competent experts from around the world in my Ph.D. committee. In particular, I am grateful to Prof. Yao Wang from the Polytechnic Institute of New York University, USA, for being my opponent in the Ph.D. defense. Not less appreciation goes to my committee members, Prof. Damon Chandler from Oklahoma State University, USA, Prof. Helmut Hlavacs from University of Vienna, Austria, Dr. Kjell Brunnström from Acreo AB, Sweden, and Prof. Bo Schenkman from Kungliga Tekniska Högskola (KTH) and Blekinge Tekniska Högskola (BTH), Sweden.

Thanks goes out to the Graduate School of Telecommunications administered by KTH, Stockholm, Sweden, for partly funding my thesis work and to the European Networks of Excellence EuroNGI, EuroFGI, and EuroNF, for funding my attendance to meetings and Ph.D. courses.

Special thanks also goes to my colleagues and friends at the department who have made working at BTH and living in Sweden a joyous experience. I will always look back at the wonderful things we did together and to the many fun parties and BBQs we had. I would like to especially thank those people at the department that always kept the wheels turning. I am particularly thinking of Madeleine Jarlten, Lena Brandt Gustafsson, Marie Ahlgren, and Mansour Mojtahedi. Their willingness to always lend a helping hand is too often taken for granted.

As my work is essentially based on data collected from many human subjects, I am highly grateful to all the participants from Sweden, Australia, and France, for sparing their valuable time to help us out with the experiments.

On a more personal note, I would like to thank all my friends who shared the past years here in Sweden with me. This applies particularly to my dear friends Fredrik and Karoline who helped me to get settled and who were substantially involved in creating unforgettable memories of my stay in Sweden.

Even though my parents Rainer and Annelie never had the privilege of moving or studying abroad, their support for me has always been undoubted. No road was too long, no holiday too valuable, and no couch too heavy to get their son to wherever necessary. Thank you mum and dad for always being there for me.

Without the unconfined support of one special person I would not be here in Sweden today, my wife Melissa. Years ago, when residing in Australia I decided to spend my life with her and not leave her behind for any job in the world. This was put to the test when I received the offer from BTH. However, three simple words of hers lead us on an unexpected road to travel: "Go for it!"... and so we did. Her continuous encouragement, her endless love, and her careful proof-reading of my thesis helped me to get where I am today. Thank you so much Schatzi.

Ulrich Engelke
Karlskrona, September 2010

Publication List

Thesis:

U. Engelke, "Perceptual Quality Metric Design for Wireless Image and Video Communication," *Licentiate Thesis at Blekinge Institute of Technology*, ISSN: 1650-2140, ISBN: 978-91-7295-144-0, Ronneby, Sweden, June 2008.

Journal articles:

U. Engelke, T. M. Kusuma, H.-J. Zepernick, and M. Caldera "Reduced-Reference Metric Design for Objective Perceptual Quality Assessment in Wireless Imaging," *Signal Processing: Image Communication*, vol. 24, no. 7, pp. 525-547, 2009.

U. Engelke and H.-J. Zepernick "A Framework for Optimal Region-of-Interest Based Quality Assessment in Wireless Imaging," *Journal of Electronic Imaging, Special Section on Image Quality*, vol. 19, no. 1, ID 011005, 2010.

Peer reviewed conference papers:

U. Engelke, H.-J. Zepernick, and A. J. Maeder, "Visual Fixation Patterns in Subjective Quality Assessment: Analysing the Relative Impact of Natural Image Content and Structural Distortions," *in Proc. of IEEE International Symposium on Intelligent Signal Processing and Communications Systems (ISPACS)*, Cheng Du, China, December 2010. *Invited paper in Special Session on 'Quality of Multimedia Experience in Signal Processing'.*

U. Engelke, A. J. Maeder, and H.-J. Zepernick, "Analysing Inter-Observer Saliency Variations in Task-Free Viewing of Natural Images," *in Proc. of IEEE International Conference on Image Processing (ICIP)*, Hong Kong, China, September 2010.

U. Engelke, R. Pepion, P. Le Callet, and H.-J. Zepernick "Linking Distortion Perception and Visual Saliency in H.264/AVC Coded Video Containing Packet Loss," *in Proc. of SPIE/IEEE International Conference on Visual Communications and Image Processing (VCIP)*, Huang Shan, China, July 2010. *Invited paper in Special Session on 'Perception Based Visual Signal Analysis and Representation'.*

U. Engelke, M. Barkowsky, P. Le Callet, and H.-J. Zepernick "Modelling Saliency Awareness for Objective Video Quality Assessment," *in Proc. of International Workshop on Quality of Multimedia Experience (QoMEX)*, pp. 212-217, Trondheim, Norway, June 2010.

U. Engelke, H.-J. Zepernick, and T. M. Kusuma "Subjective Quality Assessment for Wireless Image Communication: The Wireless Imaging Quality Database," *in Proc. of International Workshop on Video Processing and Quality Metrics (VPQM)*, Scottsdale, USA, January 2010.

U. Engelke, A. J. Maeder, and H.-J. Zepernick, "Visual Attention Modelling for Subjective Image Quality Databases," *in Proc. of IEEE International Workshop on Multimedia Signal Processing (MMSP)*, Rio de Janeiro, Brazil, October 2009.

U. Engelke, A. J. Maeder, and H.-J. Zepernick, "On Confidence and Response Times of Human Observers in Subjective Image Quality Assessment," *in Proc. of IEEE International Conference on Multimedia and Expo (ICME)*, pp. 910-913, New York City, USA, June 2009.

U. Engelke, H.-J. Zepernick, and A. J. Maeder, "Visual Attention Modeling: Regions-of-Interest Versus Fixation Patterns," *in Proc. of IEEE Picture Coding Symposium (PCS)*, Chicago, USA, May 2009. *Invited paper in Special Session on 'Visual Attention, Artistic Intent, and Efficient Coding'.*

U. Engelke and H.-J. Zepernick, "Optimal Region-of-Interest Based Visual Quality Assessment," *in Proc. of IS&T/SPIE Human Vision and Electronic Imaging XIV*, vol. 7240, San Jose, USA, January 2009.

U. Engelke and H.-J. Zepernick, "Pareto Optimal Weighting of Structural Impairments for Wireless Imaging Quality Assessment," *in Proc. of IEEE International Conference on Image Processing (ICIP)*, pp. 373-376, San Diego, USA, October 2008.

U. Engelke and H.-J. Zepernick, "Multiobjective Optimization of Multiple Scale Visual Quality Processing," *in Proc. of IEEE International Workshop on Multimedia Signal Processing (MMSP)*, pp. 212-217, Cairns, Australia, October 2008.

U. Engelke, X. N. Vuong, and H.-J. Zepernick, "Regional Attention to Structural Degradations for Perceptual Image Quality Metric Design," *in Proc. of IEEE In-*

ternational Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 869-872, Las Vegas, USA, April 2008.

U. Engelke and H.-J. Zepernick, "Multi-resolution Structural Degradation Metrics for Perceptual Image Quality Assessment," *in Proc. of Picture Coding Symposium (PCS)*, Lisbon, Portugal, November 2007.

U. Engelke and H.-J. Zepernick, "Perceptual-based Quality Metrics for Image and Video Services: A Survey," *in Proc. of International Conference on Next Generation Internet Networks Design and Engineering Heterogeneity (NGI)*, pp. 190-197, Trondheim, Norway, May 2007.

U. Engelke and H.-J. Zepernick, "An Artificial Neural Network for Quality Assessment in Wireless Imaging Based on Extraction of Structural Information," *in Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1249-1252, Honolulu, USA, April 2007.

U. Engelke, A. Rossholm, H.-J. Zepernick, and B. Lövström, "Quality Assessment of an Adaptive Filter for Artifact Reduction in Mobile Video Sequences," *in Proc. of IEEE International Symposium on Wireless Pervasive Computing (ISWPC)*, pp. 360-366, San Juan, Puerto Rico, February 2007.

U. Engelke and H.-J. Zepernick, "Quality Evaluation in Wireless Imaging Using Feature-Based Objective Metrics," *in Proc. of IEEE International Symposium on Wireless Pervasive Computing (ISWPC)*, pp. 367-372, San Juan, Puerto Rico, February 2007.

U. Engelke, H.-J. Zepernick, and T. M. Kusuma, "Perceptual Evaluation of Motion JPEG2000 Quality over Wireless Channels," *in Proc. of IEEE Symposium on Trends in Communications (SympoTIC)*, pp. 92-96, Bratislava, Slovakia, June 2006.

U. Engelke, T. M. Kusuma, and H.-J. Zepernick, "Perceptual Quality Assessment of Wireless Video Applications," *in Proc. of International ITG-Conference on Source and Channel Coding (SCC)*, Munich, Germany, April 2006.

Other publications in conjunction with this thesis:

U. Engelke, A. J. Maeder, and H.-J. Zepernick, “The Effect of Spatial Distortion Distributions on Human Viewing Behaviour when Judging Image Quality,” *in Proc. of European Conference on Visual Perception (ECVP)*, pp. 22, Regensburg, Germany, August 2009.

U. Engelke and H.-J. Zepernick, “Perceptual Quality Measures for Image and Video Services,” *in Proc. of Euro-NGI Workshop on Socio-Economic Aspects of Next Generation Internet*, pp. 15-19, Lyngby, Denmark, October 2006.

Co-authored publications:

T. Q. Duong, H.-J. Zepernick, and U. Engelke, “Cooperative Wireless Communications with Unequal Error Protection and Fixed Decode-and-Forward Relays,” *in Proc. of IEEE International Conference on Communications and Electronics (ICCE)*, Nha Trang, Vietnam, August 2010.

M. I. Iqbal, H.-J. Zepernick, and U. Engelke, “Perceptual-based Quality Assessment of Error Protection Schemes for Wireless JPEG2000,” *in Proc. of IEEE International Symposium on Wireless Communication Systems (ISWCS)*, pp. 348-352, Siena, Italy, September 2009.

T. Q. Duong, U. Engelke and H.-J. Zepernick, “Unequal Error Protection for Wireless Multimedia Transmission in Decode-and-Forward Relay Networks,” *in Proc. of IEEE Radio and Wireless Symposium (RWS)*, pp. 703-706, San Diego, USA, January 2009.

M. I. Iqbal, H.-J. Zepernick, and U. Engelke, “Error Sensitivity Analysis for Wireless JPEG2000 Using Perceptual Quality Metrics,” *in Proc. of International Conference on Signal Processing and Communication Systems (ICSPCS)*, pp. 1-9, Gold Coast, Australia, December 2008.

H.-J. Zepernick and U. Engelke, “On Perceptual Quality Evaluation of Video Applications for Wireless Ad-Hoc Networks,” *in Proc. of Scandinavian Workshop on Wireless Ad-Hoc Networks*, Stockholm, Sweden, May 2007.

Contents

Acronyms

ACJ	Adjectival Categorical Judgement
ACR	Absolute Category Rating
ANN	Artificial Neural Network
ANOVA	Analysis of Variance
AUC	Area Under the ROC Curve
AVC	Advanced Video Coding
AWGN	Additive White Gaussian Noise
BCH	Bose-Chaudhuri-Hocquenghem
BER	Bit Error Rate
BG	Background
BIT	Blekinge Institute of Technology
BPSK	Binary Phase Shift Keying
CI	Confidence Interval
CRT	Cathode Ray Tube
CS	Confidence Score
CV	Cross Validation
dB	Decibel
DCR	Degradation Category Rating
DCT	Discrete Cosine Transform
DF	Discarded Feature
DL	Discarded Level
DMOS	Differential Mean Opinion Score
DoF	Degrees of Freedom
DPCM	Differential Pulse Code Modulation
DSCQS	Double Stimulus Continuous Quality Scale
DSIS	Double Stimulus Impairment Scale
dva	Degrees of visual angle
DWT	Discrete Wavelet Transform
E1	Experiment 1
E2	Experiment 2
E3	Experiment 3
E4a	Experiment 4a

E4b	Experiment 4b
E5	Experiment 5
EBP	Error Backpropagation
FFNN	Feed-Forward Neural Network
FN	False Negative
FoA	Focus of Attention
FP	False Positive
FPR	False Positive Rate
fps	Frames per second
FR	Full-Reference
GDA	Gradient Descent Algorithm
GNA	Gauss-Newton Algorithm
GOP	Group of Pictures
GP	Gaze Point
GSM	Gaussian Scale Mixtures
HD	High Definition
HIQM	Hybrid Image Quality Metric
HM	Heat Map
HVS	Human Visual System
IA	Image Activity
IAM	Image Activity Measure
IEC	International Electrotechnical Commission
IQM	Image Quality Metric
IRCCyN	Institut de Recherche en Communications et en Cybernétique
ISO	International Organization for Standardization
ITU	International Telecommunication Union
ITU-R	ITU Radiocommunication Sector
ITU-T	ITU Telecommunication Sector
IVC	Image and Video Communication
JND	Just Noticeable Difference
JPEG	Joint Photographic Experts Group
JVT	Joint Video Team

KLD	Kullback-Leibler Distance
LIVE	Laboratory for Image and Video Engineering
LMA	Levenberg-Marquardt Algorithm
MB	Macro Block
MCS	Mean Confidence Score
MICT	Media Information and Communication Technology
MOO	Multiobjective Optimisation
MOS	Mean Opinion Score
MPEG	Moving Picture Experts Group
MQS	Mean Quality Score
MRT	Mean Response Time
MS	Mean Squares
MSE	Mean Squared Error
MSFQM	Multiple-Scale Feature-Based Quality Metric
NHIQM	Normalised Hybrid Image Quality Metric
NR	No-Reference
OR	Outlier Ratio
PSNR	Peak Signal-to-Noise Ratio
QCIF	Quarter Common Intermediate Format
QM	Quality Metric
QOE	Quality of Experience
QOS	Quality of Service
QP	Quantisation Parameter
QS	Quality Score
RMSE	Root Mean Squared Error
ROC	Receiver Operating Characteristic
ROI	Region-of-Interest
RR	Reduced-Reference
RRIQA	Reduced-Reference Image Quality Assessment
RT	Response Time

SD	Standard Definition
SE	Standard Error
SI	Spatial Information
SM	Saliency Map
SMI	SensoMotoric Instruments
SNR	Signal-to-Noise Ratio
SS	Sum of Squares
SSCQE	Single Stimulus Continuous Quality Evaluation
SSE	Sum of Squared Errors
SSIM	Structural Similarity
TetraVQM	Temporal Trajectory Aware Video Quality Measure
TI	Temporal Information
TN	True Negative
TP	True Positive
TPR	True Positive Rate
UWS	University of Western Sydney
VA	Visual Attention
VAIQ	Visual Attention for Image Quality
VFP	Visual Fixation Pattern
VIF	Visual Information Fidelity
VQEG	Video Quality Experts Group
VQM	Video Quality Metric
VSNR	Visual Signal-to-Noise Ratio
WATRI	Western Australian Telecommunications Research Institute
WIQ	Wireless Imaging Quality

1 Introduction

The human visual system (HVS) is often considered to be the most prominent of our sense organs to obtain information from the outside world [1]. Without our sight we would live in darkness and we would not be able to appreciate the beauty of the world around us. During all phases of human evolution our eyes were adapted to observing a natural environment. This has changed only in recent decades with the deployment of many visual technologies, such as television, cinema, computer screens, and most recently mobile phones. These ubiquitous technologies now strongly influence our everyday work and private life and many people, especially of the younger generation, have difficulties imagining a time before these technologies were available. Hence, we are getting more and more used to not just looking at the natural environment around us, but rather at artificial reproductions of it, in terms of digital images and videos. This is especially enabled through recent advances in communication technologies, such as the Internet and third generation mobile radio networks, which allow distribution and sharing of visual content in an ubiquitous manner.

The range of image and video processing systems that facilitate visual reproductions of the real world is broad and includes image and video acquisition, compression, enhancement, and communication systems [2]. These systems are usually designed based on a compromise between technical resources and the visual quality of the output. Since we are accustomed to impeccable quality of the real world environment, we are biased to expect also a certain degree of quality from its digital visual representations. However, the quality is often reduced due to many influencing factors, including, capture, compression, transmission, and display of the image or video. These processes potentially introduce distortions into the visual content resulting in a reduction of perceived quality. This is often due to the naturalness of the visual scene being impaired, meaning, that structures are changed or introduced that are not observed when looking at a real world environment. The degradation in quality depends highly on the type and severeness of the artifact introduced by the different processing steps.

Visual content and service providers are thus particularly interested in measuring the quality loss introduced in any of the processing steps involved, which is instrumental for guaranteeing a certain level of visual experience to the observer. This is especially crucial for wireless network providers [3], as the wireless channel constitutes an unreliable and unpredictable medium that can cause severe degradations to the transmitted signal. The scarce bandwidth of the wireless channel in conjunction with the large amount of image and video data comprise a highly complex and intricate scenario. Thus, the deployment of wireless image and video communication services is considerably more difficult, compared to the traditional voice services, for which reliable communication networks have been in place for many years.

One of the major challenges in communication systems, and in particular wireless services, is therefore the design of networks that fulfill the stringent Quality of Service (QoS) requirements of wireless image and video applications to guarantee a certain Quality of Experience (QoE) to the end-user [4–6]. In order to monitor the quality of the wireless communication services, appropriate metrics are needed that are able to accurately quantify the end-to-end visual quality as perceived by the user. The resulting metrics can then be utilised to perform efficient link adaptation and resource management techniques to fulfill the stringent QoS requirements. Traditional link layer metrics, such as signal-to-noise ratio (SNR) and bit error rate (BER), have been widely used to perform this task but were found to not suitably reflect the subjectively perceived quality [7], as the impact of transmission errors on the visual signal may vary drastically depending on the location of the errors in the bit stream.

Considering the above, new paradigms in quality metric design for wireless image and video communications need to be established [8, 9]. The aim of this thesis is to contribute to this goal by developing perceptual quality metrics that are able to accurately quantify end-to-end visual quality of wireless image and video communication services. In comparison to quality assessment for applications such as compression, the communications context represents a considerably more difficult task, which is mainly due to three reasons. Firstly, the computational complexity of the quality metrics needs to be low as the processing power in mobile devices is usually limited, as compared to, for instance, desktop computers. Secondly, the original image or video is typically not available at the receiver where the quality assessment takes place. As such, the quality assessment needs to be conducted on either, just the received image/video, or based on some additional side information from the original image/video that is sent over the channel. Lastly, the distortion patterns caused by transmission errors can be highly complex with respect to the artifact types that they contain, their distributions, and their strengths, thus, drastically complicating the quality assessment prediction as compared to the usually more uniform and globally distributed source coding distortions.

The complex distortion patterns give also rise to another phenomenon that we investigate in this thesis, namely, visual attention to the distortions and their interaction with the visual content. The motivation being, that localised distortions may have a larger impact on the overall perceived quality of a visual scene if they appear in a perceptually interesting or important region. On the contrary, distortions appearing in a region that observers find of low interest may not impact as severely on the perceived quality degradations. For this reason, we also set our focus in this thesis on the effects of visual attention and their benefits for visual quality assessment.

This introduction serves to provide the reader with the necessary background to follow the work that has been conducted in this thesis. Each of the topics discussed could fill entire books and in order to not burst the scope of the thesis, we are forced to limit our focus on the information that is relevant in the context of this work. In Section 1.1, we motivate the need for perceptual quality metrics by highlighting the drawbacks of conventional image metrics. In Sections 1.2, we then discuss subjective visual quality assessment methods and previous work conducted in this field. A classification of objective visual quality assessment methods is given in Section 1.3 followed by a survey of visual quality metrics in Section 1.4. In Section 1.5, a brief introduction to visual attention is given and the potential benefits for visual quality assessment are discussed. In Section 1.6, we then discuss visual quality assessment in the context of image and video communications and define the framework that is considered in the scope of this thesis. The introduction is concluded in Section 1.7 with a summary of contributions and an overview of the thesis.

1.1 The downside of conventional image metrics

With the increasing appearance of digital visual media, the growing need for objective quality assessment that correlates well with subjectively perceived quality has been recognised as an instrumental tool for system design and optimisation. Especially in recent years, the efforts in visual quality assessment have increased considerably, leading to a number of quality metrics that have been proposed in the literature. However, this research field is considered to be still immature, as there are no widely accepted image quality metrics (IQM) and video quality metrics (VQM) that work well under a wide range of different conditions [10]. On the contrary, in the fields of speech and audio there are two standardised and widely accepted methods, the Perceptual Evaluation of Speech Quality (PESQ) [11] and the Perceptual Evaluation of Audio Quality (PEAQ) [12], respectively. One reason for this might be that the HVS, and the higher level cognitive visual data processing, is to a great part not fully understood yet and thus cannot easily be emulated by an objective algorithm. Thus, the traditional fidelity metrics such as the mean squared error (MSE) and the related peak signal-to-noise ratio (PSNR) are still predominantly used for monitoring system performance and for system optimisation. With the advances in perceptual quality assessment, however, the acceptance of visual quality metrics as an alternative to PSNR is slowly becoming a reality.

To fully understand the benefits of perceptual quality metrics, it is conducive to investigate the properties of the traditionally used metrics, such as PSNR, and identify their shortcomings in relation to prediction of perceived visual quality. In the following, we provide a short discussion, emphasising why PSNR is generally not suitable as a perceptual quality metric.

Images and videos are presented on a digital device in a pixel-based fashion, where each pixel is represented by a luminance value and corresponding chrominance values. Unless the resolution of the visual representation is really coarse (which is nowadays rarely the case), the HVS does not recognise the pixels as single entities but rather perceives structures and objects in the scene that are composed of the pixels. This does not only apply for the visual content of the scene but also for potential distortions that are introduced. For this reason, perceptual quality metrics should not aim on quantifying the perceived annoyance of visual distortions on a pixel-by-pixel basis, as this does not represent the way the HVS works. The widely used PSNR, however, assesses the fidelity between two images $I_1(x,y)$ and $I_2(x,y)$ on a pixel-by-pixel basis as

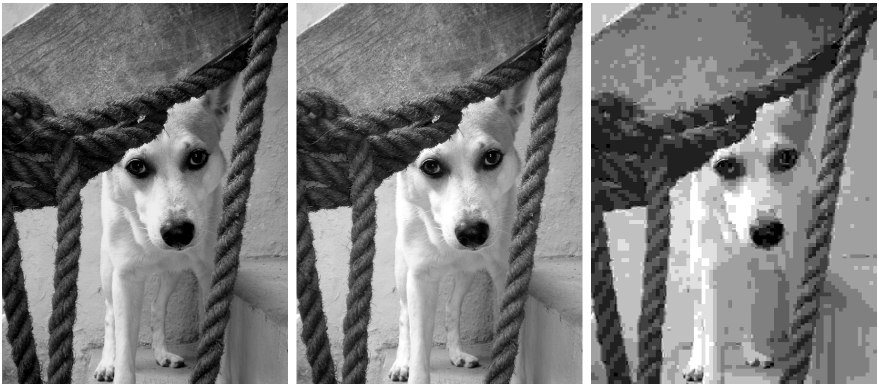
$$\text{PSNR} = 10 \log \frac{\eta^2}{\text{MSE}} \tag{1}$$

where η is the maximum pixel value, typically 255. The MSE is given as

$$\text{MSE} = \frac{1}{XY} \sum_{x=1}^{X} \sum_{y=1}^{Y} [I_1(x,y) - I_2(x,y)]^2 \tag{2}$$

with X and Y denoting the horizontal and vertical image dimensions, respectively. The simple, pixel-based difference calculation is computationally very efficient, however, it is also the main reason why PSNR and MSE exhibit in many cases a poor correlation with perceived visual quality. This is to say that PSNR does not generally perform badly, which is why it has thus far been so widely used, especially in the image and video coding community [13]. However, there are certain circumstances where PSNR fails heavily as a quality metric, which is illustrated by the following two examples.

The images in Fig. 1 show the undistorted reference image 'Boris' in the middle and two processed versions of the same image, one on either side. The image to the left has been subjected to an intensity shift, where each pixel has been darkened slightly. Clearly, this processing step does not impact much, if at



Intensity shift Reference JPEG compression

Figure 1: Reference image 'Boris' and two processed versions of it.

Table 1: Image quality metrics for image 'Boris'.

Artifact	PSNR [dB]	Δ_{NHIQM}	MOS$_{NHIQM}$
Intensity shift	23.55	0.002	99.598
JPEG compression	23.952	0.805	13.547

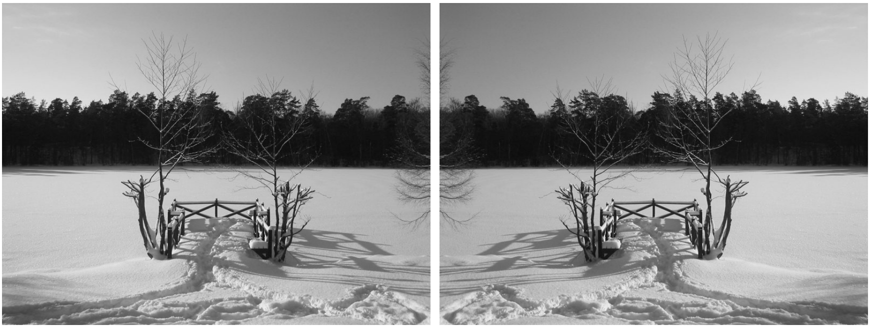
all, on the perceived quality of the image. The image to the right, on the other hand, has been compressed using the Joint Photographic Experts Group (JPEG) encoder at a compression ratio of about 0.06. The resulting distortions in terms of strong blocking artifacts are clearly visible in the image. When comparing the two processed images to the left and to the right, it is apparent that the quality loss due to the JPEG coding is substantially larger in comparison to the quality loss due to the intensity shift.

The PSNR values computed between the reference image and the respective processed images as presented in Fig. 1 are shown in Table 1. The PSNR metric is measured in decibels (dB) with a higher value indicating higher similarity between two images. It can be seen, that the PSNR values of the two processed images are almost the same, indicating that the differences between the reference image and the respective processed images are nearly the same. In fact, the slightly higher PSNR value for the JPEG coded image even suggests, that this image is more similar to the reference image than the intensity shifted image. This is obviously not the case from a perceptual point of view. The large discrepancy between the PSNR values and the perceived quality loss can be attributed to the difference in nature of the two distortions. The intensity shift did not change any of the structural properties of the image whereas the blocking artifacts, caused by the JPEG encoding, introduced highly unnatural artifacts that strongly impair the structure of the underlying image content and result in loss of spatial information [14, 15]. Due to the pixel-based analysis, this structural change is not accounted for by PSNR.

In addition to the PSNR metric in Table 1, we also provide values for the difference of the Normalised Hybrid Image Quality Metric (NHIQM) between the two images, Δ_{NHIQM}, and its related predicted mean opinion score, MOS$_{NHIQM}$. We designed this metric [16] to capture structural distortions in image and video content, in particular in the context of transmission errors. A larger Δ_{NHIQM} value indicates stronger structural differences between the images, whereas a larger MOS$_{NHIQM}$, on a scale from 0 to 100, represents better perceived quality of the processed image. It can be observed that both Δ_{NHIQM} and MOS$_{NHIQM}$ are able to distinguish well between the different levels of perceptual quality of the processed images in relation to the reference image. The metric will be explained in detail in Chapter 3.

Another simple example, highlighting the inapplicability of PSNR as a quality metric, is given with respect to the images in Fig. 2 and the corresponding metric values in Table 2. The image to the left is a visually lossless compressed version of the reference image 'Trollsjö'. The image to the right is a horizontally mirrored version of the image to the left. The process of mirroring the image obviously does not impair the perceived quality whatsoever. However, when consulting the PSNR values in Table 2 it can be observed, that the metric is much lower for the mirrored image as compared to the image with normal orientation. As with the earlier example, this is a deficit that can be attributed to the pixel-based comparison between the images, not taking into account the underlying structure of the visual content. The perceptual quality metric, Δ_{NHIQM}, and the predicted mean opinion score, MOS$_{NHIQM}$, on the other hand are largely unaffected by the mirroring of the image and predict the same perceived quality for both images.

The above examples highlight a few problems that PSNR and other pixel-based metrics experience. As a result of neglecting the visual content and the different distortion types that can occur, the pixel-based metrics generally perform poorly when quality is assessed across different visual content and across different distortion types [17]. For these reasons, pixel-based metrics usually disqualify for



Normal orientation

Horizontally mirrored

Figure 2: A perceptually lossless coded version of the image 'Trollsjö' of normal orientation and a horizontally mirrored version of it.

Table 2: Image quality metrics for image 'Trollsjö'.

Artifact	PSNR [dB]	Δ_{NHIQM}	MOS_{NHIQM}
Normal orientation	47.156	0.004	98.992
Horizontally mirrored	13.713	0.004	98.981

perceptual assessment of image and video quality.

1.2 Subjective visual quality assessment

The simple, pixel-based metrics discussed above can generally be considered as kind of a 'worst case' scenario in relation to prediction of perceived visual quality. On a scale measuring the performance in predicting perceptual quality, these metrics therefore represent the lower end. On the other hand, human observers are generally considered to be the best judges of visual quality and subjective assessment methods are considered to be the most reliable measures of perceived visual quality [18]. Subjective assessment methods are thus often considered as a 'ground truth' for quality prediction and hence, form the upper end of a quality prediction performance scale. The aim of objective visual quality measures is then to be as close as possible to the upper end of the scale, thus reflecting well the quality perception of a human observer.

For IQM and VQM to predict perceived visual quality well, subjective quality ratings are thus needed for the metric design and validation. These are usually obtained by conducting image and video quality experiments, involving a number of human observers that rate the quality of the stimuli presented to them. The resulting mean opinion scores (MOS), as an average over all observers, then constitute a subjective measure of perceived visual quality. There are several international standards that specify in detail the procedures for subjective image and video quality experiments, that should be followed to obtain valid outcomes in terms of MOS.

1.2.1 Subjective testing standards

Two of the most widely used standards are specified by the International Telecommunication Union (ITU). The Radiocommunications sector of the ITU (ITU-R) specifies procedures for television pictures in Rec. BT.500-11 [19] including both single and double stimulus methods. In the single stimulus continuous quality evaluation (SSCQE) method, the quality of the distorted stimulus is rated without any reference to the original stimulus. On the other hand, both the reference and the distorted stimuli are rated using the double stimulus continuous quality scale (DSCQS). Similarly, procedures for multimedia applications are defined in Rec. P.910 [20] by the Telecommunications sector of the ITU (ITU-T), including an absolute category rating (ACR) for single stimulus assessment and the degradation category rating (DCR) for double stimulus assessment.

It is because of these specific procedures that subjective quality experiments are widely accepted measures of perceptual quality. However, these procedures also require a careful design process, which makes subjective experiments usually tedious and time consuming. Therefore, subjective experiments are usually not feasible for deployment in most real world applications, such as quality monitoring in a video broadcasting scenario. The results of subjective quality experiments in terms of MOS are instrumental though, for the design and validation of perceptual IQM and VQM. In addition, they provide valuable insight into human visual perception of natural image and video content in the presence and absence of distortions.

1.2.2 Subjective visual quality studies

With the turn of the century it has been increasingly realised that efficient and accurate visual quality metrics can only be achieved through a thorough understanding of human visual perception in relation to visual media [21]. For this reason, several subjective studies were conducted to evaluate the impact of various system parameters on visual perception.

Yu et al. [22] studied the impact of viewing distance on quality perception and found, that there is no significant difference between two tested viewing distances. Barkowsky et al. [23] evaluated the effect of image presentation time on the final MOS. It was found that MOS from shorter presentation times can accurately be predicted from MOS given after longer presentation times. Bae et al. [24] investigated the trade-off between spatial resolution and quantisation noise and found, that human observers prefer, to some degree, a lower resolution to reduce the visibility of the compression artifacts.

Hauske et al. [25] performed early studies on the influence of different quantisation parameters (QP) and frame rates in H.264/AVC coded video on the resulting quality. More recently, De Simone et al. [26] studied the perceptual quality of H.264/AVC coded video containing packet loss. Pinson et al. [27] compared the subjective quality differences between H.264/AVC and MPEG-2 for high definition television, confirming the common belief that H.264/AVC provides similar quality at half the bit rate. It was shown though, that this is only given at bit rates below 18 Mbit/s.

Zhai et al. [28] found that perceived quality is affected in descending order of significance by the encoder type, video content, bit rate, frame rate, and frame size. The detectability of synthetic blocking, blur, ringing, and noise artifacts has been studied by Farias et al. [29] in a series of experiments. Amongst other findings, it was concluded that error visibility and perceived annoyance are highly correlated. The visibility of different types of noise in natural images has been evaluated by Winkler and Süsstrunk [30], who concluded that the noise thresholds increase significantly with image activity.

Pastrana-Vidal et al. [31] showed in their study, that overall perceived video quality can be estimated from independent spatial (sharpness) and temporal (fluidity) quality judgements. Huynh-Thu and Ghanbari [32] studied the impact of temporal artifacts in video and found that quality perception is more severely affected by jitter as compared to jerkiness artifacts. Liu et al. [33] investigated the impact of various packet loss patterns, considering the loss length, frequency, and temporal location based on PSNR measures. It was concluded that quality decreases linearly with loss length and is additive with respect to the number of losses. Cermak [34] surveyed people about the acceptable number of artifact occurrences in consumer video. It came out that on average consumers would not accept artifacts to be more frequent than once an hour, unless, the service cost is substantially reduced.

These studies highlight the many influencing factors that impact on the human perception of visual quality. Incorporating all these factors into a quality metric would likely result in a metric that reflects well the human visual perception of quality. However, such a metric would also be highly complex and computationally expensive, thus finding little use in many applications that have stringent limits on computational complexity.

1.2.3 Public subjective visual quality databases

To support reproducible research and to allow for quality metric design and validation, several image and video quality databases have been made publicly available in recent years. These databases usually consist of the stimuli that were presented during the subjective experiment and the quality scores that were obtained from the human observers.

Probably the most widely used image quality databases are the MICT database [35], the IRCCyN/IVC database [36], and the LIVE database [37], which are based on the assessment of distorted images mainly containing source coding artifacts and artificial artifacts such as white noise. More recently the elaborate TID image quality database has been made available [38], which covers a wide range of artifacts and provides MOS based on hundreds of observers. The latest image quality databases are the CSIQ [39], containing images with source coding distortions and artificial noise, and the WIQ [40] database. The latter has been made available by our group and contains images with complex distortion patterns caused by the simulation model of a wireless link. The test images and the subjective experiment procedures related to the WIQ database are explained in detail in Chapter 2. More information about the WIQ database can also be found in Appendix A.

For many years, the FRTV Phase I database [41] by the Video Quality Experts Group (VQEG) was the only available video quality database and has thus extensively been used for VQM design and validation. This has changed very recently with several video quality databases being made publicly available. The NYU database contains videos with packet loss distortions [42]. Two LIVE video quality databases are further available of which one database [43] contains videos with both compression and transmission distortions whereas the other database [44] is focused on wireless communications distortions. More recently, the EPFL-PoliMi video quality database [45] has been made available, focussing on transmission distortions. The latest release is the EPFL 3D video quality database [46], allowing for upcoming 3D VQM to be designed and validated on.

1.3 Classification of objective visual quality assessment

Both, the image metrics discussed in Section 1.1 and the subjective experiments introduced in Section 1.2 have their advantages and disadvantages. The former facilitates computationally efficient, automated assessment, which comes at the expense of low perceptual quality prediction performance. Subjective experiments, on the other hand, provide an accurate measure of perceived visual quality but are very tedious and not applicable in real-time. The aim of perceptual IQM and VQM design is to bridge the gap between the two methods and combine the advantages of automated assessment, omitting human interaction, with an accurate quality prediction performance. The design philosophies that are followed to achieve this goal are numerous and depend on the intended application of the quality metric. To shed some light on the different design methods we provide in the following a classification of visual quality metrics, in line with the philosophy of the classification presented in [47].

Visual quality metrics can generally be defined with respect to three different main factors that are considered in the metric design process:

1. The underlying knowledge and assumptions about the HVS.
2. The scope of visual distortions that are accounted for by the quality metric.
3. The information that is available from the undistorted reference stimulus.

This distinction is depicted in Fig. 3, with the three main factors being emphasised by the grey boxes.

Factor 1: Human visual system. Perceptual visual quality metrics aim to mimic the perception of a human observer and as such, it is intuitive to incorporate characteristics of the HVS into the metric design process [48, 49]. This can be done to different degrees of complexity, ranging from only simple approximations of some relevant HVS properties to very complex systems incorporating accurate models of the HVS. In general one can say, that more complex systems often result in better quality prediction performance, which comes at the cost of higher computational cost.

As suggested in [47], HVS-based metrics are generally designed with respect to either a bottom-up or a top-down philosophy. Following the former approach, the functionalities of the different HVS components [1, 50] are emulated by computational algorithms and integrated into a holistic model of perceived quality. The aim of this approach is to build a computational model that functions in a similar way as the integral parts of the HVS that are involved in quality perception.

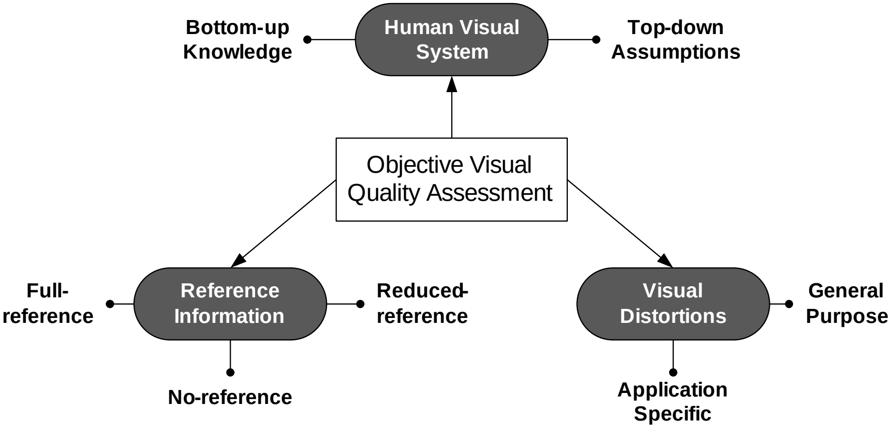


Figure 3: Classification of objective visual quality assessment methods [47].

On the other hand, metrics following the top-down philosophy do not aim to simulate each HVS component independently, but are based on high level assumptions about quality processing in the HVS. An example for this is the assumption that the HVS is adopted to extract structural information, rather than pixel information [14]. As such, the HVS is treated as a black box and the input-output relation is focused on, instead of the functionalities of the HVS.

The border between the two philosophies is blurry and quality metrics can incorporate both, specific functionalities of the HVS and also high level assumptions about the quality perception in the HVS. Considering the best of both worlds might lead to improved quality prediction performance.

Factor 2: Visual distortions. Depending on what distortion types are accounted for, perceptual quality metrics can further be classified into general purpose metrics and application specific metrics. General purpose models, also sometimes referred to as universal models, do not make any specific assumptions regarding the distortions in the visual content. As such, these metrics often focus on general features, such as natural scene statistics [51], and usually follow a HVS related design, with the aim of deployment in a wide range of applications.

On the contrary, application specific metrics have particular knowledge about distortions or make assumptions about distortions that can be expected in the visual content. This knowledge generally helps to simplify the metric design and to improve quality prediction performance for the particular application. This comes at the cost of worse performance when deployed in a different context than the one that the metric has been intended for. An example of application specific metrics are blocking metrics that are designed to specifically measure distortions in JPEG coded images. These metrics would perform poorly if used, for instance, to assess JPEG2000 coded images, which contain considerably different distortions.

Factor 3: Reference information. The amount of reference information that is available from an original image or video is a crucial design aspect of any visual quality metric. In this respect, 'original image/video' refers to an image/video that is considered to be distortion-free and of perfect quality and can, as such, be used as a reference to evaluate the quality degradations in a distorted image/video. Generally one can say that a higher amount of reference information facilitates easier metric design and promises better quality prediction performance. This is one reason why full-reference (FR) metrics are predominantly designed, where the entire original image/video is used as a reference for quality prediction of the distorted image/video. Clearly, the scope of these metrics is limited to scenarios where the reference image/video is available at the quality predictor, which is typically not the case in a communication context.

On the contrary to the FR approach, reference information is omitted entirely in no-reference (NR) metric design, where the quality is assessed solely on the distorted image/video. These methods are consequently often referred to as 'blind' metrics. Even though it is usually no problem for the HVS to judge the quality of a visual scene, it is in fact a highly difficult task for objective algorithms, as strong assumptions have to be made about what is actually considered to be perfect quality. For this reason, the efforts devoted to NR quality assessment have thus far focused on application specific metrics, such as blocking or blur metrics, and only little advances have been made towards universal NR quality predictors.

As a compromise between the FR and NR methods, reduced-reference (RR) quality metrics take into account only a subset of the reference information. As such, not the whole original image is accounted for but instead a set of extracted features. These features, along with the related features extracted from the distorted image/video, are then used for quality prediction. Metrics based on the RR approach thus combine the advantages of the FR and NR approaches, by avoiding the necessity for the entire reference image/video to be available and by considering some reference information to support the quality prediction task. In addition, RR metrics facilitate prediction of quality loss during a processing step, unlike NR metrics that quantify absolute quality. This is of particular interest in a communications context, where the quality loss during transmission can be identified. The bandwidth needed for the RR information to be sent from transmitter to receiver becomes then a crucial metric design aspect.

1.4 A brief history of objective visual quality assessment

To provide an overview of the advances in visual quality assessment thus far, we highlight in the following some of the milestones and review some of the major contributions. After surveying the early works, we look in particular at the advances in FR, NR, and RR quality assessment. In order to not burst the scope of this thesis, the survey is by no means considered to be complete and the more interested reader is referred to other survey publications [7, 52–57] for further discussions and references.

1.4.1 Early works and HVS-based metrics

Early work goes back several decades with quality models being developed for monochrome images by Mannos and Sakrison [58] and for monochrome video sequences by Lukas and Budrikis [59]. The field then experienced significant advances in the 1990's and at the turn of the century, with fundamental work on elaborate HVS-based quality models. The Visible Differences Predictor (VDP) by Daly [60] measured image fidelity as a function of display parameters and viewing conditions. The VDP was later adapted by Bradley [61] particularly for wavelet-based applications. Teo and Heeger [62] proposed a distortion measure based on the steerable pyramid transform [63] and contrast normalisation. Van den Branden Lambrecht and Verscheure [64] defined a multi-channel model of human spatio-temporal vision, specifically parameterised for video coding applications. The more general Sarnoff model proposed by Lubin [65] measures Just-Noticeable Differences (JND) in visual stimuli, based on psychophysical principles of human visual discrimination performance. Winkler proposed a Perceptual Distortion Metric (PDM) both for colour images [66] and for colour video [67], based on several properties of the HVS, including colour perception. The Digital Video Quality (DVQ) metric by Watson et al. [68] is based on an elaborate HVS model and was reported to perform similarly well to the Sarnoff JND model. Martens [69] performed multidimensional modelling to account for different factors that impact on the overall quality judgement. This was based on the hypothesis that the mapping from joint multidimensional attributes to a single overall quality judgement may vary considerably between human observers.

Parallel to the tremendous efforts towards finding plausible models of human visual perception, research efforts were still ongoing towards simple numerical metrics [70]. This was likely motivated by the high computational complexity of the HVS models and the computational constraints of most image and video processing systems. Towards the end of the 1990's, the trend then moved somewhat away from the elaborate HVS-based metrics (bottom-up approach) towards more engineering inspired metrics that often incorporate some high level assumptions about the HVS (top-down approach).

1.4.2 Advances in full-reference quality assessment

In the past decade, several FR metrics have been proposed that are expected to have a good correlation with human perception due to the availability of the reference image/video. Even though these metrics find only little use for application in a communication context, we include some major contributions in this review for completeness.

The Picture Quality Scale (PQS) by Yamashita et al. [71] is based on a number of spatial and temporal features extracted from video, including jitter, flicker, noise, and blur. A blockiness detector for MPEG coded video is proposed by Tan and Ghanbari [72]. Besides the blocking artifact extraction this metric also incorporates a simple perceptual model. The Structural Similarity (SSIM) index by Wang et al. [14] is based on the assumption that the HVS is adapted to extraction of structural rather than pixel information. The SSIM index is nowadays probably the most widely used image quality metric which can be attributed to its well balanced compromise between complexity and quality prediction performance. The Visual Information Fidelity (VIF) criterion by Sheikh and Bovik [73] approaches the quality prediction problem from an information theoretic viewpoint [74]. The VIF criterion has been developed in the same group as the SSIM index and is actually often superior to SSIM, which comes at the cost of higher computational complexity. The Visual Distortion Gauge by Lin et al. [75] is based on local contrast changes and has been found to be particularly effective in measuring blur artifacts and luminance fluctuations. The Visual Signal-to-Noise Ratio (VSNR) by Chandler and Hemami [76] deploys a two-stage approach, with the first stage determining a distortion detection threshold. If the distortions are suprathreshold, then the VSNR is computed based on perceived contrast and global precedence properties of the HVS. The Most Apparent Distortion (MAD) metric by Larson and Chandler [77] is based on the presumption that the HVS deploys different strategies for determining image quality, depending on if the visual distortions are near-threshold or suprathreshold. Thus, the model accounts for a detection-based strategy in high quality images and an appearance-based strategy in low quality images. The metric proposed by Li and Bovik [78] extends the existing SSIM index by a three-stage model to account for different categories of regional content; smooth regions, textured regions, and edge regions.

Most recently, the impact of temporal dynamics in video, both from a content and a distortion perspective, have increasingly been addressed in quality metric design. A temporal correction factor is deployed in the work by Ou et al. [79], in addition to a compression distortion metric, to account also for the impact of frame rate on the overall perceived quality. The metric by Liu et al. [33] accounts for both, degradations through source coding and packet loss. Several factors are integrated into the metric with regards to their impact on the overall perceived quality, including the loss length, the loss severity, loss location, the number of losses, and the loss patterns. The temporal variations of distortions are accounted for in the work by Ninassi et al. [80] by deploying short-term and long-term temporal pooling techniques. In particular, the short-term pooling was identified to be essential for improving the quality prediction performance of the metric. The Temporal Trajectory Aware Video Quality Measure (TetraVQM) by Barkowsky et al. [81] also mainly focuses on temporal issues, including frame freezes and skips, frame rate reduction, influence of scene cuts, and the tracking of the visibility of distorted objects. The Motion-based Video Integrity Evaluation (MOVIE) index, by Seshadrinathan and Bovik [82] involves both spatial and temporal distortion measures, but focuses in particular on evaluating motion quality along computed motion trajectories.

1.4.3 Advances in no-reference quality assessment

Given the considerably more difficult task of quality prediction without any reference, there have not been as many successful attempts to define NR quality metrics, in comparison to the number of FR models that have been proposed. In order to make the NR metric design more amenable, most models are in fact developed to serve particular applications for which the expected distortions are known.

A NR quality metric for JPEG images is proposed by Wang et al. [83]. The metric focuses on blocking artifacts, given their predominance in JPEG compressed images, and takes blur indirectly into account. The quality prediction performance of the work in [83] has been considerably improved by Horita et al. [84] through the introduction of local feature computations. A simple quality model for MPEG-4 coded video, based on frame rate and bit rate measures, has been proposed by Koumaras et al. [85]. A NR VQM based on the differences of local regions between two consecutive frames is proposed by Yang et al. [86]. These differences are weighted according to the temporal activity in the video. A quality metric based on motion characteristics and content classification is proposed by Ries et al. [87] for H.264 coded, low-resolution video sequences. Liu et al. [88] proposed a quality metric for JPEG and JPEG2000 coded images, based on localised gradient statistics.

In addition to these application specific NR quality metrics, there are also numerous metrics that are based on single feature or artifact measures. The most commonly addressed artifacts include blocking [89–91], blur [92, 93], ringing [94], and sharpness [95–97]. A metric combining synthetic blocking, blur, and noise artifacts is proposed in [98].

Artificial neural networks (ANN) have been found to perform well in predicting visual quality, based on either NR or RR features as input. Gastaldo et al. [99] developed a circular backpropagation (CBP) network for quality assessment of MPEG-2 video. Mohamed and Rubino [100] utilised a random neural network (RNN) for packet video quality assessment. Le Callet et al. [101] designed a quality predictor based on a convolutional neural network (CNN).

Another class of NR metrics [102, 103] makes unconventional use of data hiding techniques by means of watermarking [104]. A watermark is an image or pattern invisibly embedded into a host image and has been traditionally used for purposes such as copyright protection. In the context of quality assessment, however, the watermark is used to assess the quality of its host image based on the assumption that the host undergoes the same distortions as the watermark. This class of metrics is often referred to as 'pseudo NR', since no reference information is needed from the original image but instead, the watermark needs to be known.

The survey of NR quality assessment reveals that most metrics proposed thus far are application specific or even artifact specific, which shows that it is still a long way towards truly universal NR image quality metrics.

1.4.4 Advances in reduced-reference quality assessment

Reduced-reference quality assessment is of particular interest in scenarios where the reference image or video is not available, as is the case in image and video communications. Unlike NR methods, RR quality assessment allows for measurement of quality changes between an original and a distorted image or video, rather than judging the absolute quality. These might be some of the reasons why RR quality assessment received increased interest in recent years. The methods proposed thus far reported promising results in terms of quality prediction performance, being in many cases competitive with FR quality assessment with only a fraction of the reference information at hand. The amount of overhead in terms of RR information of course becomes a crucial metric design issue.

Probably the most widely used RR metric is the General Model of the Video Quality Model (VQM) by Pinson and Wolf [105]. The metric is said to be general purpose and applicable for various types of coding and transmission systems. The RR information is composed of the VQM and a set of calibration parameters. The total bandwidth needed for the RR information is 14% of the bandwidth of the uncompressed video sequence, of which 9.3% are for the VQM parameters and 4.7% for the calibration parameters. The metric proposed by Wang and Simoncelli [106] is based on a natural image statistic model in the wavelet domain. The method utilises a 3-scale and 4-orientation steerable pyramid decomposition [63] and the Kullback-Leibler distance (KLD) to quantify the difference of the wavelet coefficients. The quality prediction performance of the metric is very good and the RR side information is small, consisting of only 18 different feature measures (162 bits). However, the computational complexity of the metric is very high. The work proposed by Li and Wang [107] actually builds upon the design philosophy of the metric in [106] and improves its performance by introducing a divisive normalisation transformation that is in alignment with the de-correlation of neural responses in the early visual system.

The particularity of the metric by Masry et al. [108] is its scalability between an FR and an RR metric. As such, the bandwidth for the RR features can be traded off with the quality prediction performance, depending on the application. The metric by Yamada et al. [109] estimates the PSNR based on representative luminance values. The PSNR estimation is shown to work well but, of course, the agreement with subjectively perceived quality is questionable. The RR metric by Gunawan and Ghanbari [110] is based on local harmonic strength, focussing particularly on blocking and blur artifacts. The RR information comprises of 320 features of 8 bits each. The C4 criterion proposed by Carnec et al. [111], is maybe the only HVS-based quality model following the RR approach. The performance has been tested on several image quality databases and was found to be comparable to state-of-the-art FR metrics. However, the C4 metric is computationally highly complex. An unconventional RR quality assessment method based on distributed source coding is presented by Chono et al. [112]. Here, a feature vector is extracted from the original image and in order to reduce the RR size, only its Slepian-Wolf bit stream is transmitted to the receiver. The receiver can then correctly reconstruct the feature vector using the received image as side information. Lin et al. [113] propose a model for packet loss visibility. It was found that packet loss visibility is highly dependent on scene cuts and camera motion.

The methods discussed above all have one drawback in common, they face the problem that the RR information needs to be communicated as side information to the quality predictor. Depending on the RR size, this might be a particular problem in wireless communication networks where the bandwidth is scarce. To avoid this problem of sending the features separate from the reference over the channel, the concept of quality aware images has been introduced by Wang et al. [114] which is based on watermark embedding into the image or video frames. Unlike the other data hiding techniques discussed earlier, this method does not use the embedded watermark for quality evaluation at the receiver side. Instead the watermark itself contains the RR information which then only has to be extracted at the quality predictor. Therewith, no overhead is introduced and no ancillary channel for transmission of side information is needed. Of course, the capacity of the watermark is directly related to its visibility in the image and thus, the RR information is desired to be as small as possible.

1.5 Visual attention for quality assessment

Although the number and range of visual quality metrics that have been proposed thus far is large, most of them do not take into account an integral part of the HVS that can be assumed to have a major impact on the perception of overall perceived image and video quality. This HVS property is referred to as visual attention (VA) [115] and consists of higher cognitive processing deployed to reduce the complexity of scene analysis. For this purpose, a subset of the available visual information is selected by shifting the focus of attention across the visual scene to the most salient objects. It is because of the VA mechanisms that the HVS is able to cope with the abundant amount of visual information that it is confronted with at any instant in time.

Incorporating models of VA into visual quality assessment can thus be assumed to be very beneficial, since the viewer may be more likely to detect artifacts in highly salient regions, as compared to regions of low saliency. This is further supported by the fact that the input stage of the HVS, the retina, is highly space variant in sampling and processing of visual signals. The highest accuracy is located in the central point of focus, the fovea, and strongly diminishes towards the periphery of the visual field. As such, distortions in highly salient regions may be perceived in more detail and consequently, as being more annoying than distortions in regions of low saliency.

Given the potential relevance of VA for quality assessment, we will in the following give a brief introduction to the field and highlight in particular the aspects that are of interest in the given context of this thesis.

1.5.1 Visual attention

The main purpose of VA is to direct our gaze to the objects of interest in the visual scene, which is facilitated using rapid, saccadic eye movements. The attentional shift is guided by two main cues, namely, bottom-up and top-down. The former is fast, saliency driven, and independent of a particular task. It is understood that the bottom-up VA is performed in a pre-attentive manner across the visual field [116]. It is thus driven 'automatically' by certain low-level features that are experienced as visually salient. Top-down attention, on the other hand, is highly dependent on the viewing task and as such, it is typically slower and requires a voluntary effort to shift the gaze. Top-down attention is considered to have a modulatory effect on bottom-up attention [117] and as such, the two mechanisms together achieve that the most relevant information is continuously favoured at the expense of less relevant information.

Visual attention is guided by a large number of different low-level and high-level attributes [118]. Low-level attributes include, amongst others, colour, shape, size, and motion of objects. High-level attributes are based on semantic information and include, for instance, faces and written text [119]. Earlier work suggests that the pre-attentive, salient features are predominant in guiding attention [120], however, more recent work indicates that higher-level objects in fact have a stronger impact on VA [121].

Besides the visual attributes, VA has also been found to be highly dependent on the viewing task [122]. For instance, if a visual scene is observed without any task given, then the viewing behaviour is different as compared to the case where a particular search goal is followed. In the context of visual quality assessment, such a search goal could be the detection of visible distortions in natural scenes. Top-down attention, which mainly accounts for the task influence [123], has been investigated comparably less as compared to bottom-up attention, and is thus not as well understood. This is partly due to top-down cues being strongly driven by higher cognitive processes, whereas the saliency of the visual stimulus considerably supports the understanding of bottom-up attention.

It is well known that what we look at does not necessarily represent what we attend [118]. We can, for instance, gaze at a particular point in the visual field, but consciously attend another point in the periphery. Despite this fact, eye tracking and VA were found to be strongly interlinked [116] and thus, eye tracking experiments [124] are widely used to measure overt VA of human observers. Saliency maps (SM) created from eye tracking data are instrumental as a ground truth for the design and validation of VA models.

1.5.2 Visual attention models

Visual attention models aim to predict the attentional behaviour of human observers when viewing a visual scene. Generally, these models are not able to predict the sequential order of human fixations, the scanpath, but are limited to predicting the locations and objects that humans focus on [125, 126].

Many VA models were inspired by early works such as the feature integration theory by Treisman and Gelade [127], the guided search by Wolfe et al. [120], or the neural-based architecture by Koch and Ullman [128]. Especially the latter model constituted a theoretical basis for biologically plausible models incorporating characteristics of the HVS known to contribute to VA, such as multiple-scale processing, contrast sensitivity, and center surround processing. Probably the most widely used bottom-up VA model following this paradigm is the one by Itti et al. [129], which is based on the neuronal architecture of the early visual system, where multiple-scale image features are combined into a topographical SM. Le Meur et al. [130] also proposed a biologically inspired bottom-up VA model that predicts SM based on a three stage model including a visibility, a perception, and a grouping stage. A spatio-temporal model is proposed by Marat et al. [131] by accounting for the static and temporal pathways in the HVS.

Despite the plausibility of designing VA models inspired by the HVS, there is a strong trend towards content-based models [132–137]. These approaches usually incorporate different visual factors that are known to attract attention, such as the low-level and high-level attributes discussed in Section 1.5.1. Other VA models are based on Bayesian [138, 139], information theoretic [140, 141], or statistical approaches [142]. A nonparametric model based on only few assumptions about the VA mechanisms in the HVS and entirely trained from eye movement data is described in [143].

Only few models thus far have focused on top-down attentional processes [144, 145], mainly because they are relatively less well understood compared to bottom-up attentional processes. The model proposed by Ma et al. [146] accounts for bottom-up and top-down processes by combining spatio-temporal features with higher level semantic attributes through the deployment of a face detection algorithm.

There has also been several works that predict the level of perceived interest [147–149] or the level of importance [150,151]. Regions or objects that receive a high level of interest are often referred to as region-of-interest (ROI) or object-of-interest. These models are typically based on a segmentation of the image or video frames into areas of different levels of interest. In the simplest case, the visual scene is separated into a ROI and a background. Each of these regions is then assigned a level of interest. The design and validation of these models is performed in fundamentally different ways. Osberger et al. validate their importance prediction model [150] and their ROI prediction model [148] using gaze patterns recorded from eye tracking. Pinneli and Chandler [149], on the other hand, instructed a number of observers to rate the perceived levels of interest of different objects in an image, where the different objects were defined using segmented images of the Berkeley Segmentation Dataset and Benchmark [152]. The former approach assumes that the level of interest is strongly related to bottom-up, saliency driven visual cues, whereas the latter approach assumes a more top-down, task-driven relationship. The connection between the overt attention process of eye movements and the conscious decision process of interest selection has been studied in [153–156]. All works agree that there is a strong correlation between eye movements and the ROI selections. For this reason, ROI are sometimes determined from SM by defining appropriate thresholds for the different levels of interest [157].

In summary, the majority of the proposed models is concerned with bottom-up rather than top-down VA cues. Unlike with visual quality metrics, many of the VA models actually account for colour in images and video. This is essential for a VA model to perform well, as it has been shown [158] that human fixation locations differ considerably between coloured and grey scale versions of the same image. It should also be noted, that none of the VA models for video take into account auditory attention, even though auditory and visual information can be expected to have a strong interaction [159, 160].

1.5.3 Integration of visual attention into quality metrics

The integration of VA models into quality assessment is motivated by the generally accepted fact that VA is one of the most important features of the HVS and should thus not be neglected in visual quality metrics. As humans usually focus on highly salient regions in an image or video, outside these regions our sensitivity to distortions is considerably reduced. As such, they may be perceived as less annoying and may have a lower impact on the overall perceived quality. As a consequence, integrating visual saliency and perceptual distortion features may be crucial for improving IQM and VQM. However, most of today's visual quality metrics ignore the influence of these factors and weight distortions equally over the entire visual space.

In recent years, however, there has been increased efforts to evaluate the potential benefits of VA and saliency models for quality assessment [161]. The results reported generally agree that the incorporation of saliency information into quality metrics results in a significant improvement of the agreement with perceived visual quality. Cavallaro and Winkler [162] improved an image quality metric based on low-level features by integrating semantic information in terms of face segmentation. Maeder [163] and Moorthy and Bovik [164] reported quality prediction performance improvements by taking into account importance maps and appropriate pooling techniques. The modulatory aftereffects of VA have been found by Lu et al. [165] to improve quality assessment techniques. Barland and Saadane [151] improved a blur and ringing based quality metric for JPEG2000 by integrating an importance map into the distortion feature extraction. The Itti model [129] together with higher-level sematic information, such as face and text detection, have been integrated into a quality metric by You et al. [166], resulting in a very competitive performance to other contemporary quality metrics. Ma et al. [167] used a saliency predictor [136] to improve existing IQM and VQM considerably. Feng et al. [168] and Liu et al. [169], respectively, improved visual quality metrics and distortion visibility prediction models in the presence of packet loss. The importance of incorporating auditory cues into attention models was revealed in the subjective study on audiovisual attention by Lee et al. [170]. It was found, that the sound source attracts VA and as a consequence, the visual distortions in the regions far from the source are less perceived as compared to the distortions in the sound-emitting regions.

Despite the general agreement regarding the added value of VA models in quality assessment, there was also work that reported that no clear benefits from VA towards improved quality models could be identified [171]. This was tested for, using various pooling functions to integrate the saliency information with the distortion features. However, the study was based on gaze patterns obtained during quality assessment task and as such, the SM that were used do not directly reflect the content saliency but instead the viewing strategy of human observers when judging image quality. This may have a strong impact on the outcomes of the study and in fact, Larson et al. [172] have shown that greater improvement of quality metrics can be achieved when using eye tracking data recorded under task-free rather than task-based (quality assessment task) condition.

These results show that the incorporation of VA into quality metrics is a delicate process that needs to be carefully conducted. Personally, we believe that the benefits are also highly application dependent. More improvement is, for instance, expected in the case of localised rather than global distortions. We have also found that the added value is considerably higher in the case of video [173] rather than image [174] applications. This is mainly due to the continuous changes of the visual content in video and thus the more dynamic attention shifts and distortion fluctuations as compared to images.

One major challenge with the incorporation of VA into quality assessment is also the availability of a reliable saliency ground truth. For this reason, many of the above methods use possibly erroneous VA models instead of more reliable SM based on eye tracking data. To overcome this problem and to further advance this field of research, two eye tracking databases have recently been made freely available to the research community. The Visual Attention for Image Quality (VAIQ) database [175] has been released by our group at the Blekinge Institute of Technology, Sweden, in cooperation with the University of Western Sydney, Australia. The 'TUD Image Quality Database: Eye-Tracking Release 1' [176] has been released by Delft University of Technology, The Netherlands. The VAIQ database provides gaze patterns for three well known image quality databases, the MICT [35], the IRCCyN/IVC [36], and the LIVE [37] database. The TUD database provides gaze patterns for the LIVE database [37]. The overlap between the images used in the two eye tracking experiments further permits the comparison of SM created from gaze patterns of two independent viewer populations in different countries. These SM can, for instance, be analysed regarding their consistency between the two groups of viewers, which may lead to more insight regarding the reliability of the SM as a ground truth for quality metric design.

1.6 Quality assessment framework for image and video communications

In this thesis, we focus on subjective and objective quality assessment methods for image and video communications, and in particular wireless communications. The scarce channel bandwidth, the low computational power of mobile handheld devices, and the complex error patterns caused by the unreliable wireless channel constitute a difficult scenario with respect to measuring visual quality degradations as perceived by the end user. The aim is thus not to develop general purpose models but rather to limit our efforts to models that perform well in the communications context. As we are interested in estimating the quality loss that occurred during transmission, rather than the absolute quality of the received image or video, we thus focus on the design of RR quality metrics that use only a small amount of low-bandwidth features as reference information.

The general framework considered in this thesis is given in Fig. 4. Here, the integral parts of a wireless communication system are illustrated, comprising of a source en-/decoder, channel en-/decoder, (de)modulator, and the wireless channel. The dark-grey boxes determine the system components that need to be deployed to conduct the RR visual quality assessment. At the transmitter, a set of

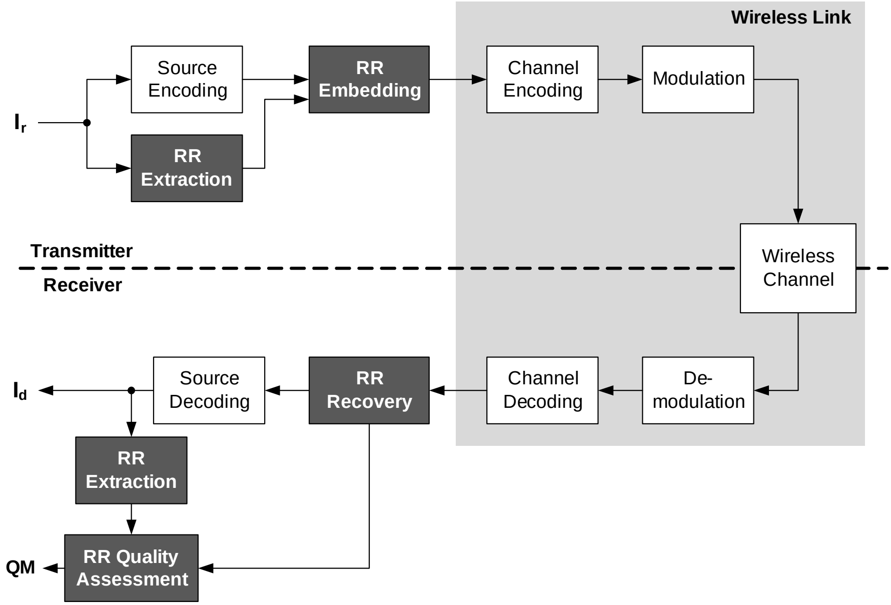


Figure 4: Framework for reduced-reference visual quality assessment in a wireless communication system.

low-bandwidth features is extracted from the reference image or video, I_r, which are sent as side information over the channel either, in-band as an additional header or in a dedicated control channel. A corresponding set of features is extracted from the received image or video, I_d, and used along with the recovered reference features to assess the quality degradation incurred during transmission. The RR quality metric (QM) may then facilitate link adaption techniques such as adaptive coding and modulation, power control, or automatic repeat request strategies.

Clearly, the diverse nature of the transmission errors and the strict bandwidth limitations of the channel are crucial aspects of the model design. A competitive quality metric is thus required to account for the complex distortion patterns while keeping the amount of RR as low as possible. Suitable pooling functions at the transmitter may thus be deployed to further condense the RR information. The complex error patterns due to the transmission errors further suggest the consideration of visual content saliency to be integrated into the metric design, as distortions in highly salient regions may be perceived to be more annoying as compared to distortions in regions of low saliency.

The general framework in Fig. 4 is adopted in Chapters 3, 6, 7, and 8, for the different models that are designed in the respective chapters. The components that are common in all cases are emphasised in Fig. 4 with the grey background. For convenience, these components are in later figures summarised by a single block named 'Wireless link'.

1.7 Thesis overview

The diagram in Fig. 5 provides an overview of the thesis and the interdependencies between the different chapters. The content of the thesis can be broadly divided into four parts. Part I consists of Chapters 2-4 and presents the design and validation of RR image quality metrics for wireless communications. Part II comprises of Chapters 5-8 and is concerned with the optimisation of the RR metrics presented in Part I. Chapter 8 also extends the metric design to incorporate visual saliency by means of ROI. Thus, Chapter 8 leads into Part III, which addresses in Chapters 9-11 VA and confidence of human observers during image quality assessment. Part IV comprises of Chapters 12-14 and addresses the impact of visual content saliency on the perceived annoyance of packet loss distortions in natural video sequences.

The following section shortly discusses the research methodology that is commonly deployed throughout the thesis. Section 1.7.2 then briefly introduces each chapter and summarises its contributions. Finally, a few remarks regarding notation in this thesis are given in Section 1.7.3.

1.7.1 Research methodology

The research methodology that is applied in each of the three parts generally involves the following three steps:

1. **Subjective experiments:** As human observers are still considered to be the most accurate source of perceptual data, we conducted several subjective experiments to obtain reliable ground truths for the model design. In particular, two image quality experiments were conducted for Part I and an ROI experiment was performed within Part II. A combined eye tracking and image quality experiment was undertaken for Part III and a combined eye tracking and video quality experiment was conducted to obtain a subjective ground truth for Part IV.

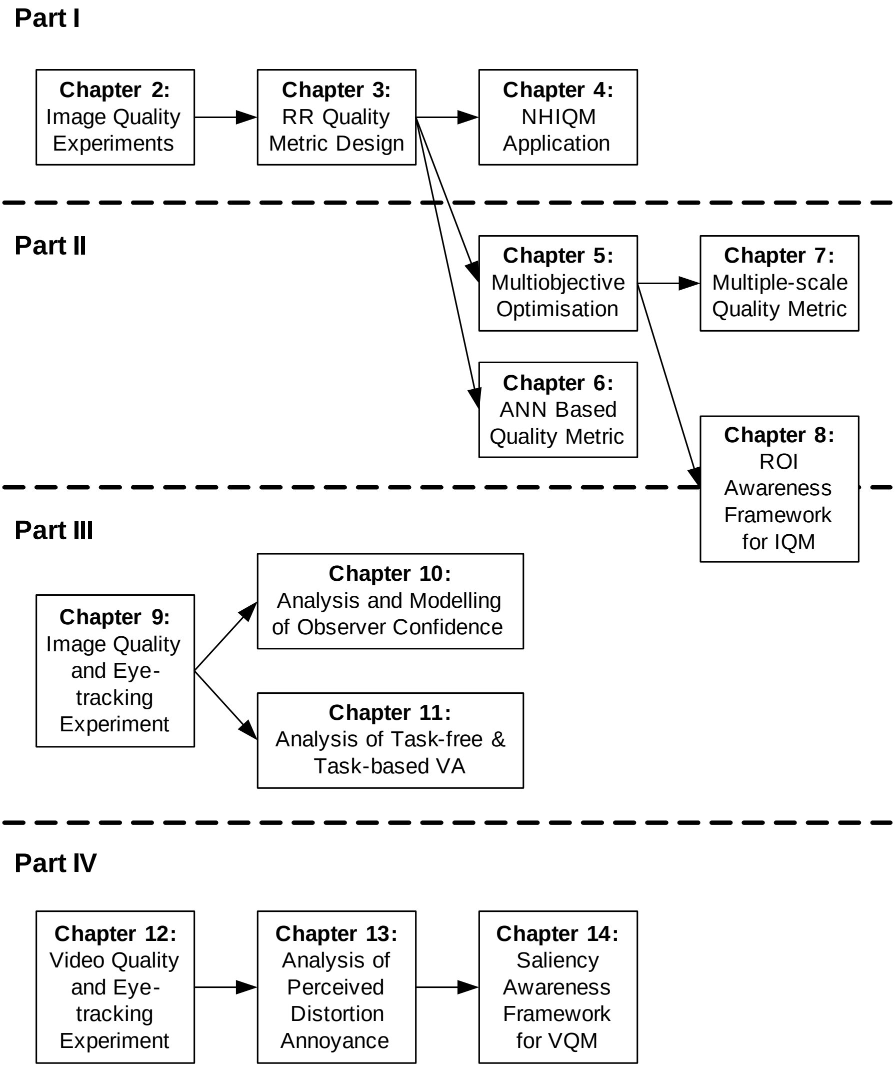

Part I
Chapter 2: Image Quality Experiments
Chapter 3: RR Quality Metric Design
Chapter 4: NHIQM Application
Part II
Chapter 5: Multiobjective Optimisation
Chapter 7: Multiple-scale Quality Metric
Chapter 6: ANN Based Quality Metric
Chapter 8: ROI Awareness Framework for IQM
Part III
Chapter 9: Image Quality and Eye-tracking Experiment
Chapter 10: Analysis and Modelling of Observer Confidence
Chapter 11: Analysis of Task-free & Task-based VA
Part IV
Chapter 12: Video Quality and Eye-tracking Experiment
Chapter 13: Analysis of Perceived Distortion Annoyance
Chapter 14: Saliency Awareness Framework for VQM

Figure 5: Overview of the four parts and the interdependencies between the chapters contained in this thesis.

2. **Subjective data analysis:** The outcomes of the subjective experiments are instrumental for the design of visual quality metrics and saliency models. However, they also provide valuable insight into the human visual perception of natural image and video content in the presence of structural distortions. As such, a detailed analysis of the data obtained from the subjective experiments generally precedes the objective modelling in each of the four parts.

3. **Objective modelling:** The final goal of each part is to determine objective models that are able to predict human ratings, typically quality ratings. In this respect, we develop original quality prediction models but also models that aim on improving existing quality metrics, for instance, by incorporating saliency information into metrics that do not account for this phenomenon.

1.7.2 Summary and contributions

The content of all chapters in this thesis are briefly summarised in the following and the respective contributions of each chapter are highlighted:

- **Chapter 1**: The introduction provides a current overview of the field of visual quality research. In particular, an up-to-date survey of visual quality assessment is presented, highlighting the advances from early works to the latest contributions in the field. Additionally, recent work related to the benefits of VA in quality assessment are briefly discussed. Parts of Chapter 1 have been published as [55].

- **Chapter 2**: Two subjective image quality experiments were conducted involving a total of 60 human observers. The experiment outcomes reveal valuable insight into the human visual perception of transmission distortions in natural image content. The test images and the MOS are made freely available to the research community in the Wireless Imaging Quality (WIQ) database (see also Appendix A). Unlike any of the other publicly available databases (see Section 1.2.3), the test images in our database contain complex distortion patterns caused by a simulation model of a wireless link. Parts of Chapter 2 have been published as [16, 40].

- **Chapter 3**: Reduced-reference image quality metrics, namely, the Normalised Hybrid Image Quality Metric (NHIQM) and the perceptual relevance weighted L_p-norm, are designed and validated using the subjective experiment results from Chapter 2. This work is a continuation of the work presented in [177]. The extensions to the previous work include modifications of feature extraction algorithms, the deployment of alternative feature pooling strategies, an extreme value normalisation of perceptual relevance weights, a detailed statistical analysis of the objective features in the image content, and the derivation of individual mapping functions for the designed metrics. Furthermore, the concept of metric training and validation has not been considered in the previous work and the metrics were not compared to other contemporary quality metrics. Both these shortcomings are addressed in this chapter. Parts of Chapter 3 have been published as [16, 178].

- **Chapter 4**: The applicability of NHIQM for de-blocking filter design of H.263 coded video is analysed to highlight the effectiveness of the metric. It is shown, that NHIQM is able to distinguish between different quality levels and to quantify changes of different artifacts due to the filtering process. Parts of Chapter 4 have been published as [179].

- **Chapter 5**: A multiobjective optimisation framework is proposed to determine the optimal perceptual relevance weights of feature-based image quality metrics. The effectiveness of the framework is demonstrated with NHIQM showing that the quality prediction performance can be considerably improved while generalisation ability is maintained. Parts of Chapter 5 have been published as [180].

- **Chapter 6**: An artificial neural network (ANN) using structural features as input is proposed. The ANN predicts well the perceived visual quality of natural images with transmission distortions while having a strong generalisation ability. It is shown that the ANN performs equally well with RR features as input as well as with NR features. The network is of low computational complexity as only few neurons are needed for the metric computation. Parts of Chapter 6 have been published as [181].

- **Chapter 7**: Multi-resolution structural feature extraction using the Gaussian pyramid decomposition is proposed for objective quality assessment. The resulting metric shows a superior performance in comparison to its single-resolution versions presented in Chapter 3. Parts of Chapter 7 have been published as [182, 183].

- **Chapter 8**: A subjective experiment is reported that we conducted to identify ROI in the reference images of the quality experiments introduced in Chapter 2. The experiment outcomes provide interesting insight into the ROI selection behaviour of human observers and the ROI selections are made publicly available to the research community (see also Appendix B). A ROI awareness framework is further proposed that is shown to improve the prediction performance of existing image quality metrics. Parts of Chapter 8 have been published as [174, 184, 185].

- **Chapter 9**: Two eye tracking experiments are reported in which gaze patterns of human observers were recorded when viewing natural image content. One experiment was conducted under natural viewing conditions (task-free) and the other experiment under image quality assessment task. In the latter experiment, additional information was collected in terms of confidence scores and response times. The recorded gaze patterns of both experiments constitute valuable information with respect to human viewing behaviour, both under task-free and task-based conditions. The gaze patterns from the task-free eye tracking experiment are made freely available in the Visual Attention for Image Quality (VAIQ) database (see also Appendix C). Parts of Chapter 9 have been published as [186].

- **Chapter 10**: The confidence of human observers during image quality assessment is analysed in detail. For this purpose, the quality scores, confidence scores, and response times of the task-based eye tracking experiment in Chapter 9 are analysed in detail. A prediction model of human observer confidence is further proposed that provides additional reliability information about MOS, as a complement to widely computed confidence intervals. Parts of Chapter 10 have been published as [187].

- **Chapter 11**: The task-free and task-based eye tracking data are analysed in detail. The former data is analysed with respect to the inter-observer variability when viewing natural image content. The latter data is investigated regarding the relative impact of the perceived level of interest and of structural distortions on the viewing behaviour during image quality assessment. Parts of Chapter 11 have been published as [153, 156, 188, 189].

- **Chapter 12**: A combined eye tracking and quality assessment experiment is reported that we performed to determine the perceived level of annoyance of localised packet loss distortions in relation to the underlying content saliency in natural video sequences. Parts of Chapter 12 have been published as [190].

- **Chapter 13**: The annoyance scores and the eye tracking data recorded in the experiment from Chapter 12 are analysed in detail. The results discussed in this chapter are considered to be highly valuable in two respects. Firstly, to gain a better understanding of the interdependence between content saliency and localised transmission distortions onto perceived visual quality. Secondly, to incorporate saliency information into existing and future VQM for better agreement with human perception. Parts of Chapter 13 have been published as [190].

- **Chapter 14**: We present a simple saliency awareness model for existing VQM. The model is non-intrusive, meaning that the actual metric does not need to be changed. The improvement of quality prediction performance based on the saliency awareness model is shown on a contemporary VQM. Parts of Chapter 14 have been published as [173].

1.7.3 Some remarks

In general, the aim of perceptual visual quality assessment is to evaluate the quality of a visual signal, either as an absolute measure or relative to a reference signal. In relation to the latter, the quality does not necessarily have to be worse than the reference, for instance, in the case of processing through image and video enhancement systems. However, in this thesis, we consider models that only estimate the quality loss of one visual signal as compared to a reference signal, based on the valid assumption that there is generally no quality enhancement performed during transmission. Therefore, the reference signal is considered to be distortion free and of perfect quality and any changes in the received signal are considered to degrade the perceived level of quality to some degree.

In this respect, the reference signal is in the remainder of the thesis labelled with r and the distorted signal is labelled with d. In the scope of this thesis, the distorted stimuli all contain degradations to some degree. In a real world application, not every transmitted signal would necessarily contain distortions.

The reader may have noticed the terms 'distortion' and 'artifact' that have been used throughout the introduction. The term 'distortion' is here considered to be a general degradation of the reference signal, not specifying any particular type. The term 'artifact', on the other hand, is here used to denote a particular kind of distortion, for instance, blocking, blur, or noise.

2 Subjective Wireless Imaging Quality Assessment

For the RR image quality metric design, a ground truth is needed in terms of subjective quality scores. However, the publicly available image quality databases (see Section 1.2.3) generally focus on compression distortions and artificial artifacts, such as white noise, and are thus not suitable for quality metric design in wireless image communication. For this purpose, we conducted subjective image quality experiments in two independent laboratories to obtain quality scores for a number of natural images with transmission distortions.

The first experiment was conducted at the Western Australian Telecommunications Research Institute (WATRI) in Perth, Australia, and is in the following referred to as experiment E1. The second experiment took place at the Blekinge Institute of Technology (BIT) in Ronneby, Sweden, and is in the following referred to as experiment E2. Both experiments were designed according to the guidelines outlined in Rec. BT.500-11 [19] of the ITU-R.

In the following sections, the creation of the test images used in the experiments is discussed, the experiments are explained in detail, and the experiments' outcomes are presented along with a statistical analysis.

2.1 Creation of test images

2.1.1 System under test

To create the distorted images from a number of undistorted reference images, we consider in the scope of this thesis a particular realisation of the wireless link model as shown in Fig. 4. In particular, the JPEG format was chosen to source encode the images prior to transmission. It is noted that JPEG is a lossy image coding technique using a block discrete cosine transform (DCT) based algorithm, thus, facilitating an easy transition to state-of-the-art DCT-based video codecs, such as H.264/AVC [191]. Due to the quantisation of DCT coefficients, artifacts may already be introduced during source encoding. A $(31, 21)$ Bose-Chaudhuri-Hocquenghem (BCH) code was then used for error protection purposes [192] and binary phase shift keying (BPSK) for modulation. An uncorrelated Rayleigh flat fading channel in the presence of additive white Gaussian noise (AWGN) was implemented as a simple model of the wireless channel [193]. Severe fading conditions may cause bit errors or burst errors in the transmitted signal which are beyond the correction capabilities of the channel decoder and as a result, distortions may be induced in the decoded image in addition to the ones purely caused by the source encoding. To produce severe transmission conditions, the average bit energy to noise power spectral density ratio E_b/N_0 was chosen as 5 dB.

It should be noted, that the system under test is a particular realisation of the wireless link outlined in Section 1.6. However, the image quality metrics proposed in this thesis can be easily adopted to other specific system components, given that suitable test images and subjective data (MOS) are available, which are crucial for the metric design. This may for instance include an extension from JPEG to JPEG2000 or to measuring spatial artifacts in video, such as H.264/AVC.

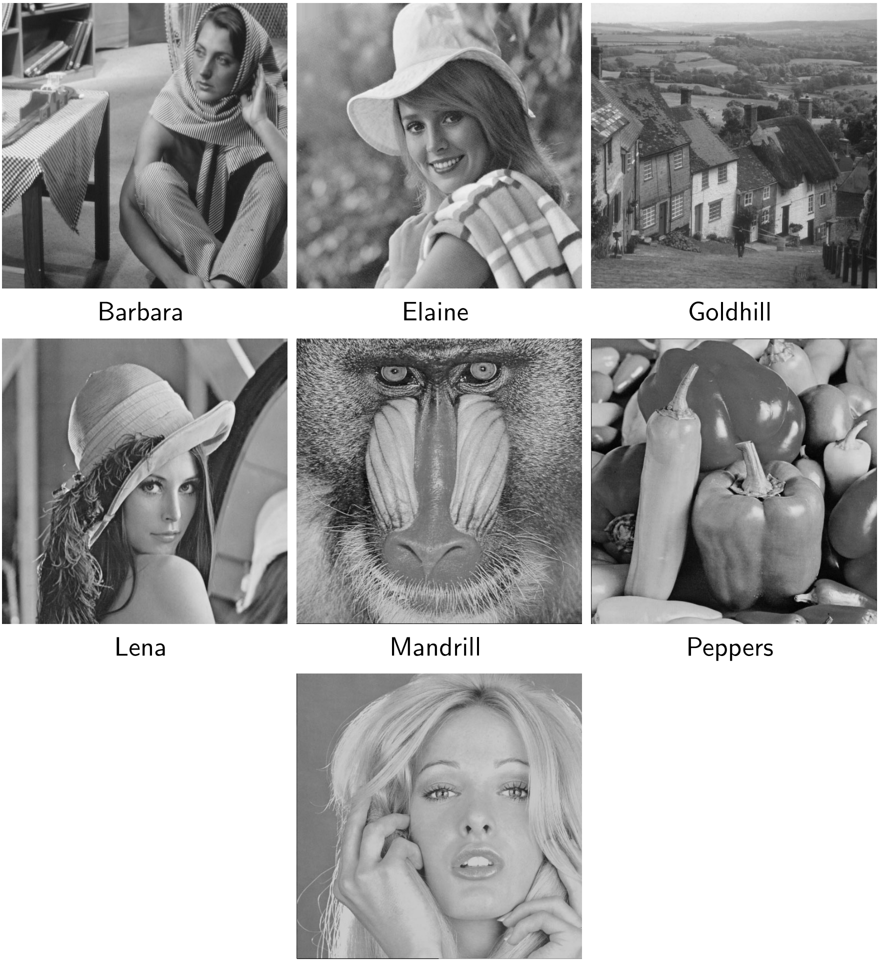
2.1.2 Reference and distorted images

A set of seven reference images, $\mathcal{I}_R$, of dimension 512×512 pixels and represented in grey scale was chosen to cover a variety of textures, complexities, and contents. The reference images are shown in Fig. 6. The system under test, as explained in Section 2.1.1, was utilised to create a large number of distorted images. Two sets of fourty distorted images each, $\mathcal{I}_1$ and $\mathcal{I}_2$, were then selected to be used in the two subjective experiments E1 and E2, respectively. The images were chosen such as to cover a wide variety of artifacts and also a broad range of severities for each of the artifacts, from almost invisible to highly distorted. Thus, the sets of test images incorporate distortions near the JND regime to artifacts widely covering the suprathreshold regime [194].

2.1.3 Artifacts observed in the distorted images

The system under test as outlined in Section 2.1.1 turned out to be beneficial with respect to generating impaired images with a broad range of distortion types, severities, and distributions. Specifically, the range of artifacts spanned beyond those typically induced by source encoding such as blocking and blur but also comprised of ringing, block intensity shifts, lost blocks, and combinations thereof. These artifacts are briefly discussed in the following. In addition, some example images are shown in Fig. 7 to illustrate the observed artifacts.

Blocking: Blocking artifacts are inherent with block-based image and video compression techniques, such as JPEG or H.264/AVC. Blocking (sometimes also referred to as blockiness) can be observed as surface discontinuity at block boundaries and is a consequence of the independent quantisation of the individual blocks of pixels. In particular, in DCT-based image and video compression, blocking is present on the 8×8 block borders due to independent quantisation of the DCT coefficients. Blocking artifacts are illustrated in Fig. 7 (a) and (b).



Barbara

Elaine

Goldhill

Lena

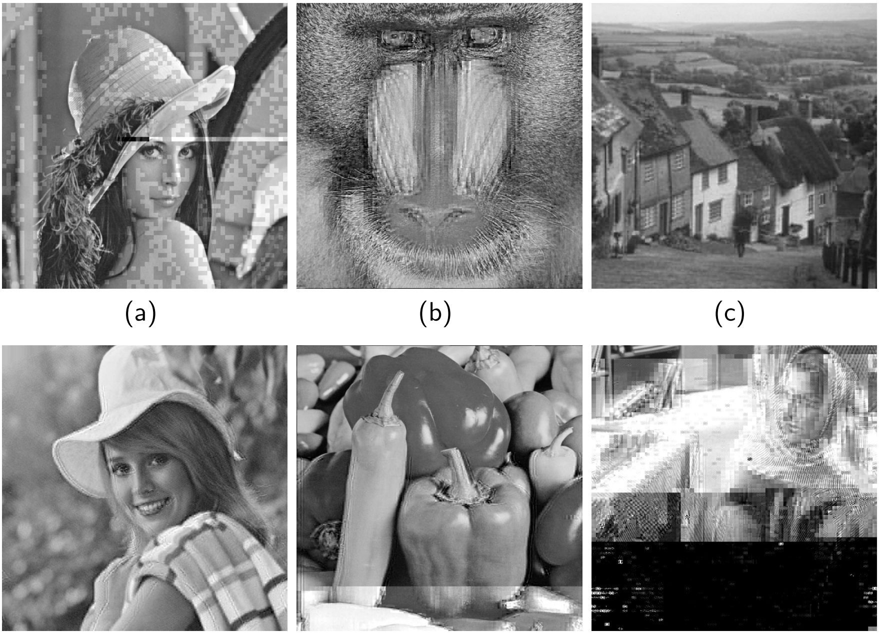
Mandrill

Peppers

Tiffany

Figure 6: Reference images used in experiments E1 and E2.

Blur: Blur artifacts relate to the loss of spatial detail of visual content and are typically observed as texture blur. In addition, blur (sometimes also referred to as blurriness) may be observed due to a loss of semantic information that is carried



(f)

Figure 7: Examples of distorted test images showing different artifacts: (a) 'Lena' with blocking, (b) 'Mandrill' with blocking, (c) 'Goldhill' with blur in 8×8 blocks, (d) 'Elaine' with ringing, (e) 'Peppers' with ringing and block intensity shifts, and (f) 'Barbara' with a combination of severe artifacts.

by the shapes of objects in an image. In this case, edge smoothness relates to a reduction of edge sharpness and contributes to blur. In relation to compression, blur is a consequence of the coarse quantisation of frequency components and the associated suppression of high-frequency coefficients. Global blur artifacts are prevalent in discrete wavelet transform (DWT) based codecs such as JPEG2000. In case of JPEG compression, blur is usually observed within the 8×8 blocks rather than on a global scale. Blur artifacts are illustrated in Fig. 7 (c).

Ringing: The artifact of ringing appears to the human observer as periodic pseudo edges around the original edges of the objects in an image. Ringing is caused by improper truncation of high-frequency components, which in turn can be noticed as high-frequency irregularities in the reconstruction. Ringing is usually more evident along high contrast edges, especially if these edges are located in areas of smooth textures. Ringing artifacts are illustrated in Fig. 7 (d) and (e).

Block intensity shifts and lost blocks: In general, masking occurs when the visibility of a stimulus is reduced due to the presence of another stimulus [18]. In this context, intensity shifts in parts of an image, or the whole image, may result in either a darker or brighter appearance of the area as compared to the original image and thus cause a change in visibility of the visual content. These artifacts may appear in the presence of strong multipath fading in wireless image communication and are in the following referred to as block intensity shifts. In the worst case, entire image blocks are lost resulting in parts of the image being black. Block intensity shifts are illustrated in Fig. 7 (e) and (f).

2.2 Details of experiments E1 and E2

2.2.1 Laboratory environments

The general viewing conditions in both experiments were arranged as specified in the ITU-R Rec. BT.500-11 [19] for a laboratory environment. The room for experiment E1 was equipped with two 17" cathode ray tube (CRT) monitors of type Sony CPD-E200 and for experiment E2 a pair of 17" CRT monitors of type DELL and Samtron 75E was used. The ratio of inactive screen luminance to peak luminance was kept below a value of 0.02. The luminance ratio of the screen when displaying black in a dark room to displaying peak white was approximately 0.01. The display brightness and contrast was set up with picture line-up generation equipment (PLUGE) according to Recommendations ITU-R BT.814 [195] and ITU-R BT.815 [196]. The calibration of the screens was performed with the calibration equipment ColorCAL from Cambridge Research System Ltd. [197], England, while the DisplayMate software [198] was used as pattern generator.

Due to its large impact on the artifact perceivability, the viewing distance must be taken into consideration when conducting a subjective experiment. The viewing distance is in the range of four times (4H) to six times (6H) the height H of the stimulus, as stated in Rec. ITU-R BT.1129-2 [199]. The distance of 4H was selected here in order to provide better image details to the viewers.

2.2.2 Viewer panels

It is recommended by the ITU-R [19] that the number of observers participating in a subjective image quality experiment should not be lower than 15. The VQEG even recommends to involve at least 24 viewers [200, 201]. To satisfy these constraints, 30 viewers participated in each experiment, E1 and E2. All participants were non-experts, meaning, that they were not professionally involved in image quality assessment at their work. In order to support consistency and eliminate systematic differences among results at the different testing laboratories (WATRI and BIT), similar panels of test subjects in terms of occupational category, gender, and age were established. In particular, 25 males and 5 females, participated in experiment E1. They were all university staff and students and their ages were distributed in the range of 21 to 39 years with the average age being 27 years. In the second experiment, E2, 24 males and 6 females participated. Again, they were all university staff and students and their ages were distributed in the range of 20 to 53 years with the average age being 27 years.

2.2.3 Test procedures

Different test methodologies are provided in detail in [19] to best match the objectives and circumstances of the assessment problem. The methodologies are mainly classified into two categories, as double-stimulus and single-stimulus. In double-stimulus, the reference image is presented to the viewer along with the test image. On the other hand, in single-stimulus, the reference image is not explicitly presented and may be shown transparently for the experimenter to evaluate judgement consistency of the subject. As we focus on RR metric design in this thesis, where partial information related to the reference image is available, we chose to deploy a double-stimulus method, the double-stimulus continuous quality scale (DSCQS). Moreover, DSCQS has been shown to have low sensitivity to contextual effects [19, 202]. Contextual effects occur when the subjective rating of an image is influenced by presentation order and severity of impairments. This relates to the phenomenon that test subjects may tend to give an image a lower score than it might have normally been given if its presentation was scheduled after a less distorted image.

In both experiments E1 and E2, the test sessions were divided into two sections. The duration of each section was well under 30 minutes and consisted of a stabilisation and a test trial. The stabilisation trials served as a warm-up to the actual test trial in each section for the participants to familiarise themselves with the test mechanism. In addition, one training trial was conducted at the very beginning of the test session to demonstrate the range of artifacts to be expected during the actual test and to explain the test procedure to the viewers. The scores obtained during the training and stabilisation trials were not processed but only the scores given during the test trials. In order to reduce the viewers' fatigue, a 15 minutes break was given between sections.

Given the DSCQS method, pairs of images A and B were presented in alternating order to the viewers for assessment, with one image being the undistorted reference image and the other being the distorted test image. The participants were asked to judge the quality of both images. In this respect, the undistorted image served as a reference to judge the quality of the distorted image. However, the participants were not told which image was the reference. As the DSCQS method is quite sensitive to small quality differences, it is well suited to not just cope with highly distorted test images (suprathreshold level) but also with cases where the quality of the original and the distorted image is very similar (JND level).

The grading was performed on a continuous scale ranging from 0-100. As a general guide for the participants, the adjectival categories of the 5-point scale (Excellent, Good, Fair, Poor, Bad) were additionally presented along the continuous scale. Given the pair of images A and B, the viewers were requested to assess their quality by placing a mark on each quality scale. As the reference and distorted image appeared in pseudo random order, A and B may have referred to either the reference image or the distorted image, depending on the actual arrangement of images in an assessment pair.

2.2.4 Post-processing of subjective scores

Let $s_r(n,k)$ and $s_d(n,k)$ denote the scores of the n^{th} viewer for the k^{th} reference image and distorted image, respectively. Given these scores, we compute a difference score as follows

$$\Delta s(n,k) = |s_r(n,k) - s_d(n,k)| \tag{3}$$

which can be interpreted as a quality degradation of the distorted image in relation to the reference image and as such, a higher $\Delta s(n,k)$ relates to a stronger quality degradation. By subtracting this value from the maximum of the quality scale as follows

$$s(n,k) = 100 - \Delta s(n,k) \tag{4}$$

we obtain a score that is positively related to the quality of the distorted image, with a higher $s(n,k)$ indicating a higher subjectively perceived quality.

To obtain a single score for each image, the difference scores $\Delta s(n,k)$ are averaged over all N viewers for the k^{th} image as follows

$$\text{DMOS}_k = \frac{1}{N}\sum_{n=1}^{N}\Delta s(n,k) \tag{5}$$

which is known as the Differential Mean Opinion Score (DMOS). Similarly to the difference scores, the scores $s(n,k)$ are combined to the Mean Opinion Scores (MOS) as follows

$$\text{MOS}_k = \frac{1}{N}\sum_{n=1}^{N} s(n,k). \tag{6}$$

In the remainder of this thesis we use the MOS. It should be emphasized here, that the MOS typically represents the accumulated scores in single stimulus quality assessment, where only the distorted image is rated. However, we use the term MOS here to distinguish these values from the DMOS and to indicate that a higher score relates to higher subjective quality.

2.3 Evaluation of experiments E1 and E2

The outcomes of the subjective experiments are discussed in the following by means of a statistical analysis. In this respect, a concise representation of the subjective data can be achieved by calculating conventional statistics like the mean, variance, skewness, and kurtosis of the related distribution of opinion scores. The statistical analysis of this data reflects the fact that perceived quality is a subjective measure and hence may be described statistically.

2.3.1 Statistical measures

Let the MOS value for the k^{th} image in a set $\mathcal{K}$ of size K be denoted here as μ_k. Then, we have

$$\mu_k = \frac{1}{N}\sum_{n=1}^{N} s(n,k) \tag{7}$$

where N is the number of viewers. The confidence interval (CI) associated with the MOS of each examined image is given by

$$[\mu_k - \epsilon_k, \mu_k + \epsilon_k]. \tag{8}$$

The deviation term ϵ_k in (8) can be derived from the standard deviation σ_k and the number N of viewers and is given for a 95% CI according to [19] by

$$\epsilon_k = 1.96 \frac{\sigma_k}{\sqrt{N}} \tag{9}$$

where the standard deviation σ_k for the k^{th} image is defined as the square root of the variance

$$\sigma_k^2 = \frac{1}{N-1} \sum_{n=1}^{N} (s(n,k) - \mu_k)^2. \tag{10}$$

The skewness β measures the degree of asymmetry of data around the mean value of a distribution of samples and is defined by the second and third central moments m_2 and m_3, respectively, as

$$\beta = \frac{m_3}{m_2^{3/2}} \tag{11}$$

where the l^{th} central moment m_l is defined as

$$m_l = \frac{1}{N} \sum_{n=1}^{N} (s(n,k) - \mu_k)^l. \tag{12}$$

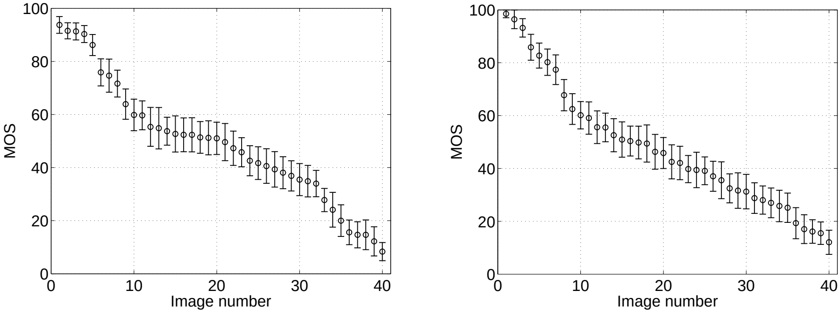
The peakedness of a distribution can be quantified by the kurtosis γ, which measures how outlier-prone a distribution is. The kurtosis is defined by the second and fourth central moments m_2 and m_4, respectively, as

$$\gamma = \frac{m_4}{m_2^2}. \tag{13}$$

It should be mentioned that the kurtosis of the normal distribution is 3. If the considered distribution is more outlier-prone than the normal distribution (leptokurtic), it results in a kurtosis greater than 3. On the other hand, if it is less outlier-prone than the normal distribution (platykurtic), it gives a kurtosis less than 3. A distribution of scores is usually considered to be approximately normal if the kurtosis is between 2 and 4.

2.3.2 Statistical analysis of the subjective data

Figures 8(a)-(b) show the scatter plots of MOS for E1 and E2, respectively. The 40 images in each experiment are ordered with respect to decreasing MOS. It can



(a)

(b)

Figure 8: Perceived quality ordered according to decreasing MOS with error bars indicating the 95% CI for: (a) E1 and (b) E2.

be seen from the figures that the material presented to the viewers resulted in a wide range of perceptual quality ratings indeed for both subjective experiments. As such, both experiments contained the extreme cases of excellent and bad image quality while the intermediate quality decreases approximately linearly in between. It is also observed that the spread of ratings around the MOS in terms of the 95% CI is generally narrower for the images at the upper and lower end of the perceptual quality scale.

Figures 9(a)-(d) show the MOS, variance, skewness, and kurtosis, respectively, for each image that was rated in the two subjective experiments. The image samples in all four figures are, as in Fig. 8, ordered with respect to decreasing MOS. In addition to the image samples the figures depict the related fits to these statistics, which reveal good agreement among the data for the two subjective experiments, as the fits progress closely in the same manner over the ordered image samples. This indicates that the two experiments have been very well aligned with each other and also that the two viewer panels, even though originating from different countries, seem to have given a similar range of quality scores for the test images they have been shown.

Figure 9(a) depicts the MOS for the impaired images along with a linear fit through this data. It can be seen from the figure, that the linear fit for both experiments are very close, suggesting that the set of image samples used in the two independent experiments at WATRI and BIT comprised of a similar range of quality impairments.

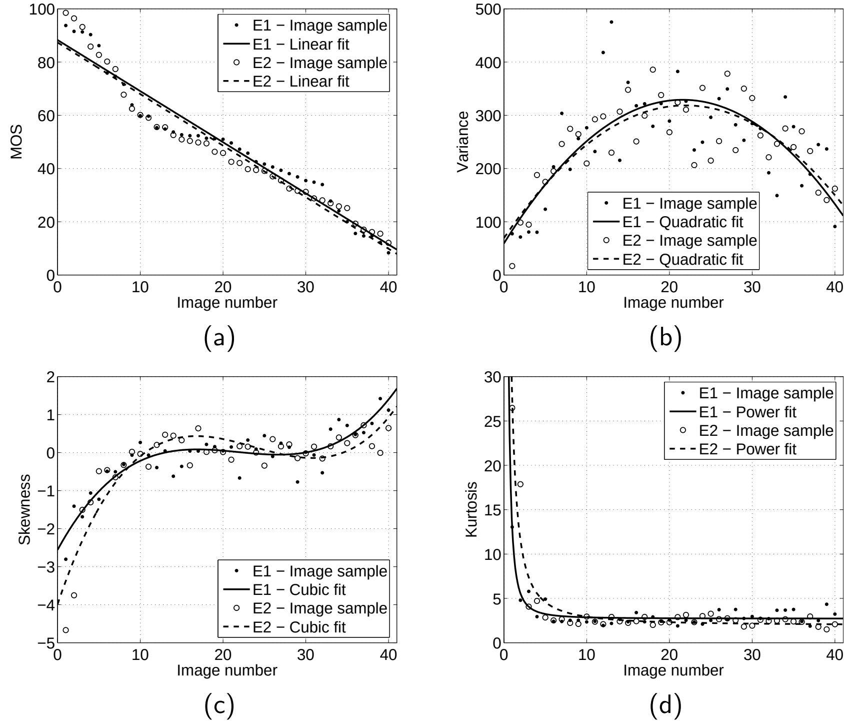


Figure 9: Statistics of opinion scores for the distorted test images: (a) MOS, (b) Variance, (c) Skewness, and (d) Kurtosis.

Figure 9(b) shows the variance of all opinion scores for each image sample. The variance can be regarded as a measure of how much the viewers agree on the perceived quality of a certain image sample. The smaller the variance, the more pronounced the agreement between all viewers. It can clearly be seen that the variance is relatively small for images that have obtained either excellent or bad subjective quality ratings. In contrast, in the region where perceptual quality of the impaired images ranges between good and poor, the variance tends to be larger with the peak at about the middle of the quality range. This result indicates that the viewers appear to be rather sure whether an image sample is of excellent or bad quality while opinions about images of average quality differ to a wider extent. These conclusions are supported by the 95% CI shown in Fig. 8(a)-(b), which are narrower for images rated as being excellent and bad. It should be noted though, that the narrower CI and analogously the lower variance, at either end of the quality scale are to some degree also a result of the scale limits [203]. In this respect, it is of interest to evaluate the impact of the observers' confidence on the quality scores given. This issue is looked at in Chapter 10, where additional confidence scores are introduced to complement the CI.

Figure 9(c) shows the skewness of the opinion score distribution for each image sample. In the context of the subjective ratings of image quality, a negative or positive skewness translates to the subjective scores being more spread towards lower or higher values than the MOS, respectively. For the images that were perceived as being of high quality, the negative skewness indicates that subjective scores tend to be asymmetrically spread around the MOS towards lower opinion scores and thus, that a number of viewers gave significantly lower quality scores as compared to the MOS. In the other extreme of image quality being perceived as bad, the positive skewness points to an asymmetrically spread around the MOS towards higher opinion scores. However, the positive skewness is not as distinct as the negative skewness at the high quality end, suggesting that the agreement of low quality was higher as compared to the agreement about high quality. The asymmetry in subjective scores for the extreme cases of excellent and bad quality is thought to be due to the rating scale being limited to 100 and 0, respectively. As such, subjective scores have to approach the maximal and minimal possible rating from below or above, respectively. The skewness of around zero for the middle range of qualities reveals that the subjective scores seem to be symmetrically distributed with respect to MOS.

Figure9(d) provides the kurtosis for each impaired image sample. It can be seen from the figure, that the distribution of subjective scores for some of the images scoring high MOS values in both experiments give kurtosis values much greater than of a normal distribution. This is evidence for outliers, meaning, that a few of the viewers gave a low image quality rating whereas the majority of viewers agreed on a high image quality. With the progression of images towards decreasing MOS, the associated kurtosis fits quickly level out around the value 3, pointing to a normal distribution of the opinion scores around MOS. It is interesting to point out, that the high kurtosis in the high quality end does not occur at the bad quality end. This means that the entire viewer panel agreed on the bad quality images with no outlier scores being present. This result is also evident in the skewness distribution where the decline towards lower values at the high quality end is much more pronounced as compared to the incline of the skewness at the low quality end.

3 Reduced-Reference Quality Metrics for Wireless Imaging

In this chapter, we discuss the design and validation of image quality metrics for deployment in wireless imaging systems. The metrics follow the RR approach (see Section 1.3), thus enabling to measure quality loss during transmission. The metrics are designed to measure the impact of the complex distortion patterns, as observed in wireless imaging, on the perceived visual quality. A low overhead in terms of RR information and a low computational complexity were further metric design issues.

We describe the development process of two feature-based image quality metrics, namely, the Normalised Hybrid Image Quality Metric (NHIQM) and the perceptual relevance weighted L_p-norm. The metrics are based on the extraction of a number of structural features related to the artifacts observed in wireless imaging applications. Both metrics follow the same design philosophy and mainly distinguish each other in the pooling of the features.

An overview of the RR visual quality assessment framework (see Fig. 4) adapted to the deployment of the feature-based metrics is shown in Fig. 10. Here, the features are extracted from the image both at the transmitter and at the receiver. In case of NHIQM, these features are combined in an additional pooling stage, whereas the pooling is omitted when using the L_p-norm. The RR information from the transmitter is communicated to the receiver either in-band as an additional header, in a dedicated control channel, or embedded in a watermark [114]. A difference computation between the reference and distorted RR information then constitutes a measure of distortions that have been induced during image communication. Curve fitting techniques are then deployed to relate the distortion measure to perceived visual quality.

In the following sections, the feature extraction and quality metric design are discussed in detail. Suitable mapping functions are additionally derived to establish a relationship between the objective artifact measures and the perceived quality degradation. Comparison of the designed quality metrics with other contemporary quality metrics reveals the superior quality prediction performance of the proposed metrics in the context of wireless imaging applications.

3.1 Structural feature extraction algorithms

Given the set of artifacts observed in the distorted images (see Section 2.1.3), algorithms for feature extraction are deployed to capture the amount by which

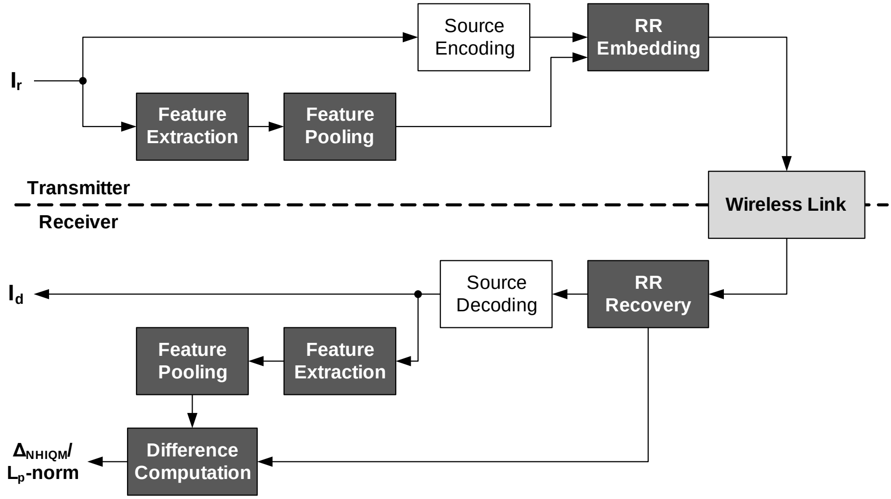


Figure 10: Reduced-reference quality assessment using NHIQM or the perceptual relevance weighted L_p-norm.

each of the artifacts is present in the images. The selection of the algorithms is driven by three constraints; a reasonable accuracy in capturing the characteristics of the associated artifact, a representation of the feature that incurs low overhead in terms of RR information (conserve bandwidth), and computational inexpensiveness (conserve battery power). Within these constraints we selected a number of feature extraction algorithms to measure and quantify the presence of the related artifacts. The feature metrics, along with the artifacts that they account for, are listed in Table 3 and are described in the following sections.

3.1.1 Feature $\tilde{f}_1$: block boundary differences

The first feature metric $\tilde{f}_1$ is based on the algorithm by Wang et al. [83] and comprises of three measures. The first measure, B, estimates blocking as average differences between block boundaries. Two image activity measures (IAM), A and Z, are applied as indirect means of quantifying blur. The former IAM computes absolute differences between in-block image samples and the latter IAM computes a zero-crossing rate. All three measures are computed in both horizontal and

Table 3: Image features, related artifacts, and feature extraction algorithms.

Feature		Related artifact	Ref.
$\tilde{f}_1$	Block boundary differences	Blocking	[83]
$\tilde{f}_2$	Edge smoothness	Blur	[204]
$\tilde{f}_3$	Edge-based image activity	Ringing	[205]
$\tilde{f}_4$	Gradient-based image activity	Ringing	[205]
$\tilde{f}_5$	Image histogram statistics	Block intensity shifts	[177]

vertical direction over the whole image and are combined in a pooling stage as

$$\tilde{f}_1 = \alpha + \beta B^{\gamma_1} A^{\gamma_2} Z^{\gamma_3} \tag{14}$$

where the parameters $\alpha = -245.9$, $\beta = 261.9$, $\gamma_1 = -0.024$, $\gamma_2 = 0.016$, and $\gamma_3 = 0.006$ were estimated in [83] using MOS from subjective experiments. Despite the two IAM incorporated in $\tilde{f}_1$, we found that this metric accounts particularly well for blocking artifacts in JPEG compressed images. This might be due to the magnitude of γ_1, being relatively large compared to γ_2 and γ_3, giving the blocking measure a higher impact on the metric $\tilde{f}_1$.

3.1.2 Feature $\tilde{f}_2$: edge smoothness

The extraction of feature metric $\tilde{f}_2$ relates purely to measuring blur artifacts and follows the work of Marziliano et al. [204]. It accounts for the smoothing effect of blur by measuring the distance between edges. It was found that it is sufficient to measure the blur along vertical edges as compared to computation on all edges, which allows for saving computational complexity. A Sobel filter is applied to detect vertical edges in the image and the edge image is then horizontally scanned. For pixels that correspond to an edge point, the local extrema in the corresponding image are used to compute the edge width. The edge width then defines a local measure of blur. Finally, a global blur measure is obtained by averaging the local blur values over all edge locations. This metric was chosen to complement the IAM in $\tilde{f}_1$ since it does not just account for in-block blur but rather contributes a global blur measure.

3.1.3 Features $\tilde{f}_3$ and $\tilde{f}_4$: image activity

Ringing artifacts are observed as periodic pseudo-edges around original edges, thus increasing the activity within an image. The feature metrics $\tilde{f}_3$ and $\tilde{f}_4$ provide an indirect means of measuring ringing artifacts and are based on two IAM by Saha and Vemuri [205].

Here, $\tilde{f}_3$ quantifies image activity (IA) based on normalised magnitudes of edges in an edge image $B(i, j)$ as

$$\tilde{f}_3 = \frac{100}{X \cdot Y} \sum_{x=1}^{X} \sum_{y=1}^{Y} B(x, y) \tag{15}$$

where X and Y denote the image dimensions. Since $\tilde{f}_3$ does not depend on the direction of the edge, it also very well complements the blocking measure in $\tilde{f}_1$, which is purely designed to measure on the 8×8 block boundaries in JPEG coded images.

On the other hand, $\tilde{f}_4$ measures IA in an image $I(x, y)$ based on local gradients in both vertical and horizontal direction as

$$\begin{aligned} \tilde{f}_4 = \frac{1}{X \cdot Y} \Bigg(& \sum_{x=1}^{X-1} \sum_{y=1}^{Y} |I(x, y) - I(x+1, y)| \\ & + \sum_{x=1}^{X} \sum_{y=1}^{Y-1} |I(x, y) - I(x, y+1)| \Bigg). \end{aligned} \tag{16}$$

In [205], the IAM were evaluated and $\tilde{f}_4$, in particular, has been found to quantify IA very accurately. We have further identified that both $\tilde{f}_3$ and $\tilde{f}_4$ account well for measuring ringing artifacts and also other high frequency changes within the image.

3.1.4 Feature $\tilde{f}_5$: image histogram statistics

Finally, feature metric $\tilde{f}_5$ accounts for block intensity shifts and lost blocks. Block intensity shifts may result in parts of the image or the whole image to appear either darker or brighter as compared to the original image. As such, we found that a simple computation of the standard deviation in the first-order image histogram provides an adequate measure of both block intensity shifts and lost blocks. We have thus deployed feature metric $\tilde{f}_5$ as

$$\tilde{f}_5 = \frac{1000}{X \cdot Y} \sqrt{\frac{1}{G} \sum_{g=0}^{G} (h_g - \overline{h})^2} \tag{17}$$

where h_g denotes the number of pixels at grey level g, $\overline{h}$ denotes the mean grey level, and G denotes the maximum grey level, here 255.

Feature metric $\tilde{f}_5$ is an adapted version of the algorithm by Kusuma [177]. In particular, the algorithm was modified to be less dependent on the image size by including a normalisation factor based on the number of pixels in the image.

3.2 Feature normalisation

The magnitudes of the different feature extraction algorithms $\tilde{f}_i$ are generally in very different ranges and thus, particular feature values may have a different meaning. As a consequence, weighting of the features to explore their perceptual relevance is not straightforward as the feature magnitude ranges inherently impact on the values of the feature weights. Therefore, we perform an extreme value normalisation [206] of the features which then allows for a more convenient and meaningful comparison of the contribution of each normalised feature f_i to the overall metric, as they are then taken from the same value range as

$$0 \leq f_i \leq 1. \tag{18}$$

Specifically, let us distinguish among I different image features. The related feature values $\tilde{f}_i$, $i = 1, 2, \ldots, I$, shall be normalised as

$$f_i = \frac{\tilde{f}_i - \min\limits_{k=1,2,..,K} (\tilde{f}_{i,k})}{\delta_i}, \qquad i = 1, 2, \ldots, I. \tag{19}$$

These features were extracted from all K images used in the subjective experiments E1 and E2, including all reference images and distorted images. Furthermore, the normalisation factor δ_i in (19) is given by

$$\delta_i = \max_{k=1,2,\ldots,K} (\tilde{f}_{i,k}) - \min_{k=1,2,\ldots,K} (\tilde{f}_{i,k}). \tag{20}$$

As far as the extreme value normalised features defined by (19) are concerned, it should be mentioned that the boundary conditions apply to those normalised feature values $f_{i,k}$ which are associated with the feature values $\tilde{f}_{i,k}$ of the images

Table 4: Relative contribution of the feature metrics to the overall computational cost of the quality metrics.

Feature metric	f_1	f_2	f_3	f_4	f_5
Computational cost	15.44 %	47.41 %	29.28 %	7.73 %	0.14 %

used in the experiments. In a practical system, it may also be beneficial to clip the normalised feature values that are actually calculated in a real-time wireless imaging application to fall in the interval $[0, 1]$ as well. For instance, severe signal fading in a wireless channel can result in significant image impairments at particular times causing the user-perceived quality to fall in a region where the HVS is saturated to notice further degradation.

3.3 Computational cost of the feature metrics

The computational complexity of the image quality metrics presented in this chapter is of particular concern in the context of wireless image communications. The relative burden of each of the feature extraction algorithms on the computational complexity of the quality metric is thus of great interest. Therefore, we present in Table 4 the relative contribution of each feature metric to the computational cost of the overall quality metric. The table shows the percentage of computation time of each feature metric in relation to the overall computation time of the quality metric.

It can be observed that feature f_2 takes up about half of the computation time of the quality metric. This is followed by feature f_3, which contributes almost a third of the computational complexity. Feature f_1 exhibits about half of the cost of f_3 and feature f_4 has approximately half of the complexity of feature f_1. These features all have a considerable contribution to the computational complexity of the quality metric. This is different for feature f_5, which impacts only marginally on the overall computational cost.

3.4 Feature metrics performance analysis

In order to gain deeper knowledge and understanding about the feature extraction, we examine the extent to which different features are present in the stimuli and quantify a relationship between the feature metrics and MOS. Given the context of RR metric design in wireless imaging, where we are interested in the difference between the quality of the received image as compared to the quality of the transmitted image, we consider in the following the magnitude of normalised feature differences

$$\Delta f_i = |f_{r,i} - f_{d,i}|, \qquad i = 1, 2, \ldots, 5 \tag{21}$$

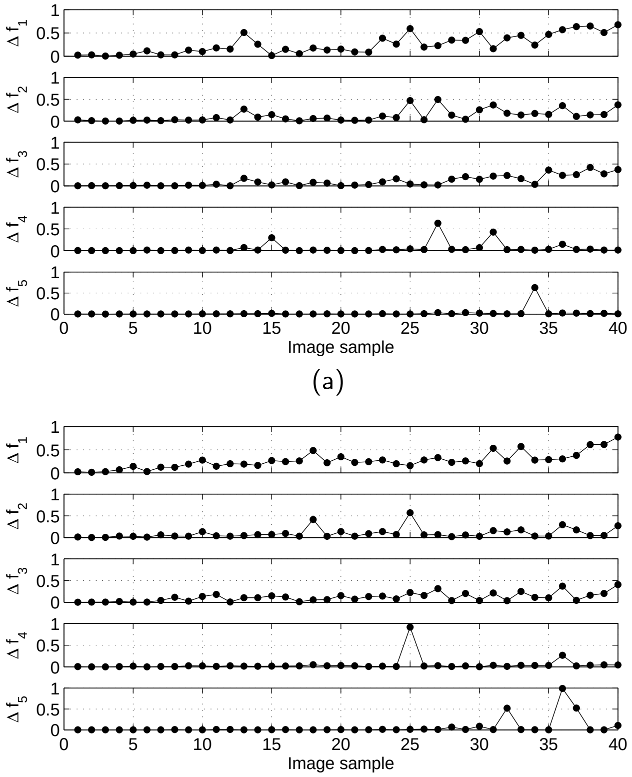
where $f_{r,i}$ and $f_{d,i}$ denote the i^{th} feature value of the reference image and the distorted image, respectively.

3.4.1 Feature magnitudes over MOS

Figures 11(a)-(b) show the magnitudes of the normalised feature differences Δf_i for the image samples that were presented in E1 and E2. For each experiment, the related 40 feature differences are ranked with respect to image samples of decreasing MOS (see the MOS in Fig. 8). It can be seen from Fig. 11(a)-(b) that the wireless link scenario indeed induced all five features but with different degrees of severity. While feature differences are almost absent for the image samples of high perceptual quality ratings, the feature differences tend to increase with decreasing MOS. Especially the level of Δf_1, relating to blocking contained in the image samples, shows the widest spread and becomes more pronounced when progressing from images of excellent to bad perceptual quality. A similar behavior is observed for edge-based image activity Δf_3 but appears not as pronounced as for Δf_1. As far as the remaining three features are concerned, these become less prevalent for most of the images but large for some of the stimuli. In particular, gradient-based image activity Δf_4 and intensity masking Δf_5 occur very distinctively with selected image samples while being almost absent from the majority of image samples.

3.4.2 Feature statistics

As with the MOS gathered from the subjective experiments, the statistical analysis may be extended to the actual feature differences in order to obtain a better understanding of the underlying objective quality degradations. However, overall statistics for the whole set of data, instead of image dedicated statistics, are presented hereafter. For all five feature differences Δf_i, the mean, variance, skewness, and kurtosis have been computed over all images that have been shown in experiments E1 and E2. The results of all statistics are presented for both experiments in Tables 5 and 6.



(a)

(b)

Figure 11: Magnitude of differences between the normalised features for all images ranked according to decreasing MOS: (a) E1 and (b) E2.

From comparison of the two tables one can observe that for all four statistics and for all five feature differences, the magnitudes of the values are very much in alignment between the two experiments E1 and E2. This indicates that the stimuli, in terms of the distorted test images, had similar characteristics in both experiments. Thus, not only the subjective data is in alignment but also the composition of objective features among the test material. In particular, it can be seen from both tables that the mean of the blocking differences dominates over the other features. This is a direct result of the JPEG source encoding of

Table 5: Statistics of magnitudes of feature differences Δf_i for E1.

	Δf_1	Δf_2	Δf_3	Δf_4	Δf_5
Mean	0.253	0.120	0.102	0.053	0.022
Variance	0.043	0.017	0.014	0.015	0.009
Skewness	0.627	1.425	1.124	3.518	6.015
Kurtosis	2.082	4.120	3.241	15.010	37.466

Table 6: Statistics of magnitudes of feature differences Δf_i for E2.

	Δf_1	Δf_2	Δf_3	Δf_4	Δf_5
Mean	0.263	0.094	0.115	0.049	0.061
Variance	0.029	0.013	0.010	0.021	0.035
Skewness	1.066	2.495	1.072	5.461	3.785
Kurtosis	4.056	9.531	3.843	32.434	17.063

which it is well known that blocking artifacts are dominant over other artifacts such as blur. The mean values of feature differences Δf_4 and Δf_5 are particularly small, however, these features exhibit instead a very high skewness and kurtosis as compared to the other features. Clearly, this quantifies the progression of feature differences in the stimuli as shown in Fig. 11(a)-(b) with Δf_4 and Δf_5 being either negligibly small or distinctively large.

3.4.3 Feature cross-correlations

Even though the feature metrics were selected to account for a particular artifact, one may expect some overlap in quantifying the different artifacts. To further understand the performance of the feature metrics in comparison to each other, Tables 7 and 8 show the Pearson linear correlation coefficient between each of the feature metrics for both E1 and E2, respectively. In this context, the cross-correlation measures the degree to which two features are simultaneously affected by a certain type and severity of an artifact. As expected, the auto-correlation of a feature with itself exhibits the maximum magnitude of 1.

It can be seen from the tables that the cross-correlations between the features vary strongly in their magnitudes. A particularly pronounced cross-correlation can

Table 7: Correlations between feature differences for E1.

	Δf_1	Δf_2	Δf_3	Δf_4	Δf_5
Δf_1	1.000	0.625	0.821	0.016	0.027
Δf_2		1.000	0.440	0.649	0.112
Δf_3			1.000	0.056	-0.061
Δf_4				1.000	0.000
Δf_5					1.000

Table 8: Correlations between feature differences for E2.

	Δf_1	Δf_2	Δf_3	Δf_4	Δf_5
Δf_1	1.000	0.376	0.640	-0.014	0.115
Δf_2		1.000	0.486	0.753	0.316
Δf_3			1.000	0.323	-0.272
Δf_4				1.000	0.170
Δf_5					1.000

be observed between feature metrics Δf_1 (block boundary differences) and Δf_3 (edge-based IA) for both E1 and E2. This is thought to be due to both metrics being based on measuring edges of an image. However, it should be noted again that feature metric Δf_1 only considers the 8×8 block borders of the JPEG encoding whereas feature metric Δf_3 quantifies image activity based on edges in all spatial locations and directions. Furthermore, feature metrics Δf_2 (edge smoothness) and Δf_4 (gradient-based IA) exhibit pronounced cross-correlations in the test sets of both experiments which may be a result of both metrics being designed to quantify smoothness in images based on gradient information. As for feature metric Δf_5 (image histogram statistics), it can be seen that this metric is only negligibly correlated to any of the other feature metrics. This is a highly desired result since the feature metrics other than Δf_5 should be widely unaffected by intensity shifts.

3.5 Reduced-reference quality metric design

In the following sections, we describe in detail the RR quality metric design which is based on the feature extraction algorithms outlined in Section 3.1. In this

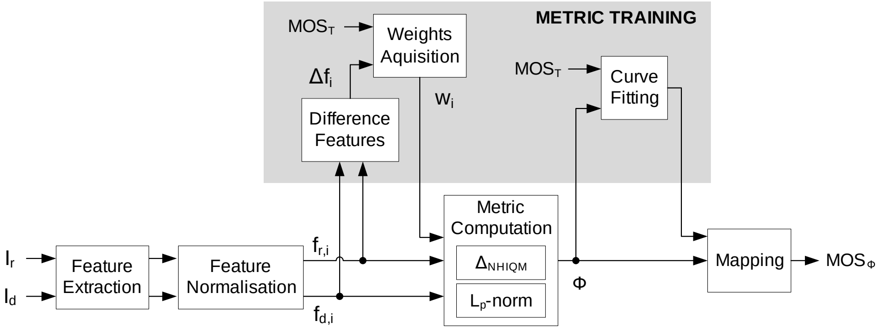


Figure 12: Overview of the reduced-reference feature-based image quality metric design. The blocks outside the grey area constitute the integral parts of the quality metric whereas the blocks inside the grey area are only deployed during metric training.

respect, the quality ratings obtained in the subjective experiments are instrumental for the transition from subjective to objective quality assessment.

3.5.1 Metric training and validation

As foundation of the metric design, the 80 images in $\mathcal{I}_1$ (E1) and $\mathcal{I}_2$ (E2) from the two experiments were organised into a training set $\mathcal{I}_T$ containing 60 images and a validation set $\mathcal{I}_V$ containing 20 images. For this purpose, 30 images were taken from $\mathcal{I}_1$ and 30 images from $\mathcal{I}_2$ to form $\mathcal{I}_T$ while the remaining 10 images of each set compose $\mathcal{I}_V$. Accordingly, a training set and a validation set were established with the corresponding MOS, here referred to as MOS$_T$ and MOS$_V$. The training sets, $\mathcal{I}_T$ and MOS$_T$, are then used for the actual metric design. The validation sets, $\mathcal{I}_V$ and MOS$_V$, are used to evaluate the metrics ability to generalise to unknown images.

3.5.2 Metric design overview

The different parts of the RR perceptual quality metric design are shown in the block diagram in Fig. 12. Here, the blocks outside the grey area constitute the integral parts of the perceptual image quality metric, whereas the blocks enclosed by the grey area are deployed only during metric training. A brief summary of the design process is given in the sequel with reference to this figure.

The first key operation in the transition from subjective to objective perceptual image quality assessment is executed within the process of feature weights acquisition. As a prerequisite of weights acquisition, the different features of the transmitted and received image are extreme value normalised to allow for a meaningful weight association. As the RR design is focused on detecting distortions between the reference image and the distorted image, the weights acquisition is performed with respect to the feature differences Δf_i, $i = 1, 2, \ldots, 5$. Given the MOS values MOS_T for the images in the training set $\mathcal{I}_T$ and the related feature differences Δf_i for each image, correlation coefficients between subjective ratings and feature differences are computed as weights w_i, $i = 1, 2, \ldots, 5$, to reveal the relevance of each feature to the subjectively perceived quality. It is then straightforward to compute a feature-based image quality metric Φ by applying a pooling function to condense the information. Here, two metrics are proposed, namely Δ_{NHIQM} and the perceptual relevance weighted L_p-norm. The former metric is particularly efficient in reducing the RR information into a single value before transmission. The latter metric consolidates the features at the receiver and therefore allows for tracking of individual artifacts at the cost of larger RR information.

The second essential component in moving from subjective to objective quality assessment relates to the curve fitting block as shown in Fig. 12. Its inputs are the MOS values MOS_T for the images in the training set $\mathcal{I}_T$ and the values of the perceptual quality metric Φ for each of these images. The relationship between subjective quality given by MOS_T and objective quality represented by Φ, is then modeled by a suitable mapping function $f(\Phi)$. The parameters of potential mapping functions can be obtained by using standard curve fitting techniques. The selection of suitable mapping functions is typically based on both goodness of fit measures and visual inspection of the fitted curve. The obtained mapping function $f(\Phi)$ can then be used to calculate predicted MOS values, MOS_Φ, for given values of the quality metric Φ.

3.5.3 Perceptual relevance of features

The Pearson linear correlation coefficient ρ_P has been chosen to reveal the extent by which the individual feature differences contribute to the overall perception of image quality. In this sense, it captures prediction accuracy referring here to the ability of a feature difference to predict the subjective ratings with minimum average error. Given a set of K data pairs (u_k, v_k), the Pearson correlation is given by

$$\rho_P = \frac{\sum\limits_{k=1}^{K}(u_k - \bar{u})(v_k - \bar{v})}{\sqrt{\sum\limits_{k=1}^{K}(u_k - \bar{u})^2}\sqrt{\sum\limits_{k=1}^{K}(v_k - \bar{v})^2}} \tag{22}$$

where u_k and v_k are the feature difference and the subjective rating related to the k^{th} image, respectively, and $\bar{u}$ and $\bar{v}$ are the means of the respective data sets.

This choice is motivated by the fact that the correlation coefficient explicitly characterises the association between two variables, which are given here by pairs of ratings and difference feature metrics. The sign of the correlation value may be neglected as it only represents the direction (increase/decrease) in which one variable changes with the change of the other variable. In view of the above, the absolute values of the Pearson linear correlation coefficients ρ_P are computed as the perceptual weights w_i of the related features. A higher correlation coefficient then corresponds to a feature that more significantly contributes to the overall quality as perceived by the viewer, while a lower correlation coefficient means less perceptual significance. Also, if the correlation coefficient approaches the zero value, the relationship between the perceptual quality and the examined feature is not strongly developed.

Table 9 shows the values of the Pearson linear correlation coefficients, or feature weights, that were obtained for each of the five feature differences Δf_i, $i = 1, 2, \ldots, 5$, for the training set when correlated to the associated MOS$_T$ values. Accordingly, block boundary differences (Δf_1) appear to be the most relevant feature followed by edge-based image activity (Δf_3), edge smoothness (Δf_2), image histogram statistics (Δf_5), and gradient-based image activity (Δf_4). This indicates that blocking is the most annoying artifact followed by ringing due to edge-based image activity, blur, block intensity shifts, and ringing due to gradient-based image activity. Similar findings have also been made by Farias et al. [207] who observed that blocking is more annoying than blur. The same group also found [208] that ringing is the least annoying artifact. This agrees with our feature metric Δf_4 which also received the smallest weight. On the other hand, the feature metric Δf_3 deployed here measures ringing as well but received a higher weight. We believe that this outcome can be related to Δf_3 having a strong correlation with Δf_1 (see Tables 7 and 8), thus not only accounting for ringing but also for blocking artifacts.

It should be noted here that the relevance weights in Table 9 were obtained for

Table 9: Perceptual relevance weights of feature differences Δf_i for the images in the training set.

Metric	Δf_1	Δf_2	Δf_3	Δf_4	Δf_5
Weight	w_1	w_2	w_3	w_4	w_5
Value	0.819	0.413	0.751	0.182	0.385

the particular case of JPEG source encoding where blocking artifacts are predominant over other artifacts such as blur. This may also contribute to the higher correlation weights for the edge-based features Δf_1 and Δf_3 as compared to the gradient-based features Δf_2 and Δf_4. Hence, the relevance weights may not be purely related to the perceptual relevance but also to the particular artifacts that are observed in the visual content. As such, one may obtain different relevance weights in case of other source encoders, such as JPEG2000. To derive these weights, appropriate JPEG2000 encoded image sets and corresponding MOS from subjective experiments would be needed, which is outside the scope of this thesis.

3.5.4 Normalised Hybrid Image Quality Metric

Unlike the Hybrid Image Quality Metric (HIQM) as described in [177], the Normalised Hybrid Image Quality Metric (NHIQM) proposed here for feature pooling uses the normalised features and the respective perceptual relevance weights instead. The NHIQM metric is defined as a weighted sum of the extreme value normalised features as

$$NHIQM = \sum_{i=1}^{I} w_i f_i \tag{23}$$

where w_i denotes the relevance weight of the associated feature f_i. Clearly, this RR metric is particularly beneficial for objective perceptual quality assessment in wireless imaging, as the RR information is represented by only one single value for a given image. Accordingly, NHIQM can be communicated from the transmitter to the receiver whilst imposing very little stress on the bandwidth resources.

In order to measure quality degradations during image communication, NHIQM can be calculated for both the reference image I_r (the transmitted image) and the possibly distorted image I_d (the received image), resulting in the corresponding values $NHIQM_r$ and $NHIQM_d$ at the transmitter and receiver, respectively.

Provided that the $NHIQM_r$ value is communicated to the receiver, structural differences between the images at both ends can then simply be represented by the absolute difference

$$\Delta_{NHIQM} = |NHIQM_r - NHIQM_d|. \tag{24}$$

Thus, not only the absolute quality of the received image can be measured but also any quality loss that occurred during image transmission.

3.5.5 Perceptual relevance weighted L_p-norm

In the Δ_{NHIQM} pooling step, all considered features are combined in a single value. This is advantageous with respect to RR overhead, however, valuable information may get lost about the degradation of each of the features deployed in the metric. For this purpose, we consider a second pooling method, the L_p-norm, also referred to as Minkowski metric. The L_p-norm is a distance measure commonly used to quantify similarity between two signals or vectors. In image processing it has been applied, for instance, with the percentage scaling method [209] and the combining of impairments in digital image coding [210]. The Minkowski summation has further been deployed as a pooling function representing a good trade-off between performance and complexity [211].

In this paper, we incorporate the relevance weighting for the extreme value normalised features into the calculation of the L_p-norm. This modification of the L_p-norm is defined as

$$L_p = \left[\sum_{i=1}^{I} w_i^p |f_{r,i} - f_{d,i}|^p \right]^{\frac{1}{p}} \tag{25}$$

where $f_{r,i}$ and $f_{d,i}$ denote the i^{th} feature value of the reference and the distorted image, respectively.

The Minkowski exponent p can be determined experimentally [209] or, alternatively, the Minkowski exponent p can be assigned a fixed value. In both cases, a higher value of p increases the impact of the dominant features on the overall metric. In the limit of p approaching infinity, we obtain

$$L_\infty = \max_{i=1,2,..,I} |f_{r,i} - f_{d,i}| \tag{26}$$

meaning that the largest absolute feature value difference solely dominates the norm. We found [178] that values beyond $p = 2$ do not improve the quality prediction performance of the modified L_p-norm given in (25). We believe that this characteristic is because of the perceptual relevance weights obtained for each feature inherently accounting for the dominance of the particular features. In the sequel, we therefore consider the modified L_p-norm for Minkowski exponents of $p = 1$ and $p = 2$ only.

Although the perceptual relevance weighted L_p-norm belongs to the class of RR metrics, it requires more transmission resources compared to Δ_{NHIQM}, as all feature values considered in the metric need to be communicated from the transmitter to the receiver. On the other hand, the information about each of the feature degradations may provide further insights into the channel induced distortions. Hence, overhead may be traded off at the expense of a reduction about structural degradation information by neglecting feature metrics that received low perceptual relevance weights.

3.5.6 Mapping to predicted MOS

In a final step, a regression analysis is performed to find suitable prediction functions to map the image quality metrics, Φ, onto predicted MOS, MOS$_\Phi$. This procedure inherently serves two important purposes. Firstly, the prediction function maps the range of a quality metric onto the range of the subjective scores from the experiment, facilitating prediction of the subjective scores. This is an important step as, typically, different metrics create predictions in different value ranges and as such, the actual metric value may not have much meaning if it is not put in relation to subjective scores.

Secondly, due to non-linear quality processing in the HVS, measured artifacts and perceived quality do not necessarily follow a linear relationship. To be more precise, within the suprathreshold regime of artifacts, human observers tend to make a more pronounced distinction between two quality levels of weakly distorted images as compared to two quality levels of two strongly distorted images. To account for this phenomenon, a mapping function is applied to translate a perceptual quality metric Φ into predicted MOS, MOS$_\Phi$, as follows:

$$\text{MOS}_\Phi = f(\Phi). \tag{27}$$

Within the scope of this thesis, we take into account different classes of mapping functions that we consider as possible candidates for the mapping to predicted MOS. In particular, we consider the following four classes of mapping functions:

$$\mathrm{MOS}_{\Phi} \quad \triangleq \quad \begin{cases} \sum\limits_{j=0}^{m} p_j \cdot \Phi^j & \text{Polynomial} \\ \sum\limits_{j=0}^{m} u_j \cdot e^{v_j \Phi} & \text{Exponential} \\ \frac{l_1}{1+e^{-l_2(\Phi - l_3)}} & \text{Logistic} \\ k_1 \cdot \Phi^{k_2} + k_3 & \text{Power} \end{cases} \tag{28}$$

where the parameters of the prediction functions are to be determined through curve fitting based on the experimental data from the training set. These four classes of mapping functions have been chosen as candidates for quality prediction due to the following reasons:

- **Polynomial functions** provide sufficient flexibility to support simple empirical prediction.
- **Exponential and power functions** are imposed to enable a good fit to experimental data over the middle-to-upper range of the quality impairment measure [59] and may be less prone to overfitting compared to functions with many parameters, such as higher order polynomials.
- **Logistic functions** facilitate the mapping of quality impairment measures into a finite interval. They produce scale compressions at the high and low extremes of quality while progressing approximately linear in the range between these extremes.

Standard curve fitting techniques have been used to deduce the parameters of the mapping functions that mathematically describe best the relationship between subjective ratings and perceptual quality metric with respect to a given goodness of fit measure. The goodness of fit between MOS and predicted MOS can be specified by either of the following statistics:

- **Squared correlation coefficient $\mathbf{R^2}$** captures the degree by which variations in the MOS values are accounted for by the fit. It can assume any value in the interval $[0, 1]$ with a good fit being close to 1.
- **Root mean squared error (RMSE)** is referred to as the standard error of the fit, with a better fit indicated by a smaller RMSE.

- **Sum of squared errors (SSE)** represents the total deviation between predicted MOS and MOS from the experiments. The smaller the SSE value, the better the fit.

The Matlab Curve Fitting Toolbox was used to find the parameters of the considered mapping functions. The mapping functions have been derived for both Δ_{NHIQM} and the perceptual relevance weighted L_p-norm, however, as they exhibited very similar properties, detailed results are in the following only presented for Δ_{NHIQM}. The results are provided in Table 10 along with the different goodness of fit measures. A visual examination of the fitted mapping functions is supported by Fig. 13-16, which also show the 95% CI for each fit. It should be highlighted here again, that a larger value of Δ_{NHIQM} corresponds to a stronger quality degradation and thus, to a lower MOS value.

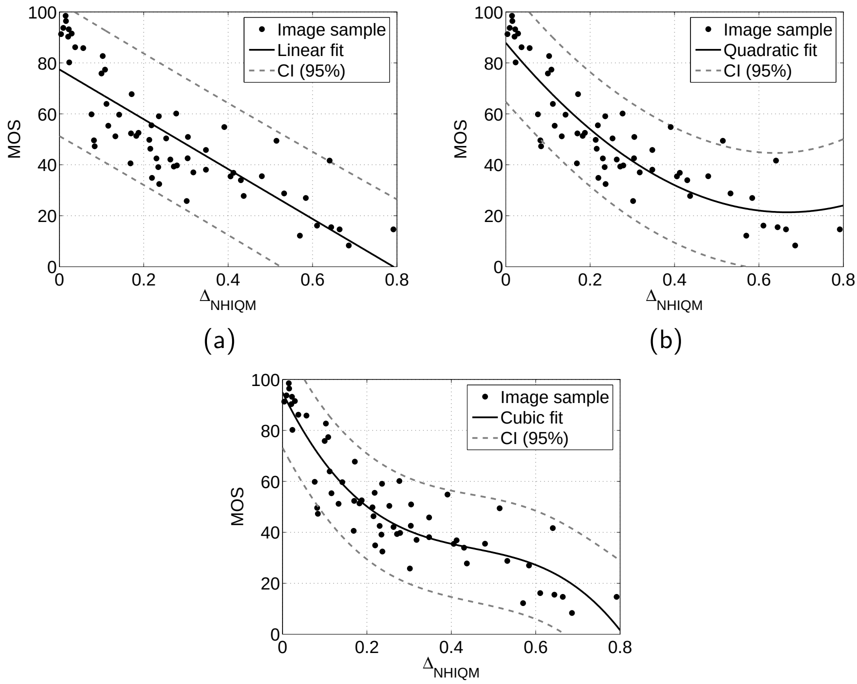
As far as the polynomial functions are concerned, it can be seen from Table 10 that the linear polynomial results in poor goodness of fit measures, as it does not take into account the non-linearity of quality processing in the HVS. Regarding the quadratic and cubic polynomials, it could be concluded at first sight, when looking only at the goodness of fit statistics, that both of them perform similarly well as the exponential functions. However, visual inspection of Fig. 13 suggests the opposite, as the good fit applies only for the given data range but tends to diverge outside this range. An increase of the perceptual quality metric beyond the value of 0.8 would actually increase the predicted MOS again in case of the quadratic polynomial (see Fig. 13 (b)) and would predict 'negative' MOS values in case of the cubic polynomial (see Fig. 13 (c)). As higher-degree polynomials may result in even more severe overfitting, the class of polynomials is in the remainder of this chapter not considered anymore as a suitable mapping function.

In contrast to the polynomial functions, favorable fitting has been obtained for all three considered exponential mapping functions, not only in terms of goodness of fit measures but also confirmed by visual inspection (see Fig. 14). However, it can be observed that the triple exponential function performs similarly to the exponential function but at the price of a larger computational complexity due to its more involved analytical expression. As such, the triple exponential function is not considered further.

As for the logistic mapping function, the goodness of fit measures indicate a rather poor fit to the data from the subjective experiments. Especially, the compression at the high end of the quality scale produces disagreement with the MOS (see Fig. 15). The power function was found to behave very similarly to the double exponential function, which is apparent both in the goodness of fit measures and also through visual inspection of the fitting curve (see Fig. 16).

Table 10: Mapping functions $f(\Phi)$ = MOS$_{NHIQM}$, $\Phi = \Delta_{NHIQM}$, and their goodness of fit.

Type	Function	Parameters	R^2	RMSE	SSE
Polynomial	$p_1\Phi+p_0$	$p_1=-97.8$ $p_0=77.45$	0.71	12.78	9472
	$p_2\Phi^2+p_1\Phi+p_0$	$p_2=149.5$ $p_1=-199.4$ $p_0=87.88$	0.79	11.07	6982
	$p_3\Phi^3+p_2\Phi^2+p_1\Phi+p_0$	$p_3=-493.9$ $p_2=672.2$ $p_1=-338.3$ $p_0=94.87$	0.82	10.17	5792
Exponential	$u_1 e^{v_1\Phi}$	$u_1=88.79$ $v_2=-2.484$	0.79	10.76	6714
	$u_1 e^{v_1\Phi}+u_2 e^{v_2\Phi}$	$u_1=69.76$ $v_1=-1.719$ $u_2=32.05$ $v_2=-17.39$	0.83	10.01	5612
	$u_1 e^{v_1\Phi}+u_2 e^{v_2\Phi}+u_3 e^{v_3\Phi}$	$u_1=63.18$ $v_1=-3.056$ $u_2=-175$ $v_2=0.1434$ $u_3=198.2$ $v_3=0.041$	0.80	11.12	6678
Logistic	$l_1/[1+e^{-l_2(\Phi-l_3)}]$	$l_1=100$ $l_2=-4.613$ $l_3=0.262$	0.72	12.63	9263
Power	$k_1\cdot\Phi^{k_2}+k_3$	$k_1=-120.5$ $k_2=0.28$ $k_3=128.2$	0.82	10.06	5764



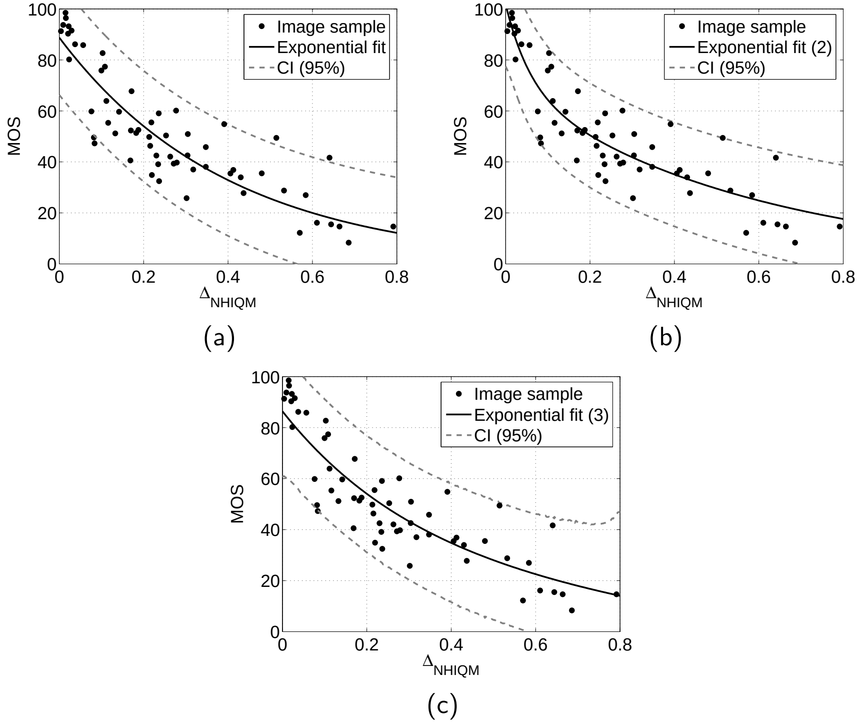
(a)

(b)

(c)

Figure 13: Polynomial mapping functions: (a) linear, (b) quadratic, and (c) cubic.

Due to the strong performance of both the double exponential and the power function, we considered these mapping functions as strong candidates for the mapping to predicted MOS. However, an analysis of the mapped metric values for the validation set revealed that the prediction performance did not match the performance on the training set and was in fact rather poor. As such, the double exponential and the power fit may to some degree introduce overfitting. For this reason, we neglect these mapping functions and consider the exponential function as the best compromise with respect to the goodness of fit measures and the resulting metric's ability to generalise to the validation set. Similar observations were also made for the perceptual relevance weighted L_1-norm and L_2-norm.



(a)

(b)

(c)

Figure 14: Exponential mapping functions: (a) exponential, (b) double exponential, and (c) triple exponential.

3.6 Quality prediction performance

In the following sections, the quality prediction performance of Δ_{NHIQM} and the perceptual relevance L_p-norm is analysed in detail for both, the actual metric values and also the corresponding predicted MOS. The metrics' quality prediction performance is further compared to the performance of a selection of state-of-the-art image quality metrics.

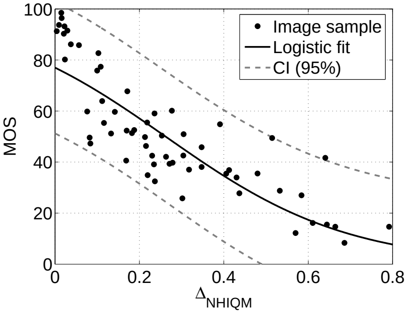


Figure 15: Logistic mapping function.

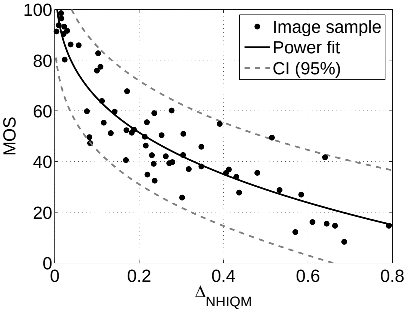


Figure 16: Power mapping function.

3.6.1 Image quality metrics for performance comparison

We selected contemporary and widely used image quality metrics for a performance comparison with the proposed feature-based Δ_{NHIQM} and the L_p-norm. Specifically, the reduced-reference image quality assessment (RRIQA) technique proposed in [106] is chosen as a prominent member of the class of RR metrics. From the class of FR metrics we chose the structural similarity (SSIM) index [14], the visual information fidelity (VIF) criterion [73], the visual signal-to-noise ratio (VSNR) [76], and the peak signal-to-noise ratio (PSNR) [212] to be used for performance comparison. It is noted that FR metrics would not be suitable for the considered image communication scenario but rather serve to benchmark prediction performance, which can be expected to be high due to the utilisation of the reference image. A brief summary of each of the metrics is given in the following.

RRIQA: This metric [106] is based on a natural image statistic model in the wavelet domain. The image distortion measure is obtained from the estimation of the KLD between the marginal probability densities of wavelet coefficients in the subbands of the reference and distorted images as

$$\mathsf{D} = \log_2\left(1 + \frac{1}{D_0}\sum_{k=1}^{K}|\hat{d}^k(p^k\|q^k)|\right) \tag{29}$$

where the constant D_0 is used as a scaler of the distortion measure, $\hat{d}^k(p^k\|q^k)$ denotes the estimation of the KLD between the probability density functions p^k and q^k of the k^{th} subband in the transmitted and received image, and K is the number of subbands. The overhead needed to represent the RR information is given as 162 bits [106].

SSIM: The SSIM index [14] is based on the assumption that the HVS is highly adapted to the extraction of structural information from the visual scene. As such, SSIM predicts structural degradations between two images based on simple intensity and contrast measures. The final SSIM index is given by

$$\mathsf{SSIM}(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{30}$$

where μ_x, μ_y and σ_x, σ_y, denote the mean intensity and contrast of image signals x and y, respectively, and σ_{xy} denotes the covariance between x and y. The constants C_1 and C_2 are used to avoid instabilities in the structural similarity comparison that may occur for certain mean intensity and contrast combinations ($\mu_x^2 + \mu_y^2 = 0$, $\sigma_x^2 + \sigma_y^2 = 0$).

VIF: The VIF criterion [73] approaches the image quality assessment problem from an information theoretical point of view. In particular, the degradation of visual quality due to a distortion process is measured by quantifying the information available in a reference image and the amount of this reference information that can be still extracted from the distorted image. As such, the VIF criterion measures the loss of information between two images. For this purpose, natural scene statistics and, in particular, Gaussian scale mixtures (GSM) in the wavelet domain, are used to model the images. The proposed VIF metric is given by

$$\mathsf{VIF} = \frac{\sum_{j \in subbands} I(\overrightarrow{C}^{N,j}; \overrightarrow{F}^{N,j} | s^{N,j})}{\sum_{j \in subbands} I(\overrightarrow{C}^{N,j}; \overrightarrow{E}^{N,j} | s^{N,j})} \tag{31}$$

where $I(\cdot)$ represents the mutual information, $\overrightarrow{C}$ denotes the GSM, N denotes the number of GSM used, s is a random field of positive scalars, and $\overrightarrow{E}$ and $\overrightarrow{F}$ denote the visual output of a HVS model, respectively, for the reference and distorted image.

VSNR: The VSNR [76] metric deploys a two-stage approach based on near-threshold and suprathreshold properties of the HVS to quantify image fidelity. The first stage determines whether distortions are visible in an image. For this purpose, contrast thresholds for distortion detection are determined using wavelet-based models of visual masking. If the distortions are below the threshold, the quality of the image is assumed to be perfect and the algorithm is terminated. If the distortions are visible, a second stage implements perceived contrast and global precedence properties of the HVS to determine the impact of the distortions on perceived quality. The final VSNR metric is then given as

$$\mathsf{VSNR} = 20 \ \log_{10}\left(\frac{C(\mathbf{I})}{\alpha \ d_{pc} + (1-\alpha)\frac{d_{gp}}{\sqrt{2}}}\right) \tag{32}$$

where $C(\mathbf{I})$ denotes the root-mean-squared contrast of the original image $\mathbf{I}$, d_{pc} and d_{gp} are, respectively, measures of perceived contrast and global precedence disruption, and α is a weight regulating the relative contributions of d_{pc} and d_{gp}.

PSNR: Image fidelity is an indication about the similarity between the reference and distorted images and measures pixel-by-pixel closeness between those pairs. The PSNR [212] is the most commonly used fidelity metric. It measures the fidelity difference of two image signals $I_r(x,y)$ and $I_d(x,y)$ on a pixel-by-pixel basis as

$$\mathsf{PSNR} = 10 \log \frac{\eta^2}{\mathsf{MSE}} \tag{33}$$

where η is the maximum pixel value, here 255. The MSE is given as

$$\mathsf{MSE} = \frac{1}{XY} \sum_{x=1}^{X} \sum_{y=1}^{Y} [I_r(x,y) - I_d(x,y)]^2 \tag{34}$$

Table 11: Computational complexity of the quality metrics and amount of reference information needed.

Metric		Computation	Reference
Type	Name	time/image	information
RR	Δ_{NHIQM}	1.55 s	17 bits
	L_p-norm	1.55 s	85 bits
	RRIQA	7.12 s	162 bits
FR	SSIM	0.37 s	Full image
	VIF	0.92 s	Full image
	VSNR	0.33 s	Full image
	PSNR	0.05 s	Full image

where X and Y denote horizontal and vertical image dimensions, respectively. Despite being an FR metric, PSNR usually does not correlate well with the visual quality as perceived by a human observer [10] as discussed in Section 1.1.

3.6.2 Computational complexity and amount of reference information

In the following, we discuss the computational complexity of the considered metrics and the amount of reference information that is needed in order to assess the quality of a test image. The details are summarised in Table 11.

The computational complexity is measured in terms of the time that each of the metrics needs to assess the quality of a single image in our sets $\mathcal{I}_1$ and $\mathcal{I}_2$. Here, we have computed each metric over all 80 images and then determined the average time. The metrics were run on a laptop computer containing an Intel T2600 Dual Core processor with 2.16 GHz and 4 GB of RAM. In order to allow for a fair comparison, the publicly available Matlab implementation of each metric was used even though there may be other implementations available for some of the metrics. It can be seen from Table 11 that the computational complexity of all FR metrics is lower as compared to the RR metrics. Amongst the FR metrics, PSNR outperforms by far the other considered metrics in terms of computational complexity. Regarding the RR metrics, it is observed that both Δ_{NHIQM} and L_p-norm are significantly less complex than RRIQA.

In the context of wireless imaging, the amount of reference information needed for quality assessment determines the overhead of data that needs to be transmitted over the channel along with the actual image. From Table 11 one can see that the reference information is significantly lower for both Δ_{NHIQM} and L_p-norm as compared to RRIQA. The particularly small reference information for Δ_{NHIQM} results from the fact that only a single value $NHIQM_r$ needs to be transmitted. On the other hand, with the L_p-norm five features need to be transmitted resulting in a five times higher overhead. However, as discussed in Section 3.5.5, the number of features used may be traded off with the transmission overhead by neglecting features of low perceptual relevance. As for the FR metrics, the reference image is needed for the quality assessment and as such, the size of the image determines the amount of reference information. Independent of the image size, however, the amount of reference information would be magnitudes higher as compared to the RR metrics.

3.6.3 Prediction performance indicators

To quantify the quality prediction performance of the metrics, we follow the recommendations by the VQEG [200,202], which define the following three quality prediction performance indicators:

- **Prediction accuracy**: the ability of a quality metric to predict subjective quality ratings with low error.
- **Prediction monotonicity**: the degree to which the objective quality predictions agree with the relative magnitudes of the subjective quality ratings.
- **Prediction consistency**: the degree to which the quality metric maintains prediction accuracy over the range of test images, thus, revealing the metrics robustness to different image content and a variety of artifacts.

As recommended in [200, 202], the prediction accuracy is determined using the the Pearson linear correlation coefficient ρ_P (see Eq. (22)) and the root mean squared error (RMSE), which is given as

$$RMSE = \sqrt{\frac{1}{K-d}\sum_{k=1}^{K}\epsilon^2(k)} \tag{35}$$

with K denoting the number of images, d representing the number of degrees of freedom (the number of coefficients) in the mapping function, and $\epsilon(k)$ being the prediction error between MOS and predicted MOS computed as

$$\epsilon(k) = \mathrm{MOS}(k) - \mathrm{MOS}_{\Phi}(k). \tag{36}$$

The Spearman rank order correlation coefficient ρ_S is adopted as a measure of prediction monotonicity [202] as follows:

$$\rho_S = \frac{\sum\limits_{k=1}^{K}(\chi_k - \bar{\chi})(\gamma_k - \bar{\gamma})}{\sqrt{\sum\limits_{k=1}^{K}(\chi_k - \bar{\chi})^2}\sqrt{\sum\limits_{k=1}^{K}(\gamma_k - \bar{\gamma})^2}} \tag{37}$$

where χ_k and γ_k denote the ranks of the predicted scores and the subjective scores, respectively, and $\bar{\chi}$ and $\bar{\gamma}$ are the midranks of the respective data sets. This measure is used to quantify if changes (increase or decrease) in one variable is followed by changes (increase or decrease) in another variable, irrespective of the magnitude of the changes.

Finally, the outlier ratio (OR), r_0, is computed to measure prediction consistency [200]. A predicted MOS, MOS$_\Phi$, is defined as an outlier if

$$|\epsilon(k)| = |\mathsf{MOS}(k) - \mathsf{MOS}_\Phi(k)| > 2 \cdot \sigma_{\mathcal{M}_n}(k) \tag{38}$$

where $\sigma_{\mathcal{M}_n}$ denotes the standard error of the MOS as follows:

$$\sigma_{\mathcal{M}_n}(k) = \frac{\sigma_{\mathcal{M}}(k)}{\sqrt{N}} \tag{39}$$

and with $\sigma_{\mathcal{M}}$ being the standard deviation of the MOS over all N viewers. The OR then relates the number of outliers R_0 to the total number of metric values R in the set as

$$r_0 = \frac{R_0}{R}. \tag{40}$$

Here, we have $R = 60$ for the training set and $R = 20$ for the validation set.

3.6.4 Analysis of prediction function parameters

Exponential prediction functions have been derived for Δ_{NHIQM}, L_1-norm, and L_2-norm to map the metric values onto predicted MOS (see Section 3.5.6). As not all metrics behave the same, the exponential function may not be the best choice for every image quality metric. For this reason, we derived prediction functions for each of the other image quality metrics introduced in Section 3.6.1 taking into account the same constraints as for Δ_{NHIQM}; the goodness of fit, visual inspection of the fitting curve, and generalisation ability to unknown images. To evaluate the benefits of the perceptual relevance weight, we have further included

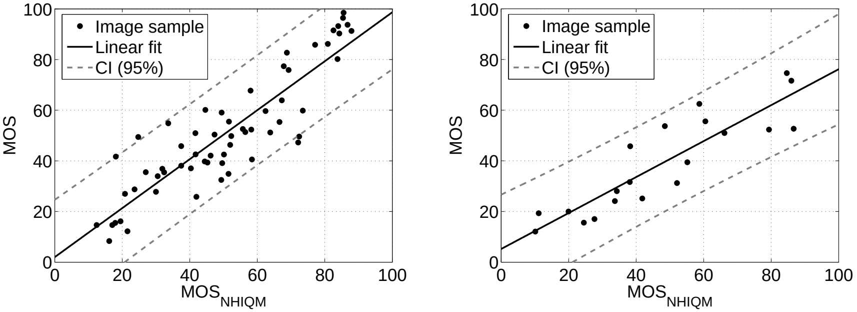
Table 12: Parameters of prediction functions for the image quality metrics (Φ in the prediction function denotes the respective metric).

Metric		Mapping		Parameters		
Type	Name	Type	Function	a	b	c
RR	Δ_{NHIQM}	Exponential	$a \cdot e^{b \cdot \Phi}$	88.79	-2.484	–
	Δ^{*}_{NHIQM}	Exponential	$a \cdot e^{b \cdot \Phi}$	85.66	-1.91	–
	L_1-norm	Exponential	$a \cdot e^{b \cdot \Phi}$	87.63	-1.839	–
	L_2-norm	Exponential	$a \cdot e^{b \cdot \Phi}$	90.21	-2.82	–
	RRIQA	Linear	$a \cdot \Phi + b$	-8.348	90.64	–
FR	SSIM	Power	$a \cdot \Phi^{b} + c$	50.96	11.7	40.76
	VIF	Exponential	$a \cdot e^{b \cdot \Phi}$	4.292	2.886	–
	VSNR	Linear	$a \cdot \Phi + b$	2.06	10.77	–
	PSNR	Logistic	$\frac{a}{1+e^{-b \cdot (\Phi - c)}}$	143	0.07	36.29

the Δ_{NHIQM} metric computed for all weights being equal to 1 ($w_i = 1$), which is denoted here as Δ^{*}_{NHIQM}.

Table 12 presents the mapping functions and its related parameters for all considered metrics. Apart from Δ_{NHIQM}, L_1-norm, and L_2-norm, the VIF metric was also found to be most suitably mapped with an exponential function. In the case of SSIM, the power function actually nicely mapped the metric onto predicted MOS without introducing any signs of overfitting on the training data. A simple linear fit was found to work best for both RRIQA and VSNR, indicating that these metrics already take into account the non-linear quality processing in the HVS. Finally, the logistic function was determined as the most suitable mapping function for PSNR.

The desired result from the mapping is a linear relationship between the MOS and the predicted MOS. As an example, Fig. 17 shows a scatter plot of the MOS versus MOS$_{NHIQM}$, the predicted MOS for Δ_{NHIQM}, for both the training and validation set. In addition, a linear function has been fitted to the data set and is presented along with the 95% CI. It can be seen that the fitting curves for both the training and validation set produce the desired linear relationship between predicted MOS and MOS.



(a) (b)

Figure 17: MOS versus predicted MOS, MOS$_{NHIQM}$, for: (a) training set and (b) validation set.

3.6.5 Quality prediction performance

The quality prediction performance of the considered quality metrics on both training and validation set is evaluated using the performance indicators introduced in Section 3.6.3. The performance indicators are presented in Table 13. Here, the prediction accuracy is quantified by the Pearson linear correlation coefficient, ρ_P, for both the actual metric values and also the predicted MOS. The generally higher Pearson correlation on the predicted MOS indicates the linearisation of the relationship between the objective and subjective quality scores. Furthermore, the prediction monotonicity is measured using the Spearman rank order correlation coefficient, ρ_S. As all mapping functions are strictly monotonic increasing or decreasing, the Spearman correlation is in fact the same for both the actual metric value and the predicted MOS. Thus, we present the Spearman correlation coefficient only for the predicted MOS. Finally, the RMSE and the OR, r_0, are presented for the predicted MOS.

The numerical results in Table 13 show strong quality prediction performance of Δ_{NHIQM}, L_1-norm, and L_2-norm in all indicators. In particular, the proposed metrics clearly outperform the other quality metrics in prediction accuracy and monotonicity. The prediction consistency is comparable to the better ones amongst the comparison metrics. This strong performance of Δ_{NHIQM}, L_1-norm, and L_2-norm is worth highlighting, as these metrics base the quality prediction on only a single or a few numerical values, as compared to the FR metrics, which use the entire reference image for quality assessment. Furthermore,

Table 13: Quality prediction performance of the image quality metrics and the corresponding predicted MOS.

	Metric		Predicted MOS							
Name	$\rho_{P,T}$	$\rho_{P,V}$	$\rho_{P,T}$	$\rho_{P,V}$	$\rho_{S,T}$	$\rho_{S,V}$	$RMSE_T$	$RMSE_V$	$r_{0,T}$	$r_{0,V}$
RR Metrics										
Δ_{NHIQM}	0.843	0.841	0.892	0.888	0.867	0.892	10.579	14.035	0.583	0.7
Δ^*_{NHIQM}	0.795	0.781	0.859	0.909	0.848	0.899	12.015	11.593	0.6	0.55
L_1-norm	0.833	0.842	0.873	0.897	0.854	0.901	11.406	12.895	0.65	0.7
L_2-norm	0.845	0.846	0.886	0.884	0.875	0.89	10.733	13.245	0.6	0.75
RRIQA	0.821	0.772	0.821	0.772	0.786	0.758	13.328	15.878	0.617	0.55
FR Metrics										
SSIM	0.582	0.434	0.697	0.628	0.558	0.347	16.733	18.498	0.7	0.9
VIF	0.713	0.727	0.789	0.788	0.813	0.729	14.486	13.892	0.817	0.7
VSNR	0.766	0.696	0.766	0.696	0.686	0.501	15.015	16.176	0.7	0.7
PSNR	0.742	0.712	0.751	0.705	0.638	0.615	15.413	14.835	0.633	0.65

the high quality prediction performance is given for both the training set and the validation set, indicating good generalisation ability of the exponential mapping function.

The higher Pearson linear correlation coefficients of Δ_{NHIQM} as compared to Δ^*_{NHIQM} on both the training and validation sets indicate the importance of the perceptual relevance weights with respect to the actual quality metric. After mapping to predicted MOS, these benefits are not as pronounced since the exponential prediction functions of these two metrics are fairly similar, as can be seen from the mapping parameters in Table 12.

Amongst the metrics used for comparison, RRIQA shows the best quality prediction performance. Thus, the four RR metrics in the test outperform the FR, which exemplifies that the presence of the reference image at the quality assessor is not necessarily a must to achieve good quality prediction performance. This, however, may come at the cost of computational complexity during RR acquisition, as can be observed from Table 11, where the FR metrics have lower complexity as compared to the RR metrics. In particular RRIQA exhibits a high computational complexity in comparison to all other metrics.

Within the set of FR metrics, VIF outperforms the other metrics in prediction accuracy and monotonicity, followed by VSNR and PSNR. In fact, PSNR shows the best prediction consistency performance amongst all FR metrics. It is interesting to note that the difference between the VIF and VSNR metrics to the PSNR metric is not as big as one might expect, even though PSNR is usually considered to perform poorly as a predictor of subjective quality. Furthermore, SSIM shows a somewhat lower quality prediction performance in all indicators in comparison to the other quality metrics.

The numerical results in Table 13 show that Δ_{NHIQM}, L_1-norm, and L_2-norm have very similar quality prediction performance in all four indicators. Considering the similarity of these performance indicators and the lower RR overhead of Δ_{NHIQM} compared to the L_p-norm, one may in fact consider Δ_{NHIQM} to provide the best overall solution for RR quality assessment in wireless imaging. However, the perceptual relevance weighted L_p-norm still offers the advantage of communicating information about particular feature differences and the related artifact severities.

Although the quality prediction performance of the proposed metrics is comparably higher to the other metrics, there may be still room for improvement. For this reason, different methodologies are discussed in Chapters 5, 6, 7, and 8, which were deployed with the aim to further enhance the quality prediction performance of the feature-based quality metrics.

4 Performance Assessment of a Deblocking Filter Using NHIQM

Before moving on to the various techniques deployed to improve the feature-based quality metrics, we conclude the first part of this thesis by discussing an application of NHIQM for deblocking filter design in video sequences. The deployment of NHIQM for deblocking filter design, as presented in this chapter, serves to illustrate the applicability of the proposed feature-based quality metrics in other contexts than image communication and to highlight the performance of NHIQM and in particular the feature metrics involved.

The most widely used video codecs, such as those based on the H.263, H.264/AVC, and MPEG-4 [213] format recommended by the ITU-T and ISO/IEC, provide efficient data compression as needed by mobile video applications. However, a common problem of these codecs is the introduction of blocking artifacts. These impairments to the original video sequence are due to quantisation of the DCT coefficients and motion estimation associated with block-based codecs. In H.264 this problem is reduced by an in-loop deblocking filter [214]. However, for other codecs like H.263, a post filter is needed to reduce blocking artifacts, such as the one recommended in H.263 App. III [215]. In this respect, image and video quality metrics play an important role to support a justified selection of suitable filters and the best filter settings.

Although NHIQM and the relevance weighted L_p-norm were designed with the aim to measure image quality degradations in a communication system, the use of these metrics is not necessarily restricted to this particular application. In this chapter, we thus use NHIQM to assess the performance of an adaptive deblocking filter [216], which was developed for H.263 coded video sequences as an alternative to the reference filter recommended in H.263 App. III [215]. The deblocking filter is designed to reduce blocking artifacts but at the same time faces the problem of introducing blur during the filtering process. Thus, the overall quality of the filtered video is a compromise between blocking and blur artifacts. Given the scope of distortions, NHIQM is considered to be a suitable metric to objectively assess the relative trade-off between the structural artifacts. In the given context of video sequences, NHIQM is used as a spatial quality assessor measuring structural distortions on a frame-by-frame basis.

In the following, we will briefly introduce the deblocking filter [216] and explain the test conditions considered here. This is followed by a visual inspection of the video sequences and by a detailed discussion of the quality prediction results obtained with NHIQM.

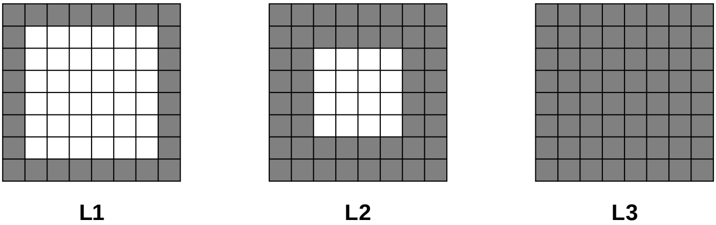


Figure 18: Filter levels with reference to 8×8 blocks of pixels.

4.1 Adaptive deblocking filter for H.263 encoded video

The adaptive deblocking filter that is examined in this chapter has been presented in [216]. This filter operates as a post-processing step after a hybrid differential pulse code modulation (DPCM) transform codec that uses 8×8 blocks for spatial decorrelation, as is given with H.263 and MPEG-4. The filter is adapted according to the level of compression, with higher compression resulting in increased filtering. The filter can be adjusted in two ways; the level of filtering and the filter strength. The level of filtering refers to the amount of pixels of each block that are being processed. Here, we apply filtering to the luminance data in three levels as follows:

L1: Only the first tier of border pixels in every 8×8 block is filtered.

L2: The first and second tier of pixels in every 8×8 block are filtered.

L3: The entire 8×8 pixels in every block are filtered.

For illustration the three levels of filtering are shown in Fig. 18.

The strength of the filter refers to how much low-pass filtering a certain compression level results in, where the compression level is related to the quantisation parameter (QP). The different filter strengths are achieved by adding an offset to the QP value at the input of the filter weight generator. The following four filter strengths S0-S3 are examined in this chapter:

S0: Deblocking filtering is not performed.

S1: Nominal strength is used.

S2: An offset of 4 is added.

S3: An offset of 6 is added.

4.2 Test conditions

For the performance assessment of the deblocking filter we selected three different video sequences, namely, 'Cart', 'Foreman', and 'Mobile', and compressed them using the H.263 codec. The video sequences comprised of 150 frames and were given in 176×144 pixels quarter common intermediate format (QCIF) at 15 frames per second (fps). The sequences have very different motion characteristics, with 'Cart' having high motion, 'Foreman' having medium motion, and 'Mobile' having low motion. Each sequence has been encoded at two different bit rates of 48 kb/s and 96 kb/s. Given the constant bit rates, one can expect a stronger presence of blocking artifacts in videos containing higher motion.

Each video sequence has been subjected to the different filtering conditions. In particular, the three filter levels (L1, L2, L3) have been deployed in all nine possible combinations with the three filter strengths (S1, S2, S3), in addition to the case where the deblocking filtering is not performed (S0). For all these sequences, the reference deblocking filter of H.263 App. III [215] has been switched off. For comparison, one additional sequence has been produced using the App. III deblocking filter instead of the adaptive filter. Thus, given the two bit rates, a total of 22 test sequences has been created for each content.

The quality of the video sequences has been measured by computing Δ_{NHIQM} (see (24)) on a frame-by-frame basis between the H.263 encoded sequences and their corresponding reference sequences. The final metric values were then mapped to predicted MOS, MOS$_{NHIQM}$, using the exponential prediction function derived in Section 3.5.6. In addition to the overall quality, each individual feature f_i has been recorded for more in depth analysis of the particular artifacts that occur in the video sequences. Both MOS$_{NHIQM}$ and the features f_i were then averaged over the total number of frames of each sequence to obtain overall quality and artifact measures for the sequences.

4.3 Visual quality of video frames

To visualise the effect of filter level on the ability of the adaptive filter to reduce blocking artifacts, Fig. 19 shows samples of the 72^{nd} frame of video sequence 'Cart' encoded at a bit rate of 48 kb/s ($QP=29$), filter strength S1, and different filter levels. The zoomed version is also given for each of the frame samples to facilitate a more detailed observation. It should be mentioned that the zoomed frames have been produced using the pixel replication technique which is a special case of nearest neighbour interpolation [217]. Distortions due to this zoom operation have not been observed for the considered frame samples. Furthermore,

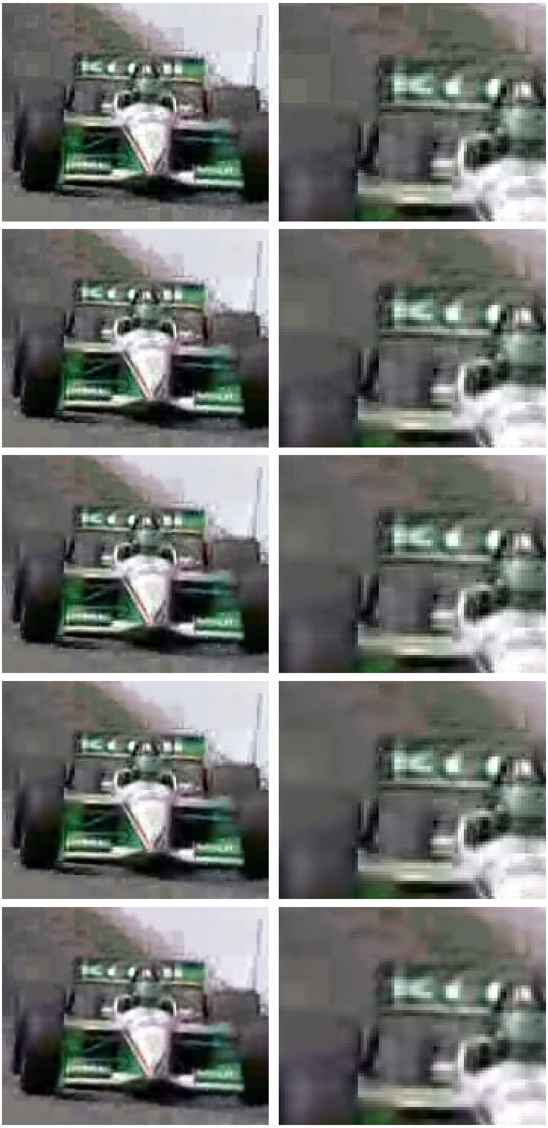


Figure 19: Frame samples of video sequence 'Cart' (left column) and their zoomed versions (right column) for bit rate $48\,\mathrm{kb/s}$, filter strength S1, and different filter levels. Top to bottom: No filtering, L1, L2, L3, H.263 App. III.

the video frame samples are used here as a means of visualisation, however, the filtering effects were in fact more pronounced when comparing the actual played back video sequences.

It can be seen in Fig. 19 that an increase in filter level from no filtering to level L1 gives the most perceptual improvement. This is especially apparent in the areas around the left back wheel of the cart. The blocking artifact can be clearly seen in the non-filtered sample in the top row of the figure. Implementing filter level L1, the samples in the second row are now much smoother, with the blocking largely reduced. An additional increase in the filter level does not appear to improve the perceptual quality for those samples. In fact, the filter level L3 increasingly starts to introduce blur to the video, as can be seen from the fourth row of the figure. Finally, it can be observed that with the H.263 App. III deblocking filter some degree of blocking still remains but overall the performance is comparable to the examined adaptive filter suggested in [216].

4.4 Structural feature metrics

From the visual inspection of the video frames one could observe that the adaptive deblocking filter indeed reduces the blocking artifacts but at the same time introduces blur to some degree at higher filtering levels. To quantify the degree to which these artifacts are present in the videos we recorded the difference of the five features that are part of NHIQM; blocking f_1, blur f_2, edge-based IA f_3, gradient-based IA f_4, block intensity shifts f_5 (see Section 3.1).

The feature differences computed on the sequences 'Cart', 'Foreman', and 'Mobile' are presented in Fig. 20, Fig. 21, and Fig. 22, respectively. It can be observed that the most significant reduction in the blocking feature f_1 is obtained when changing from no deblocking filtering being performed (S0) to nominal filter strength (S1). This is true for all three contents but applies particularly to the 'Cart' sequence, where blocking was most severe due to the high motion. Further increase in filter strength produces only minor blocking decrease. Analogously to the decrease in blocking, the blur feature metric f_2 increases most with the transition from no filtering (S0) to the nominal filter strength (S1). Higher filter strengths impact only little on the blur metric f_2.

As far as the filter levels are concerned, similar conclusions can be drawn to the ones from the visual inspection of the video frames (see Section 4.3). The blur feature metric f_2 predicts larger amounts of blur with increasing filter levels. Even though the blocking reduces for all three levels with an increase of filter strength, the absolute magnitude of blocking increases with higher filter levels from, for instance, a value below 0.3 in Fig. 20(a) to a value above 0.3 in Fig. 20(c). This

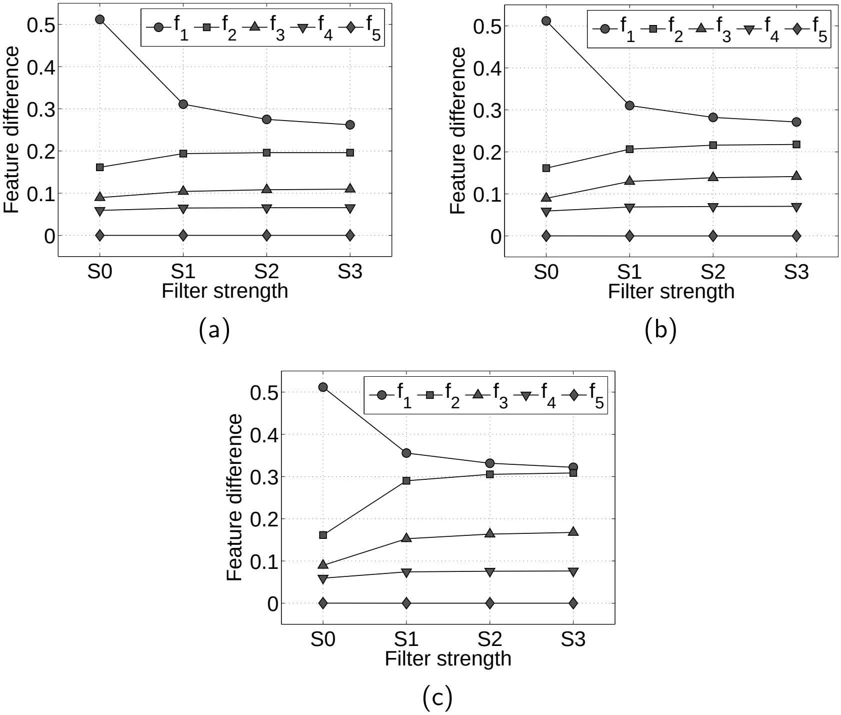
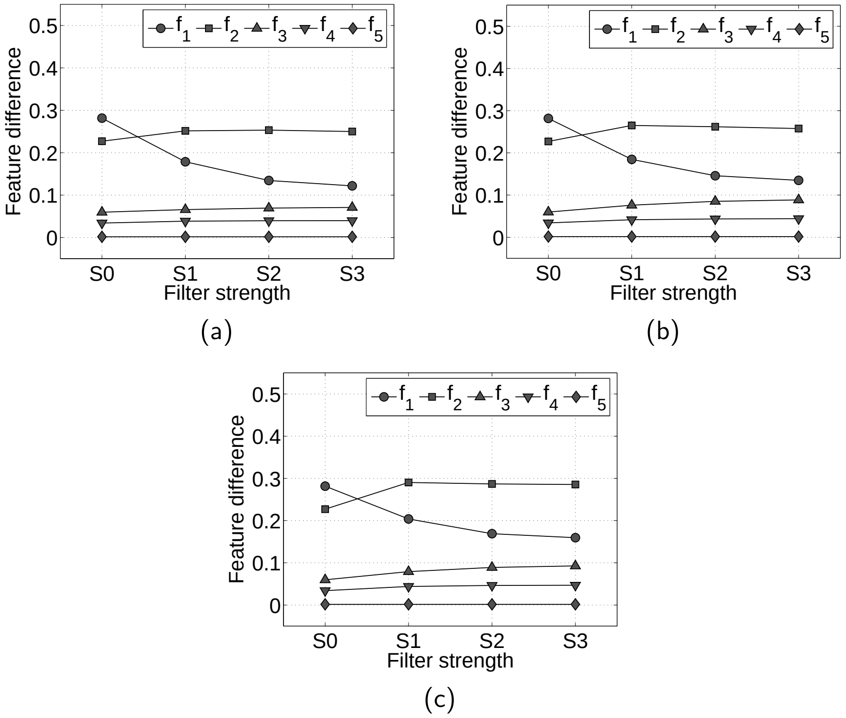


Figure 20: Extreme value normalised feature differences for the video sequence 'Cart' of bit rate 48 kb/s for the four different filter strengths S0, S1, S2, and S3 and for: (a) filter level L1, (b) filter level L2, and (c) filter level L3.

effect is due to the algorithm deployed here for the blocking artifact extraction [83] which not only accounts for blocking but also indirectly considers some degree of blur. As such, an increase of blur with higher filter levels may cause the feature metric f_1 to settle at higher values.

The three remaining features f_3, f_4, and f_5, are largely unaffected by the blocking and blur artifacts that are traded off by the different filter settings. This is particularly true for the block intensity shift feature f_5, which is not affected at all by both the filter level and the filter strength. This is a highly desirable result, as the block intensity shift measure should not be triggered by any of the



(a)

(b)

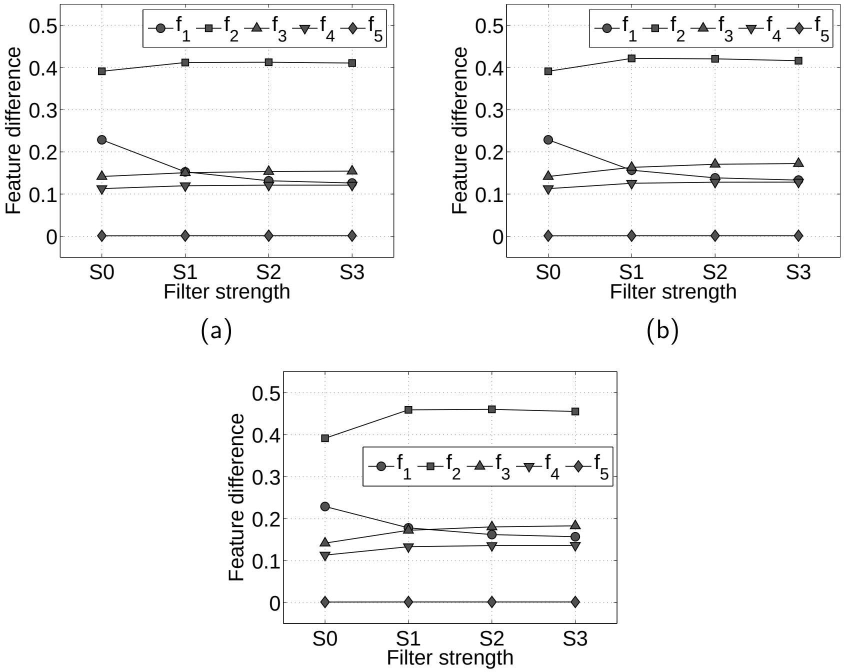
(c)

Figure 21: Extreme value normalised feature differences for the video sequence 'Foreman' of bit rate 48 kb/s for the four different filter strengths S0, S1, S2, and S3 and for: (a) filter level L1, (b) filter level L2, and (c) filter level L3.

artifacts that are present in the filtered video sequences. As such, the computation of feature metric f_5, and possibly f_3 and f_4, could be considered for exclusion from the metric computation in this particular application.

4.5 NHIQM-based quality prediction

The quality prediction results in terms of MOS$_{NHIQM}$ are presented in Fig. 23 for all three contents, all filter settings, and for the two bit rates. From this figure, it can be generally concluded that the video quality improves with the increase of



(a)

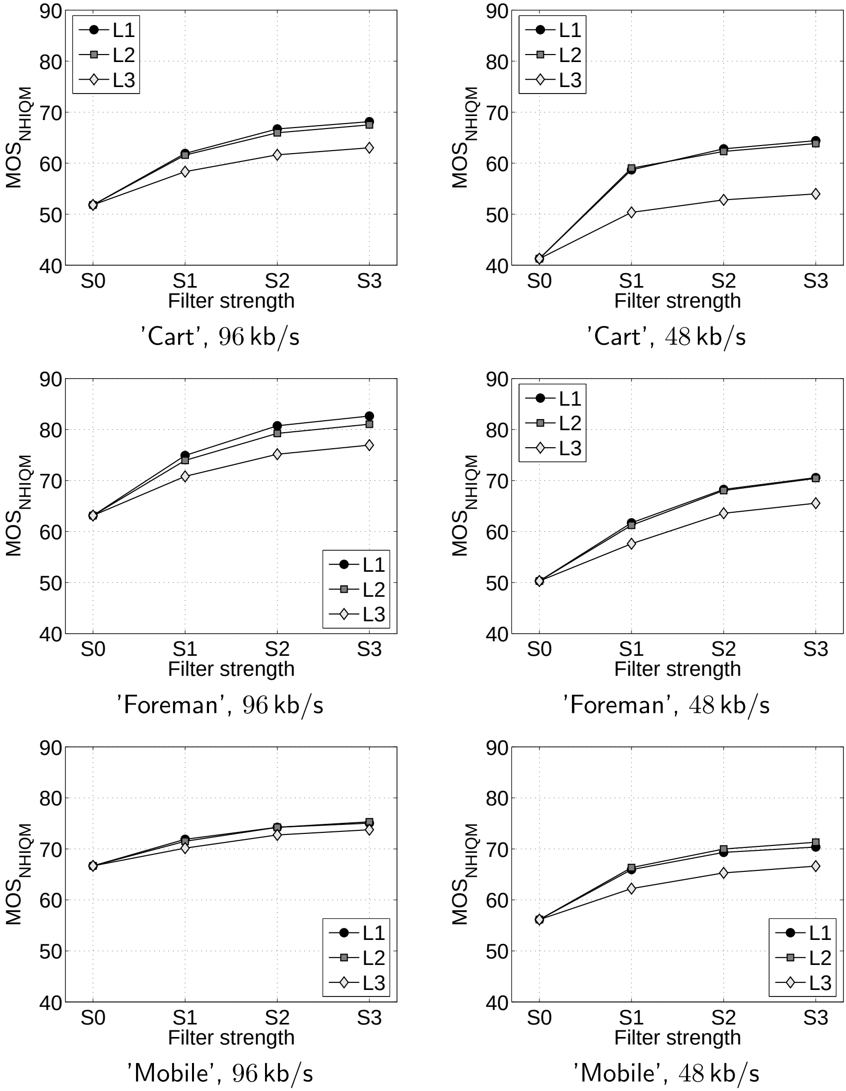
(b)

(c)

Figure 22: Extreme value normalised feature differences for the video sequence 'Mobile' of bit rate 48 kb/s for the four different filter strengths S0, S1, S2, and S3 and for: (a) filter level L1, (b) filter level L2, and (c) filter level L3.

filter strength from S0 to S3 and on the other hand, decreases with the increase of filter level from L1 to L3. The latter phenomenon can be related to the blur that is introduced with higher filter levels, which is particularly given for filter level L3. It can also be seen that the increase in quality between the different filter strengths is generally higher for the lower bit rate of 48 kb/s, which would be due to the stronger blocking artifacts at lower bit rates that provide the filter with a larger room for improvement.

For comparison, the MOS$_{NHIQM}$ for the H.263 App. III deblocking filter are provided in Table 14. It can be seen that the performance of the adaptive



'Cart', 96 kb/s

'Cart', 48 kb/s

'Foreman', 96 kb/s

'Foreman', 48 kb/s

'Mobile', 96 kb/s

'Mobile', 48 kb/s

Figure 23: Predicted MOS, MOS_{NHIQM}, for all filter levels (L1, L2, L3) and filter strengths (S0, S1, S2, S3).

Table 14: Predicted video quality for the H.263 App. III deblocking filter using MOS$_{NHIQM}$.

Bit	MOS$_{NHIQM}$		
rate	'Cart'	'Foreman'	'Mobile'
48 kb/s	62.02	70.96	88.78
96 kb/s	72.8	80.96	94.45

filter with filter level L1 and filter strength S3 (see Fig. 23) is similar to the H.263 App. III filter for the sequences 'Cart' and 'Foreman'. However, the H.263 App. III filter outperforms the adaptive filter on the 'Mobile' sequence. This particular sequence has a highly textured background which may cause the performance of the adaptive filter to drop.

It is worth highlighting that the quality prediction performed with NHIQM has been very consistent for the different video sequences, the filter parameters, and the bit rates. In particular, the improved quality with increasing filter strength and the reduced quality with increasing filter levels is consistently predicted for all three sequences and for both bit rates, as can be observed from Fig. 23. Furthermore, NHIQM consistently predicts the quality of the 96 kb/s sequences better than the corresponding 48 kb/s sequences. In this respect it is noted that the evaluation presented here was based solely on the predicted MOS, MOS$_{NHIQM}$, and no formal subjective tests have been conducted to support the validity of the results presented. However, when watching the video sequences, the four people that were involved in this particular project all agreed, that MOS$_{NHIQM}$ very well reflected the perceived quality differences due to the different filter settings. It was further agreed on that the different feature metrics very suitably captured the trade-off between the reduced blocking artifacts and the introduced blur artifacts.

5 Multiobjective Metric Optimisation

In Chapter 3, feature weights were derived for each of the feature extraction algorithms to account for the perceptual relevance of the associated artifacts that are measured. These weights were established in a very intuitive, but rather ad-hoc manner, by taking the correlation of the respective feature metric with subjectively perceived quality. As such, these weights inherently take into account that some artifacts are perceived as being more annoying than others. However, the interdependence between the different feature metrics and the related artifacts is not taken into account, as each of the weights is derived independently.

To take into account the inter-dependance between the features, we deploy in this chapter a multiobjective optimisation (MOO) approach [218] that facilitates finding the optimal weights of all features simultaneously. In this way, the mutual impact of the artifacts on perceptual image quality can be accounted for and deeper insights into the perceptual relevance of common artifacts observed in wireless imaging are gained. The ultimate goal being to improve the quality prediction performance of the feature-based image quality metrics. The optimisation of the weights on a particular set of images (the training images), however, strongly bears the risk of overfitting the quality metric. For this reason, we further aim on maintaining the metric's generalisation ability during the optimisation of the weights.

In the following sections, we first introduce the MOO framework that we propose for optimisation of feature-based image quality metrics, taking into account the design goals mentioned above. This framework is then deployed to determine the optimal weights for NHIQM, resulting in an improvement of quality prediction performance on the training set while maintaining generalisation ability to the validation set.

5.1 Multiobjective optimisation framework

Optimisation in general is concerned with minimisation of an objective, subject to a set of decision variables. However, the performance of a system cannot always be quantified by a single number. Therefore, MOO is concerned with the optimisation of multiple, often conflicting objectives [218]. Two objectives are said to be conflicting when a decrease in one objective leads to an increase in the other. A MOO problem can be transformed into a single-objective optimisation, for instance, by defining an objective as a weighted sum of multiple objectives. However, it is recommended to preserve the full dimension of the MOO and instead perform a two stage process [219]. In the first step, the design space is reduced to a set of optimal trade-offs between the objectives by determining the Pareto optimal (noninferior) solutions, which have the characteristic that one objective can only be optimised at the cost of another. However, although a Pareto optimal solution should always be a better compromise than the solutions it dominates, not all Pareto optimal solutions may be acceptable solutions. Therefore, in the second step the best trade-off solution is chosen from the set of Pareto optimal solutions under consideration of system design aspects [218].

5.1.1 Pareto optimal feature weights

Considering the above, we conduct a two step MOO, as it allows for taking into account our two conflicting objectives of optimising the quality prediction accuracy while maintaining the generalisation ability of the metric. The first step is to determine the Pareto optimal solutions and then chose the best compromise solution that best satisfies the constraints that we impose on the final metric.

We define a decision vector $\mathbf{w} = [w_1, \ldots, w_5] \in \mathbb{W} \subset \mathbb{R}^5$ containing the feature relevance weights w_i, $i \in \{1, 2, \ldots, 5\}$. The range of the weights in the decision space $\mathbb{W}$ is constrained to $w_i \in [0, 1]$. Given the above aims, we define two objectives as

- **Objective** O_A: maximise image quality prediction accuracy on a training set of images.
- **Objective** O_G: maximise generalisation performance to a validation set of images.

For this purpose, we utilise the two sets of images that we introduced earlier (see Sec. 3.5.1), the training set $\mathcal{I}_T$ containing 60 images and the validation set $\mathcal{I}_V$ containing 20 images. As with the derivation of the perceptual relevance weights in Section 3.5.3, the corresponding MOS sets, MOS$_T$ and MOS$_V$, again play a vital role in determining the optimal feature weights.

Objective O_A defines the metric's ability to predict MOS with minimal error and is measured as the Pearson linear correlation coefficient between a quality metric and the MOS as follows:

$$\rho_{P,T} = \frac{\sum_k (\Phi_T(k) - \overline{\Phi}_T)(\mathcal{M}_T(k) - \overline{\mathcal{M}}_T)}{\sqrt{\sum_k (\Phi_T(k) - \overline{\Phi}_T)^2}\sqrt{\sum_k (\mathcal{M}_T(k) - \overline{\mathcal{M}}_T)^2}} \tag{41}$$

where Φ_T and $\mathcal{M}_T$ are used here to denote the quality metric and MOS on the training set, respectively.

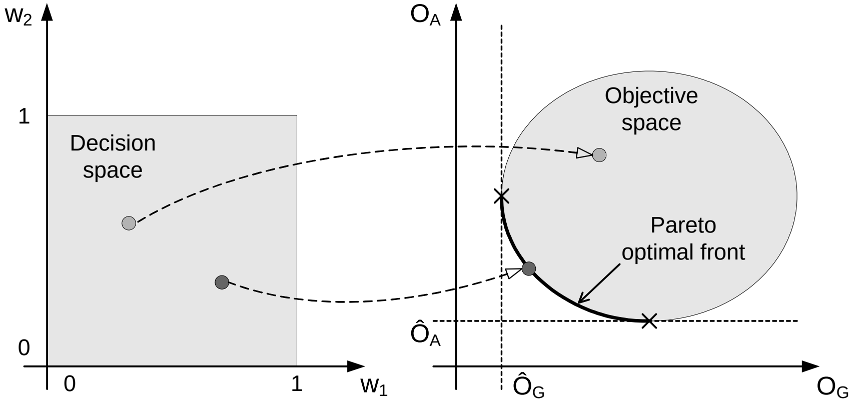


Figure 24: Multiobjective optimisation.

Optimising the weights based only on objective O_A would likely overtrain the metric, meaning, it would work very well on the training set but not on a set of unknown images. Therefore, the second objective O_G defines the metric's ability to perform quality prediction on a set of unknown images. We compute it as the absolute difference of the Pearson linear correlation coefficient on the training set, $\rho_{P,T}$, and the validation set, $\rho_{P,V}$, as follows:

$$\Delta\rho_P = |\rho_{P,T} - \rho_{P,V}|. \tag{42}$$

Thus, minimising $\Delta\rho_P$ assures that the prediction accuracy on the training and validation set are as close as possible and hence, the generalisation ability of the metric is maintained. Given the two objectives, we define the objective vector as

$$\mathbf{O}(\mathbf{w}) = \begin{pmatrix} O_A(\mathbf{w}) \\ O_G(\mathbf{w}) \end{pmatrix} = \begin{pmatrix} -|\rho_{P,T}| \\ \Delta\rho_P \end{pmatrix}. \tag{43}$$

Here, the minus sign before the objective O_A is used to adapt this objective to a minimisation problem as defined in the following section. Thus, by minimising $-|\rho_{P,T}|$ we inherently maximise the Pearson linear correlation coefficient.

The decision vector $\mathbf{w}$ is evaluated by assigning it an objective vector $\mathbf{O}$ in the objective space $\mathbb{O}$ as: $\mathbb{W} \rightarrow \mathbb{O} \subset \mathbb{R}^2$. This is illustrated in Fig. 24 for both a two-dimensional decision space and objective space (in the actual optimisation, the decision space has five dimensions, relating to the number of feature weights to be optimised). Here, a decision vector $\mathbf{w}$ is optimal in the Pareto sense, if it is assigned a noninferior solution on the Pareto optimal front. The Pareto optimal front is enclosed by $\hat{O}_A$ and $\hat{O}_G$, which are the independent optimal solutions for the respective objectives O_A and O_G. The other noninferior solutions are considered optimal trade-offs between the two objectives.

5.1.2 Goal attainment

We determine the noninferior solutions using the goal attainment method proposed in [220]. Here, goals $\mathbf{O}^* = (O_A^* \quad O_G^*)^T$ are specified, which can be interpreted as the desired level of the corresponding objectives. This requires sufficient intuitive understanding of the problem to know what values one would like to attain for each of the objectives, which we have from the earlier feature-based quality metric design (see Section 3). We then define the MOO problem as

$$P_G : \begin{cases} min & z \\ s.t. & \mathbf{O}(w) - \boldsymbol{\lambda} \cdot z \leq \mathbf{O}^* \end{cases} \tag{44}$$

where z is an unrestricted scalar variable which serves to minimise the two objectives O_A and O_G simultaneously using sequential quadratic programming [221]. The magnitude of $\boldsymbol{\lambda} = (\lambda_A \ \lambda_G)^T$ determines how close the objectives $(O_A(w) \ O_G(w))^T$ are to the goals $(O_A^* \ O_G^*)^T$. As is typically done [218], we set λ_A and λ_G to the absolute value of the goals

$$\lambda_A = |O_A^*| \quad \text{and} \quad \lambda_G = |O_G^*|. \tag{45}$$

The quantity $\boldsymbol{\lambda} \cdot z$ then corresponds to the degree of under- or overattainment of the goals $\mathbf{O}^*$.

5.2 Application to NHIQM

In this section, we apply the MOO framework to NHIQM to determine the Pareto optimal weights for the five features included in the metric. We first determine all Pareto optimal solutions within a feasible range of the two objectives and then discuss two optimal trade-off solutions.

5.2.1 Pareto optimal solutions

We use our knowledge from the previous metric design discussed in Chapter 3 to define the goals O_A^* and O_G^*. In particular, we set a fixed goal $O_A^* = -0.87$

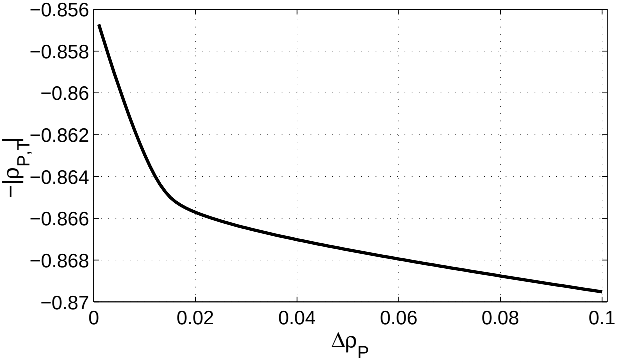


Figure 25: Pareto optimal front for NHIQM.

and define a range of goals $O_G^* \in [0.001, 0.1]$. Higher values than 0.1 were not considered for O_G since generalisation would be too weak. In fact, we found that the weights can be optimised to have a prediction accuracy on the training set that is well beyond 0.87, however, this is achieved at the cost of a very poor generalisation ability of the metric.

The noninferior solutions in terms of the Pareto optimal front are shown in Fig. 25 with the generalisation objective ($\Delta\rho_P$) given on the abscissa and the prediction accuracy objective ($-|\rho_{P,T}|$) given on the ordinate. One can see that prediction accuracy on the training set improves as the generalisation objective is relaxed. This is an expected result, as the optimisation process overtrains the metric on the training set when a lower prediction accuracy on the validation set is permitted. It should be noted though, that the loss in generalisation over the length of the abscissa is large ($|0.001 - 0.1| = 0.099$) compared to the gain in prediction accuracy over the corresponding interval on the ordinate ($|0.869 - 0.857| = 0.012$).

The Pareto optimal weights corresponding to the noninferior solutions are shown in Fig. 26. The weights of three features are clearly dominating, namely, feature f_1 (relating to blocking artifacts), feature f_3 (relating to ringing artifacts measured through edge-based IA), and feature f_5 (relating to block intensity shifts). The weights for feature f_4 (relating to ringing artifacts measured through gradient-based IA) are small over the whole range. The magnitudes of all these weights are fairly in line with the correlation weights obtained in Section 3.5.3. More unexpected, however, are the weights of the feature metric f_2 (relating

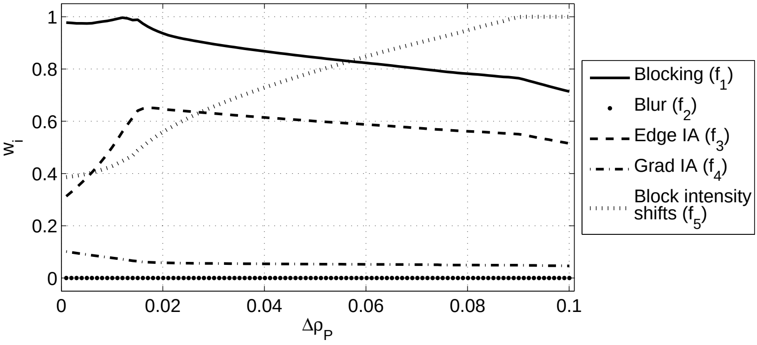


Figure 26: Pareto optimal weights for NHIQM.

to blur artifacts), which are zero or negligibly small over the whole range of noninferior solutions. This might be due to blocking being usually perceived as more annoying than blur but also due to the nature of the JPEG codec, that mainly produces blocking rather than blur artifacts. This result is particularly interesting since, due to the negligible size of the weights w_2, one may disregard the feature metric f_2 in the computation of the overall quality metric. Thus, given the complexity of the feature metric f_2 (see Section 3.3), approximately 47 % savings in computational cost can be achieved. In the case of deploying the L_p-norm, the overhead is further reduced, as the feature metric f_2 is not included in the RR information.

Given the noninferior solutions it is up to the system designer to make a choice as to which solution represents the most suitable trade-off under consideration of the system constraints. In the following, we discuss two representative solutions referred to as S1 and S2 resulting in the metrics $\Delta_{NHIQM}^{(S1)}$ and $\Delta_{NHIQM}^{(S2)}$, respectively. For both metrics, we further derived predicted MOS, $\text{MOS}_{NHIQM}^{(S1)}$ and $\text{MOS}_{NHIQM}^{(S2)}$, following the procedure as outlined in Section 3.5.6.

5.2.2 Solution S1: Optimal trade-off for Δ_{NHIQM}

Solution S1 was selected with respect to an optimal trade-off for Δ_{NHIQM}, disregarding the exponential mapping to MOS_{NHIQM}. When consulting the Pareto optimal front in Fig. 25, it can be seen that the gain in prediction accuracy $|\rho_{P,T}|$ is small in comparison to the loss in generalisation $\Delta\rho_P$, as one proceeds

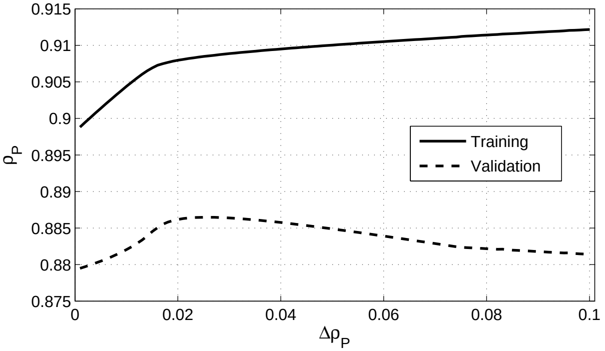


Figure 27: Pearson correlation ρ_P of MOS and MOS$_{NHIQM}$.

along the abscissa. For this reason, and also since we want to prevent the metric from overfitting on the training set, the optimal trade-off for Δ_{NHIQM} has been chosen at $\Delta\rho_P = 0.001$, representing the smallest value for $\Delta\rho_P$ and thus, the best generalisation ability of the metric $\Delta^{(S1)}_{NHIQM}$. The corresponding Pareto optimal weights $w_i^{(1)}$ are then obtained as

$$w_1^{(1)}=0.98, \quad w_2^{(1)}=0, \quad w_3^{(1)}=0.31, \quad w_4^{(1)}=0.1, \quad w_5^{(1)}=0.39. \qquad (46)$$

This set of weights provides direct insight into the perceptual relevance of wireless imaging artifacts, since we only considered the linear relationship between the weights, as given in Δ_{NHIQM}, rather than the non-linearity of MOS$_{NHIQM}$.

5.2.3 Solution S2: Optimal trade-off for MOS$_{NHIQM}$

Given the previous results, solution S2 was chosen with respect to an optimal trade-off for MOS$_{NHIQM}$. Here, for each Δ_{NHIQM} relating to the noninferior solutions in Fig. 25, a curve fitting has been conducted to derive a prediction function and map the metric to predicted MOS. The prediction accuracies, ρ_P, for all predicted MOS are presented in Fig. 27 for both the training and the validation set. One can see that ρ_P for the training set continuously increases with $\Delta\rho_P$. The validation set, however, has a maximum of ρ_P at $\Delta\rho_P = 0.026$ and has thus been chosen as the best trade-off for MOS$^{(S2)}_{NHIQM}$. The corresponding Pareto

Table 15: Quality prediction performance of the optimised metrics $\Delta_{NHIQM}^{(S1)}$ and $\Delta_{NHIQM}^{(S2)}$ in comparison to Δ_{NHIQM}.

Name	Metric		Predicted MOS							
	$\rho_{P,T}$	$\rho_{P,V}$	$\rho_{P,T}$	$\rho_{P,V}$	$\rho_{S,T}$	$\rho_{S,V}$	$RMSE_T$	$RMSE_V$	$r_{0,T}$	$r_{0,V}$
$\Delta_{NHIQM}^{(S1)}$	0.857	0.856	0.899	0.879	0.875	0.898	10.245	13.61	0.55	0.65
$\Delta_{NHIQM}^{(S2)}$	0.866	0.84	0.909	0.886	0.876	0.892	9.76	14.051	0.583	0.6
Δ_{NHIQM}	0.843	0.841	0.892	0.888	0.867	0.892	10.579	14.035	0.583	0.7

optimal weights $w_i^{(2)}$ are given by

$$w_1^{(2)}=0.91, \quad w_2^{(2)}=0, \quad w_3^{(2)}=0.64, \quad w_4^{(2)}=0.06, \quad w_5^{(2)}=0.62. \tag{47}$$

5.2.4 Evaluation of the optimal trade-off solutions

The quality prediction performance of the metrics $\Delta_{NHIQM}^{(S1)}$ and $\Delta_{NHIQM}^{(S2)}$ and their corresponding predicted MOS is shown in Table 15. The metrics are also compared to Δ_{NHIQM} to evaluate the gain of the optimal weights as compared to the correlation-based weights.

It can be seen from the table that both $\Delta_{NHIQM}^{(S1)}$ and $\Delta_{NHIQM}^{(S2)}$ are improved in prediction accuracy, as compared to Δ_{NHIQM}. Especially $\Delta_{NHIQM}^{(S1)}$ exhibits a nice generalisation performance, as the Pearson correlation in the training set and validation set are almost the same. The fairly small improvement indicates that the correlation weights incorporated in Δ_{NHIQM}, in fact, already performed well in representing the perceptual relevance of the corresponding features.

As for the other performance indicators, it can be observed that they are fairly comparable between the three considered metrics. However, it should be emphasised here that the comparable performance of $\Delta_{NHIQM}^{(S1)}$ and $\Delta_{NHIQM}^{(S2)}$ has been achieved without contribution of the blur metric. Hence, by disregarding the blur metric, valuable computational complexity can be saved, without sacrificing quality prediction performance.

The presented MOO framework has been shown to successfully improve the quality prediction performance of NHIQM while considerably reducing its computational complexity. However, the framework is considered to be generally applicable to feature-based image quality metrics. In fact, in Chapters 7 and 8 it will be shown that the MOO framework is also beneficial for the weight determination of a multiple-scale and a ROI-based image quality metric, respectively.

6 Artificial Neural Network Based Quality Metric

The RR quality metrics described in Chapter 3 deploy three steps to perform the transition from the structural feature metrics to the objectively predicted quality. Perceptual relevance weights are determined independently for each feature metric, in a pooling stage the weighted features are then condensed into a single quality metric, and finally the metric is mapped to predicted MOS using an exponential prediction function determined through regression analysis. To better account for the interdependence of the features and the related artifacts, the MOO framework, discussed in Chapter 5, was then deployed to simultaneously determine the feature relevance weights. In this chapter, we take this one step further by combining the simultaneous determination of the feature weights with the subsequent pooling and mapping stage. This is realised using an artificial neural network (ANN) to perform the quality prediction task [222, 223].

In analogy to the biological nervous system, an ANN consists of a number of neurons that communicate with each other through weighted connections. Each neuron in an ANN consists of an input (the equivalent to the dendrite in the nervous system) to receive information and an output (the equivalent to the axon in the nervous system) to forward information to other neurons. The level with which the data is 'fired' is determined by an activation function. Given this structure, ANN can be trained to find associations between an input signal and the corresponding desired response, by adjusting the weights between the network elements, the neurons. In this chapter, we consider a particular ANN, the feed-forward neural network (FFNN), also known as the multilayer perceptron [224]. Such an FFNN consists, in fact, of multiple layers of logistic regression models that are successively interconnected with each other. Thus, the mapping to predicted MOS, independently conducted in our earlier metrics, is an integral part of the ANN-based metric.

In the context of quality assessment, ANN have previously been used to map a set of objective features to predicted MOS. Mohamed and Rubino [100] have deployed random neural networks for video quality prediction based on codec and channel specific input parameters, such as bit rate, frame rate, and packet loss. Gastaldo et al. [223] designed a neural network based on simple input features such as signal energy, covariance, and entropy. Both works concluded that the deployed ANN performed well in the quality prediction task. In our work, we use the structural features (see Section 3.1) extracted from the image content as network input to an FFNN to predict the MOS from the subjective experiments E1 and E2.

An overview of the RR visual quality assessment framework (see Fig. 4)

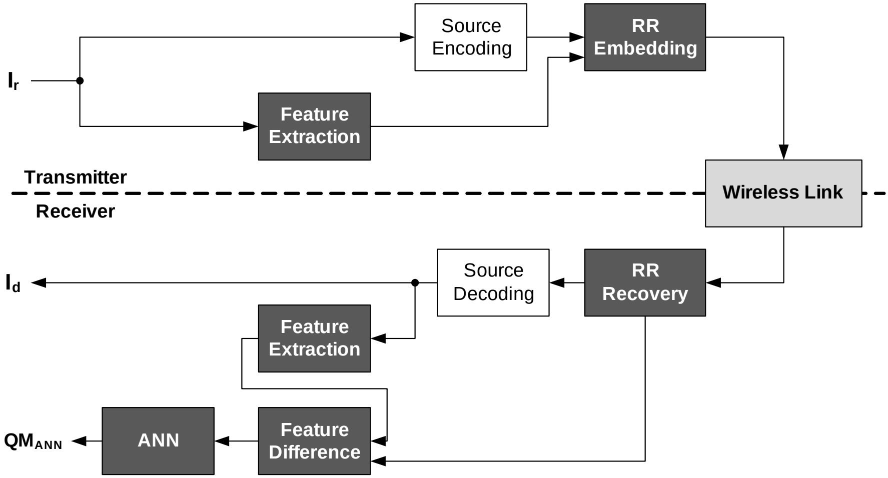


Figure 28: Reduced-reference quality assessment using an artificial neural network.

adapted to the deployment of the ANN-based metric is shown in Fig. 28. As with the perceptual relevance weighted L_p-norm, features are extracted both at the transmitter and at the receiver, disregarding a subsequent pooling step. The differences between the transmitted and received features are then fed into the ANN to perform the quality prediction task. Alternatively, the distorted features can be used as network input, thus, omitting the feature extraction at the transmitter. It is shown that this NR design of the ANN results in similar performance to the RR approach.

In the following sections, we discuss the FFNN architecture, the network training, and the evaluation of the prediction performance results. Since ANN are a very involved topic, covering a wide range of different network topologies, learning paradigms, and application scenarios, we focus here on the crucial facts needed to understand the particular feature-based ANN quality metric design deployed in this thesis.

6.1 Feed-forward neural network architecture

In general, an FFNN consists of multiple layers, in particular, an input layer, an output layer, and one or several hidden layers. Each of the layers contains various amounts of hidden units, the neurons. These are processing units composed of

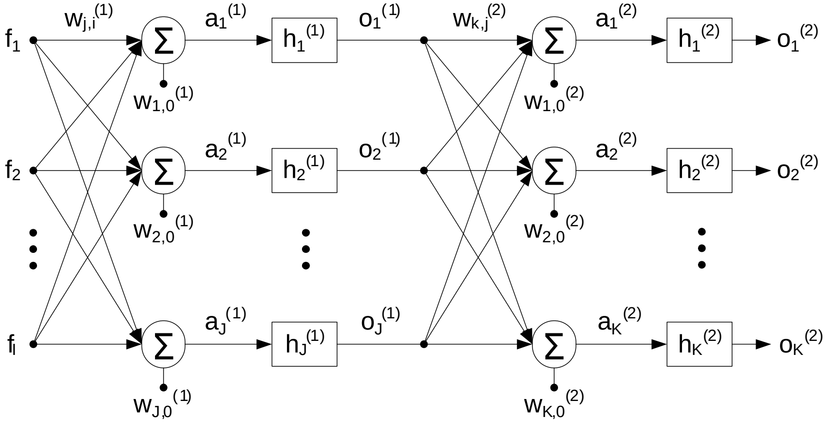


Figure 29: Fully-connected two-layer neural network structure.

a summation part and an activation function. In a fully-connected network, all neurons in a hidden layer have a weighted interconnection to the neurons in the previous and successive layer. The outputs of the $(n-1)$-th layer, $o_l^{(n-1)}$, then serve as input to the n-th layer, where they are combined as

$$a_l^{(n)} = w_{l,0}^{(n)} + \sum_{i=1}^{I} w_{l,i}^{(n)} \cdot o_l^{(n-1)} \tag{48}$$

where $w_{l,i}^{(n)}$ and $w_{l,0}^{(n)}$ denote the weights and the biases, respectively, of the l-th neuron in the n-th layer, and $a_l^{(n)}$ are referred to as the corresponding activations. Each of the activations is then transformed using a differentiable activation function $h_l^{(n)}(\cdot)$ to yield

$$o_l^{(n)} = h_l^{(n)}(a_l^{(n)}). \tag{49}$$

A fully-connected two-layer network architecture with J neurons in the first layer and K neurons in the second layer is illustrated in Fig. 29. Here, f_i denotes the i^{th} out of I features at the network input.

The choice of the right number of layers in an ANN and the number of neurons within each layer is crucial with respect to the network's prediction performance. Networks of too high complexity tend to overfit the data to the training set and thus, exhibit weak generalisation performance. On the other hand, networks of too low complexity might result in large errors for both training and generalisation. It is, however, known that any continuous function can be approximated sufficiently well by a two-layer network architecture, given a non-linear, differentiable transfer function and sufficient neurons in the first layer, and a linear transfer function in the second layer [225]. In view of this finding, we designed a fully-connected two-layer FFNN containing one hidden layer with multiple neurons and one output layer with a single neuron. The differentiable bipolar sigmoid function was chosen as the activation function for all neurons in the hidden layer

$$h^{(1)}(a_j^{(1)}) = \frac{e^{a_j^{(1)}} - e^{-a_j^{(1)}}}{e^{a_j^{(1)}} + e^{-a_j^{(1)}}}. \tag{50}$$

A linear activation function was used for the single output neuron

$$h^{(2)}(a_1^{(2)}) = a_1^{(2)}. \tag{51}$$

Given the above network topology, the predicted MOS, MOS$_\Phi$, is computed by the FFNN as

$$\mathsf{MOS}_\Phi \triangleq o_1^{[2]} = h^{(2)}\Big(\mathbf{w}^{(2)} \cdot h^{(1)}\big(\,\mathbf{W}^{(1)}\mathbf{f} + \mathbf{w}_0^{(1)}\,\big) + w_{1,0}^{(2)}\Big) \tag{52}$$

with $\mathbf{f} = [f_i]_{5\times1}$ being the feature input vector, $\mathbf{W}^{(1)} = [w_{j,i}^{(1)}]_{J\times5}$ being the matrix of weights in the first layer, $\mathbf{w}^{(2)} = [w_{1,j}^{(2)}]_{1\times J}$ being the vector of weights in the second layer, and $\mathbf{w}_0^{(1)} = [w_{j,0}^{(1)}]_{J\times1}$ and $w_{1,0}^{(2)}$ representing the biases of the respective layers.

There is no strict design rule regarding the number J of neurons in the hidden layer and thus, we considered different numbers of neurons to find the best prediction performance of the network with respect to both prediction accuracy and generalisation ability.

6.2 Network training

Unlike for optimisation, where defined mathematical expressions are used to solve a given problem, iterative numerical procedures are used in the case of ANN to train the network's performance of associating a given input to a desired output. In unsupervised learning, the weights and biases are modified in response to the network inputs only, whereas in supervised learning, the network outputs o are compared to targets t. The procedure for supervised learning is illustrated in

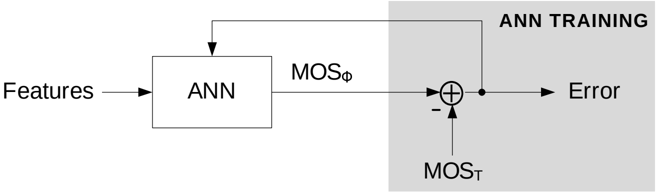


Figure 30: Supervised training of the artificial neural network.

Fig. 30. Here, the error between the output and the target is fed back into the network and is used in an iterative procedure to adjust the weights and biases of the network to decrease the errors. We use the MOS from the training set of the subjective experiments E1 and E2, MOS$_T$, as targets for comparison with the predicted MOS, MOS$_\Phi$, obtained as network output. As such, the network is trained using dedicated input-target pairs $(\mathbf{f}_T(m), \text{MOS}_T(m))$, with $\mathbf{f}_T(m)$ being a vector of all 5 features corresponding to training image m.

6.2.1 Gradient descent optimisation using error backpropagation

Error backpropagation (EBP) was created by generalising the Widrow-Hoff learning rule [226] to multiple-layer networks and non-linear differentiable activation functions. It utilises efficiently the gradient descent algorithm to minimise the error between the k-th network output $o(m,k)$ and the target $t(m,k)$ as

$$e(m,k) = o(m,k) - t(m,k). \tag{53}$$

Similar to regression, the sum-of-squares error function for a particular training sample m and for all K network outputs

$$E_m = \frac{1}{2}\sum_{k=1}^{K} e^2(m,k) \tag{54}$$

is used as a performance measure of the error. When EBP is deployed in batch mode, the error functions E_m are summed up over all M training samples before further processing as follows:

$$E(\mathbf{w}) = \sum_{m=1}^{M} E_m(\mathbf{w}). \tag{55}$$

In each iteration, one can then distinguish between two major stages that are performed in the EBP algorithm. In a first stage, the gradient of the sum-of-squares error function is evaluated with respect to the network weights **w** as

$$\nabla E(\mathbf{w}) = \frac{\delta E}{\delta \mathbf{w}}. \tag{56}$$

For this purpose, the errors are step-by-step propagated back through the network and the derivatives of the errors are computed at each step. In the second stage, the derivatives are then used to adjust the network weights to move a small step into the steepest direction of the negative gradient

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} - \gamma \nabla E(\mathbf{w}^{(\tau)}) \tag{57}$$

with γ being the learning rate that regulates the step length of each iteration in the direction of the negative gradient.

Although the gradient descent based weight updates intuitively seem like a reasonable methodology, there are in fact other algorithms using Newton-based methods that are more robust and faster than the gradient descent based methods. The Newton-based methods also have the advantage that the error function decreases in each iteration, unless the weight vector arrives at a local or global minimum. A Newton-based algorithm that we used for network training is discussed in the following section.

6.2.2 Levenberg-Marquardt algorithm

The Levenberg-Marquardt algorithm (LMA) [227, 228] interpolates between the Gauss–Newton algorithm (GNA) and the gradient descent algorithm (GDA). Newton-based methods in general use the Hessian matrix for the weights update, which contain computationally challenging second order derivatives. The GNA is an adapted Newton method that avoids the computation of second order derivatives by approximation of the Hessian matrix **H** through Jacobian matrices **J** as

$$\mathbf{H} = \mathbf{J}^T \mathbf{J}. \tag{58}$$

The LMA further has the advantage that it is more robust than the GNA, which means that it usually finds a solution, even if it starts far off the final minimum. For these reasons we deployed the LMA to perform the minimisation of the error function.

The difference of the LMA to the GDA explained earlier mainly lies in the weights update, which is given for the LMA as

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} - [\mathbf{J}^T \mathbf{J} + \mu \mathbf{I}]^{-1} \mathbf{J}^T \mathbf{e} \tag{59}$$

with **J** being the Jacobian matrix, **I** being the identity matrix, **e** being a vector of network errors, and μ being a damping factor. In analogy to (57), the learning rate of the LMA is given as

$$\gamma = [\mathbf{J}^T \mathbf{J} + \mu \mathbf{I}]^{-1} \tag{60}$$

and the gradient is computed as

$$\nabla E = \mathbf{J}^T \mathbf{e}. \tag{61}$$

The damping factor μ given in (59) regulates the trade-off between the Gauss-Newton like behaviour and the gradient descent like behaviour of the LMA. If μ is zero, the method boils down to the GNA. On the other hand, if μ is large, the LMA converts to a GDA with a small step size. Generally, the GNA is faster and more accurate near an error minimum, so the aim is to shift towards Newton's method as quickly as possible. For this purpose, μ is decreased in each iteration when the error function is decreased and μ is only increased when a tentative step would result in an increase of the error function. This way, the performance function will always be reduced at each iteration of the algorithm.

6.2.3 Bayesian regularisation

Due to the relatively low number of images and related MOS that we have available in our training set, the networks capability to generalise to unknown data is restricted. Therefore, special methods have to be used to improve the generalisation of the network. The most widely used techniques are early stopping and Bayesian regularisation. The former method is typically recommended for larger data sets, as the data needs to be divided into three subsets, a training, validation, and test set. On the other hand, Bayesian regularisation only needs a training and a test set and is therefore preferably used on smaller data sets. For this reason, we used Bayesian regularisation in conjunction with the LMA to train our network.

The aim of deploying regularisation is to create smoother network outputs that are less prone to overfitting on the training data. This is encouraged by adding a penalty term Ψ to the error function to yield the regularised error function

$$E_{reg} = E + \nu \Psi \tag{62}$$

where the parameter ν controls the influence of the penalty term Ψ in relation to the error function E. Networks that provide a good fit on the training data will result in a small value for E whereas networks that produce smooth outcomes will exhibit a small value of Ψ. To trade-off these conflicting characteristics, the network training is performed using the regularised error function E_{reg}.

A simple and widely adopted penalty term is referred to as weighted decay [229]. It consists of the sum of squares of all Q weights w_q (including biases) in the network and is given as follows:

$$\Psi = \frac{1}{2} \sum_{q=1}^{Q} w_q^2. \tag{63}$$

To get the best performance with Bayesian regularisation during training, we scaled both network inputs and targets to fall in the range $[-1, 1]$. In a post-processing step the MOS have been reverted to fall into their original interval $[0, 100]$.

6.3 Network performance evaluation

In this section, we first evaluate the performance of the designed ANN with respect to a varying number of neurons in the hidden layer. We then present the trained network weights and evaluate the quality prediction performance of the network using the prediction performance indicators introduced in Section 3.6.3. The evaluation is discussed with respect to two different network inputs that we consider here; the feature differences Δf between the reference and distorted features (RR deployment of the ANN) and the distorted features f_d (NR deployment of the ANN). The resulting networks are in the following referred to as ANN$_{\Delta f}$ and ANN$_{f_d}$, respectively.

6.3.1 Impact of the number of hidden neurons

We varied the number of neurons in the hidden layer between 1 and 8 and trained the network for each case using the methodology outlined in Section 6.2. The features and MOS corresponding to the training set $\mathcal{I}_T$ (see Section 3.5.1) were used here for the network training and the related quantities from the validation set $\mathcal{I}_V$ were used to validate the network. The quality prediction performance of the resulting ANN was then analysed by computing the Pearson linear correlation coefficient ρ_P between the network output, the predicted MOS, and the corresponding MOS. The results are shown in Fig. 31(a) and Fig. 31(b) for ANN$_{\Delta f}$ and ANN$_{f_d}$, respectively.

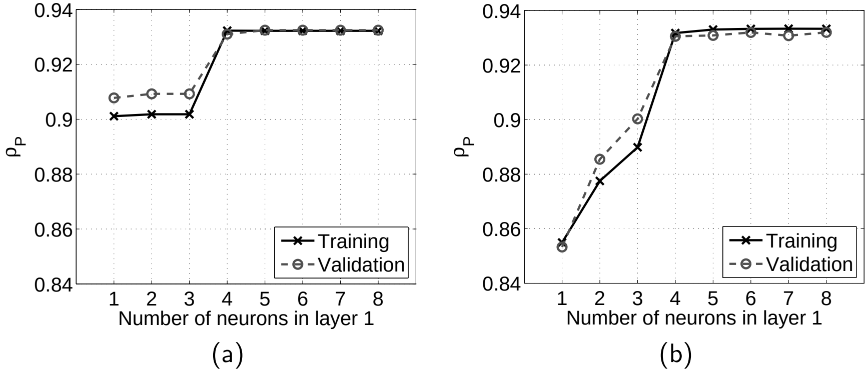


Figure 31: Pearson linear correlation coefficient ρ_P between predicted MOS and MOS for varying numbers of neurons in the first layer and two different network inputs: (a) difference features Δf and (b) distorted features f_d.

It can be seen that for both cases, the prediction accuracy raises by increasing the number of neurons from 1 to 4 and levels out for a number of neurons larger than 4. Furthermore, the performance on the training set and the validation set are very close, which is true for both ANN$_{\Delta f}$ and ANN$_{f_d}$ and also for all considered numbers of neurons. This is a strong indication that the Bayesian regularisation performed well in preventing the network from overfitting.

6.3.2 Trained network weights

Given the very similar prediction performance at a lower network complexity, we consider the networks ANN$_{\Delta f}$ and ANN$_{f_d}$ with $J = 4$ neurons in the hidden layer as the preferred choice over the other networks with $J > 4$ neurons in the hidden layer. Thus, only the matrices and vectors containing the final weights and biases of the trained networks with $J = 4$ neurons are presented in the following. The weights and biases for ANN$_{\Delta f}$ are given in (64)-(66) as follows:

- Weights of the first layer for ANN$_{\Delta f}$:

$$\mathbf{W}^{(1)}_{\Delta f} = \begin{pmatrix} -0.934 & -0.861 & -0.453 & 0.035 & 0.742 \\ 0.323 & -1.43 & 0.373 & -0.855 & -0.009 \\ 0.435 & -0.888 & 0.862 & -0.098 & 0.416 \\ 1.496 & 0.274 & -0.565 & 0.315 & -0.44 \end{pmatrix} \tag{64}$$

- Weights of the second layer for ANN$_{\Delta f}$:

$$\mathbf{w}_{\Delta f}^{(2)} = \begin{pmatrix} -0.768 & 1.031 & -1.415 & -1.5 \end{pmatrix} \tag{65}$$

- Biases of the first and second layer for ANN$_{\Delta f}$:

$$\mathbf{w}_{0,\Delta f}^{(1)} = \begin{pmatrix} -0.459 \\ -0.364 \\ 0.399 \\ 1.126 \end{pmatrix}, \qquad w_{1,0,\Delta f}^{(2)} = 0.133 \tag{66}$$

The weights and biases for ANN$_{f_d}$ are given in (67)-(69) as follows:

- Weights of the first layer for ANN$_{f_d}$:

$$\mathbf{W}_{f_d}^{(1)} = \begin{pmatrix} 0.24 & 0.627 & 0.932 & 0.253 & -0.07 \\ -1.79 & -0.333 & -0.178 & 2.231 & -0.568 \\ 0.049 & 1.087 & 0.174 & -0.205 & -0.253 \\ 0.74 & 0.444 & 1.138 & -2.443 & 0.492 \end{pmatrix} \tag{67}$$

- Weights of the second layer for ANN$_{f_d}$:

$$\mathbf{w}_{f_d}^{(2)} = \begin{pmatrix} -0.752 & 1.837 & 0.608 & 1.577 \end{pmatrix} \tag{68}$$

- Biases of the first and second layer for ANN$_{f_d}$:

$$\mathbf{w}_{0,f_d}^{(1)} = \begin{pmatrix} -0.453 \\ 0.242 \\ -0.357 \\ 0.474 \end{pmatrix}, \qquad w_{1,0,f_d}^{(2)} = -0.863 \tag{69}$$

The presented weights are not as easily interpretable as, for instance, the weights obtained through optimisation (see Section 5.2). However, they are given here for completeness and also to show that in fact for both ANN$_{\Delta f}$ and ANN$_{f_d}$, all network weights and biases are contributing to the overall computation of the predicted MOS. This does not necessarily have to be the case, as we found that for $J > 4$ neurons in the hidden layer, the number of redundant weights increased. This was apparent in the weights either being trained to be equal to zero or in multiple rows of weights having exactly the same values. This is another indicator for the good performance of the regularisation method and can be related to the constant prediction accuracy for $J \geq 4$.

Table 16: Quality prediction performance of the artificial neural network based metric using as input either the difference features Δf (reduced-reference) or the distorted features f_d (no-reference).

Network	$\rho_{P,T}$	$\rho_{P,V}$	$\rho_{S,T}$	$\rho_{S,V}$	$RMSE_T$	$RMSE_V$	$r_{0,T}$	$r_{0,V}$
ANN$_{\Delta f}$	0.932	0.931	0.919	0.936	8.087	9.78	0.417	0.55
ANN$_{f_d}$	0.932	0.931	0.926	0.937	8.543	8.812	0.45	0.55

6.3.3 Quality prediction performance

The prediction performance indicators computed between the predicted MOS and the MOS are given for both networks ANN$_{\Delta f}$ and ANN$_{f_d}$ in Table 16. It can be seen that both networks perform in all performance indicators comparably better to the previously discussed metrics, Δ_{NHIQM}, L_p-norm, $\Delta_{NHIQM}^{(S1)}$, and $\Delta_{NHIQM}^{(S2)}$. This may be attributed to the more complex relationship between the features that is accounted for by the larger number of weights. It may further be a result of the combined feature weighting, feature pooling, and metric mapping, which has been performed in independent steps in the previous metrics.

It should also be highlighted, that ANN$_{\Delta f}$ and ANN$_{f_d}$ perform similarly well in all indicators. This is particularly interesting, since ANN$_{f_d}$ found this association between the input features and the subjectively perceived quality solely on the distorted features, not being given any reference information regarding the original image content. This suggests that information about the changes in the structural information is not necessarily needed to enhance prediction performance of the ANN. Thus, one may deploy the ANN as an NR image quality predictor to save the feature extraction on the reference image, as well as transmission overhead in terms of the reference feature values. However, this comes at the cost of not being able to predict the quality loss that occurs during transmission as no information about the original image quality is given.

7 Multiple-Scale Quality Assessment

The feature-based quality metrics presented thus far incorporate multiple high level assumptions of the HVS, such as, the extraction of structural information, the perceptually varying significance of different distortion types, and the non-linear quality processing. In this chapter, we extend these models with an integral characteristic of the HVS, namely, multiple-scale processing.

The integration of multiple-scale processing into quality assessment is motivated by the well known fact that the HVS is adapted to visual information processing at different scales [1]. This evolutionary adaption arises from the nature of the environment around us, which contains objects of all different sizes. Objects may also be located at varying distances, changing the size of their appearance. The multiple-scale processing in the HVS suggests also an objective multiple resolution analysis of images, to not miss information in the image due to single resolution analysis. In relation to image quality, one may suspect that distortions are also perceived differently at given scales. The Gaussian pyramid is a convenient and computationally efficient multi-resolution image representation that mirrors the multiple scales of processing in the HVS well [230]. It has therefore been widely used before to perform image analysis at multiple resolutions.

In [183], we proposed a multiple-scale extension of NHIQM by computing Δ_{NHIQM} in each level of the Gaussian pyramid decomposition, with a subsequent pooling across pyramid levels. Within the level pooling, weights were deployed to account for the perceptual relevance of each pyramid level. This approach resulted in a considerable improvement in quality prediction performance for NHIQM and was found to also improve the prediction performance of other quality metrics, such as SSIM and PSNR. However, the performance gain of NHIQM was achieved at the cost of having to compute an elaborate amount of feature metrics, in fact, 5 features for each level involved. Furthermore, this approach did not take into account that not every feature computation may be necessary within each pyramid level. Finally, the interdependence between the perceptual relevance of the features and the pyramid levels was disregarded as the feature weights and level weights were obtained independently from each other.

In contrast to this work, we utilise in this chapter the MOO framework outlined in Chapter 5.1 to achieve three goals: 1) create a multiple-scale feature-based quality metric with high quality prediction performance and good generalisation ability, 2) lower the metric's complexity as compared to the metric proposed in [183] by reducing the number of features that need to be computed in each Gaussian pyramid level, and 3) account for the interdependence between feature and level relevance weights.

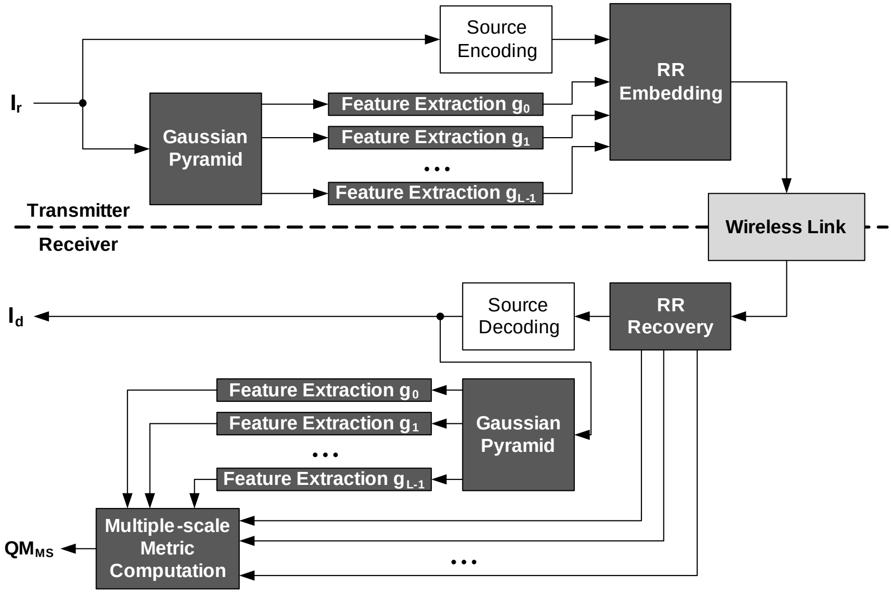


Figure 32: Reduced-reference quality assessment using the multiple-scale feature-based quality metric.

An overview of the RR visual quality assessment framework (see Fig. 4), adapted to the multiple-scale feature-based quality metric (MSFQM), is presented in Fig. 32. In analogy to the perceptual relevance weighted L_p-norm the features need to be transmitted over the channel. On the contrary to the L_p-norm, however, a multi-resolution decomposition is deployed in terms of the Gaussian pyramid before the actual feature extraction. The pooling stage at the transmitter is omitted, as we found that it is crucial to keep information regarding the structural degradations for each pyramid level [183]. The respective Gaussian pyramid decomposition and feature extraction is conducted at the receiver and the reference and distorted multiple-scale features are then pooled, yielding the MSFQM.

In the following, we summarise the Gaussian pyramid decomposition and discuss the level-based feature extraction, the pooling, and the optimisation of the perceptual relevance weights. A thresholding is introduced that was deployed to further reduce the computational complexity of the metric at minimal impact on

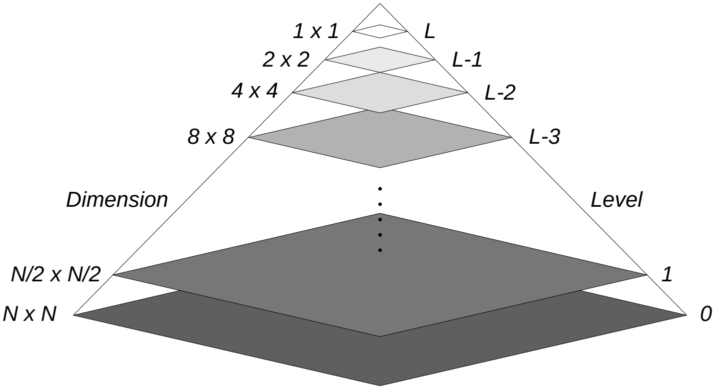


Figure 33: Full Gaussian pyramid decomposition.

the quality prediction performance.

7.1 Gaussian pyramid creation

The Gaussian pyramid is a convenient multi-resolution image representation that mirrors the multiple scales of processing in the HVS [230]. A full Gaussian pyramid decomposition is illustrated in Fig. 33 along with the level numbering and image dimensions for each level. In the following, an efficient iterative algorithm for the pyramid generation is summarised from [231].

The pyramid consists of $L+1$ levels with the image g_0 in the bottom being the original image in full resolution $N \times N$. The higher level images g_l, $l = 1, 2, \ldots, L$, are low-pass filtered and sub-sampled versions of the underlying images. The low-pass filtering is performed using a generating kernel $\sigma(m,n)$ of size 5×5 pixels. The size has been chosen with respect to filtering performance and low computational cost. Sub-sampling is done by a factor of two. Therewith, each image g_l is obtained from its predecessor g_{l-1} as

$$g_l(u,v) = \sum_{m=-2}^{2} \sum_{n=-2}^{2} \sigma(m,n) \cdot g_{l-1}(2u+m, 2v+n). \tag{70}$$

For simplicity, the generating kernel is made separable

$$\sigma(m,n) = \sigma(m) \cdot \sigma(n). \tag{71}$$

Furthermore, the one-dimensional patterns $\sigma(m)$ and $\sigma(n)$ are constrained to be normalised

$$\sum_{m=-2}^{2} \sigma(m) = \sum_{n=-2}^{2} \sigma(n) = 1 \tag{72}$$

and must be symmetric

$$\sigma(i) = \sigma(-i). \tag{73}$$

The density of image pixels is reduced by four from one level to the next level up. Hence, an additional constraint called equal contribution requires all pixels at a given level to contribute the same total weight of $1/4$. The above constraints are satisfied when

$$\begin{aligned} \sigma(0) &= a \\ \sigma(1) &= \sigma(-1) = \frac{1}{4} \\ \sigma(2) &= \sigma(-2) = \frac{1}{4} - \frac{a}{2} \end{aligned} \tag{74}$$

where $a = 0.4$. It should be noted that the algorithm was slightly modified to fit our original image size of 512×512 pixels.

For the multi-resolution analysis we considered a maximum of six Gaussian pyramid levels. Taking the original image resolution and the sub-sampling of factor two into account, the highest level in the pyramid has a resolution of 16×16 pixels. Images of higher levels were not taken into account since the feature extraction algorithms do not work on such a small number of pixels. Figure 34 shows an illustrative example of the multi-resolution levels for a distorted 'Lena' image. Corresponding error maps are presented in Fig. 35 to further visualise the impact of the Gaussian pyramid on the distortions in the content. For better visualisation the downsampled images were expanded to original size using the pixel replication technique [217]. As one would expect, the Gaussian filter introduces some blur into higher level images.

7.2 Multiple-scale feature extraction

The five structural feature metrics f_i (see Section 3.1) are independently computed in all Gaussian pyramid levels, resulting in a large number of features. For this reason, we will in the following adopt a feature notation as $f_{i,l}$. Here, the subscript i, $i = 1, 2, \ldots, I$, denotes the i^{th} out of $I = 5$ feature metrics and subscript l, $l = 0, 1, \ldots, L$, denotes the l^{th} out of $L + 1 = 6$ Gaussian pyramid

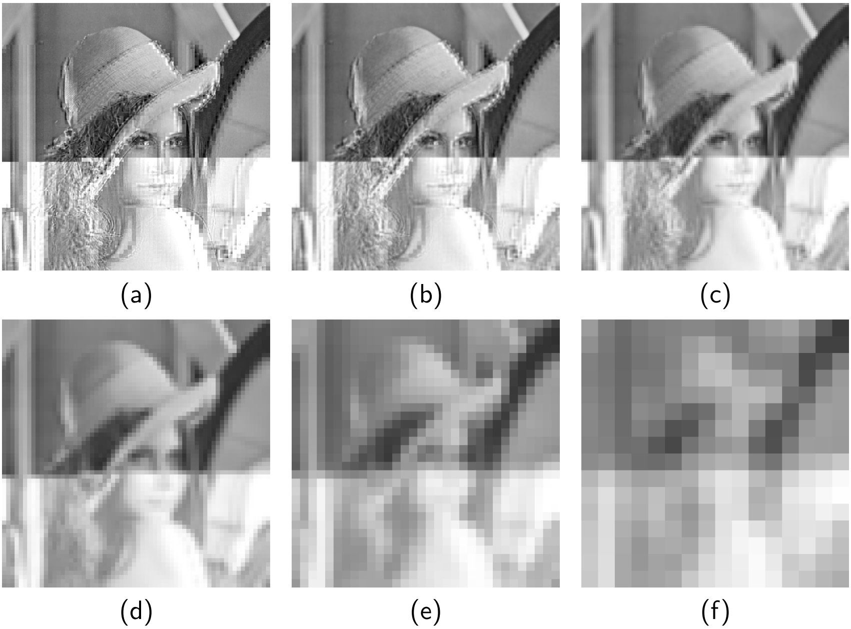


Figure 34: First six Gaussian pyramid levels of a distorted 'Lena' image: (a) g_0 (512×512), (b) g_1 (256×256), (c) g_2 (128×128), (d) g_3 (64×64), (e) g_4 (32×32), and (f) g_5 (16×16)).

levels. For instance, feature $f_{2,0}$ would be the blur metric f_2, computed in the base level g_0 of the Gaussian pyramid.

To assess the structural degradation in each pyramid level, the features $f_{r,i,l}$ and $f_{d,i,l}$ are, respectively, computed in the reference image and the distorted image and subsequently combined using a pooling function. In order to keep the RR information small it would be sensible to pool the reference features, $f_{r,i,l}$, independently from the distorted features, $f_{d,i,l}$. However, we found [183] that it is crucial to preserve the information about structural degradation for each pyramid level. For this reason, we compute difference features within all Gaussian pyramid level as follows

$$\Delta f_{i,l} = |f_{r,i,l} - f_{d,i,l}|. \tag{75}$$

Given the above, we have a total of 30 difference features $\Delta f_{i,l}$ which may

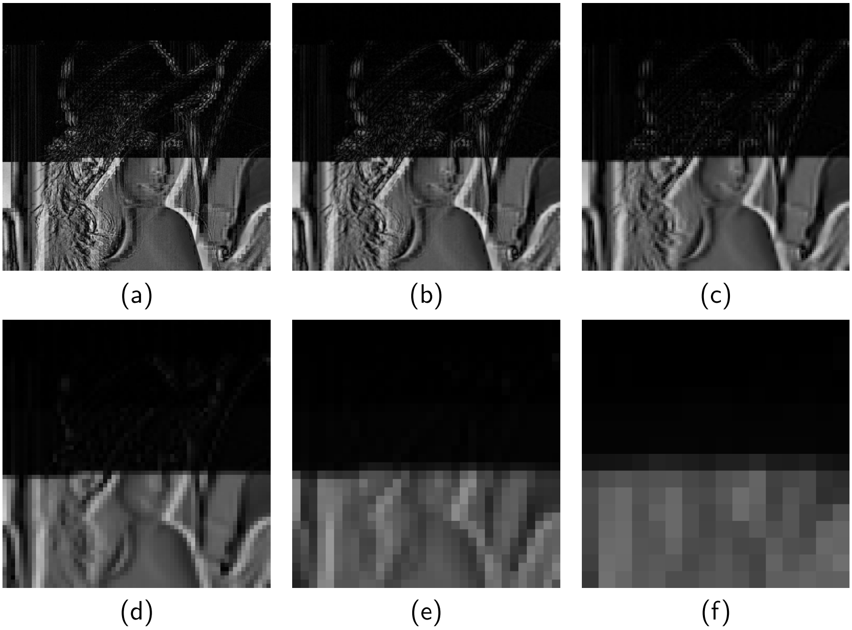


Figure 35: Error maps between the distorted images in Fig. 34 and the corresponding reference images: (a) g_0 (512×512), (b) g_1 (256×256), (c) g_2 (128×128), (d) g_3 (64×64), (e) g_4 (32×32), and (f) g_5 (16×16)).

contribute to the multiple-scale quality metric MSFQM. The number of features included in the metric impacts directly on the computational complexity of the metric and also determines the size of the RR information. Thus, it is desired to reduce the number of features by only taking into account the most relevant features at a given pyramid level.

7.3 Multiple-scale feature-based quality metric

The feature differences are pooled into the MSFQM metric which is in the following denoted as Θ_{MSF}. Similar to the NHIQM and L_p-norm metrics, we introduce perceptual relevance weights $w_{i,l}$ for all difference features $\Delta f_{i,l}$, taking into account that not all features, and thus the related structural degradations, have the same impact on perceived quality. The feature weights will also serve to eliminate the features that are of low relevance to MSFQM. Finally, a simple pooling function is defined as a weighted L_1-norm over all features across all pyramid levels to yield

$$\Theta_{MSF} = \sum_{l=0}^{L} \sum_{i=1}^{I} w_{i,l} \cdot \Delta f_{i,l}. \tag{76}$$

The pooling function in (76) has intentionally been kept simple to put no restrictions and bias on the optimisation process which will in the following be used to determine the optimal feature weights $w_{i,l}^{(opt)}$.

7.4 Optimisation of the perceptual relevance weights

The MOO framework as outlined in Section 5.1 is deployed here to determine the optimal feature relevance weights $w_{i,l}^{(opt)}$ for MSFQM. The MOO is conducted with respect to the same objectives O_A and O_G of improving the prediction accuracy on the training set $\mathcal{I}_T$ while maintaining the generalisation ability to the validation set $\mathcal{I}_V$, respectively. The Pearson linear correlation is here computed between MSFQM and the MOS from the subjective experiments as

$$\rho_P = \frac{\sum_k (\Theta_{MSF,k} - \overline{\Theta}_{MSF})(\mathcal{M}_k - \overline{\mathcal{M}})}{\sqrt{\sum_k (\Theta_{MSF,k} - \overline{\Theta}_{MSF})^2} \sqrt{\sum_k (\mathcal{M}_k - \overline{\mathcal{M}})^2}}. \tag{77}$$

Unlike in Section 5.1, we do not define a range of goals for objective O_G, since we want this objective to be very small to obtain a good generalisation for MSFQM. Thus, the goals are defined here as $O_A^* = -0.9$ and $O_G^* = 0.001$ and the magnitudes λ_A and λ_G are as in (45) set to the absolute value of the goals.

The optimal weights $w_{i,l}^{(opt)}$ determined from the MOO are shown in Fig. 36. It should be emphasised again, that these weights are considered to be optimal in a Pareto sense as we considered two conflicting objectives O_A and O_G in the MOO, where one objective can only be optimised at the cost of the other.

One can see from the Fig. 36, that in each level there are some features dominating over the other features providing, to some extent, insight into the impact on perceived quality of a feature at a given scale. For instance, the feature f_1 (relating to blocking artifacts) and the edge-based IA f_3 (relating to ringing artifacts) seem to have a strong impact up to level g_3. The feature f_2 (relating to blur artifacts) on the other hand dominates in the levels g_4 and g_5 which may be due to the blur induced through the Gaussian filtering in the higher pyramid levels. However, in general it seems that the feature metrics which are

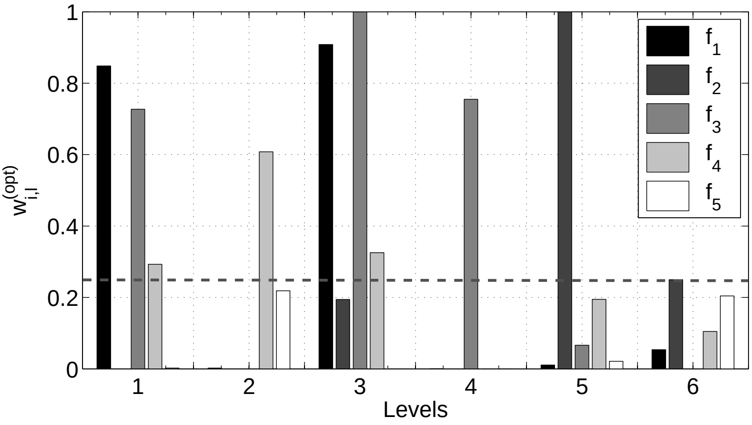


Figure 36: Optimal feature weights for six Gaussian pyramid levels (the dashed line indicates a threshold $\tau = 0.25$).

based on sharp edges, f_1 and f_3, dominate over the feature metrics that are based on measuring smooth transitions and gradients, f_2 and f_4. This is in line with the previous results for NHIQM (see Section 3.5.3) and its optimised versions (see Section 5.2). More surprising are the small weights for the feature f_5 which may be due to this block intensity shifts being indirectly accounted for by the other metrics in higher levels of the Gaussian pyramid.

7.5 Thresholding and feature elimination

It can be observed from Fig. 36 that many of the 30 weights are either zero or very close to zero. This may be related to the phenomenon that too many features would cause overfitting of the data to the training set [224]. The small magnitudes of some of the weights suggests, that the related features are of low importance to the overall metric MSFQM and thus, one can consider them for further elimination. This in turn results in computation of less features and therewith savings in valuable computational complexity and RR overhead. For this purpose, we introduce a weight thresholding as follows

$$w_{i,l}^{(opt)} = 0 \quad \text{for} \quad w_{i,l}^{(opt)} \leq \tau. \tag{78}$$

Thus, all weights $w_{i,l}^{(opt)}$ below threshold τ are set to zero and the related features $f_{i,l}$ are omitted from the metric computation. The resulting metric, that was subjected to a particular threshold τ, is denoted as $\Theta_{MSF}^{(\tau)}$. The compromise between prediction performance and savings in computational complexity of the metric with regards to different thresholds τ is discussed in the following section.

7.6 Quality prediction performance

The quality prediction performance of MSFQM is evaluated for different thresholds τ, using the quality prediction performance indicators from Section 3.6.3. For each $\Theta_{MSF}^{(\tau)}$ an exponential prediction function has been derived through curve fitting and the related predicted MOS, MOS$_{MSFQM}^{(\tau)}$, have been computed. The quality prediction performance indicators for both $\Theta_{MSF}^{(\tau)}$ and MOS$_{MSFQM}^{(\tau)}$ are presented in Table 17 for the training (T) and validation (V) sets. In addition to the quality prediction performance, the table also presents the number of discarded features (DF) and the number of discarded levels (DL) for a particular threshold τ. A discarded feature corresponds to a weight that has been set to zero by either the MOO or the thresholding. A level is discarded if all features within the level are discarded and if there is no higher pyramid level.

Table 17 shows that only 4 features are zero if no thresholding is used. As the threshold τ increases, more features are discarded from the metric computation. It should be observed that up to $\tau = 0.25$, the loss in prediction performance is minimal as compared to the number of discarded features. In particular, for $\tau = 0.25$ only 9 out of 30 features need to be computed. We also do not need to compute any feature in level g_5 and can thus discard this level. This means for $\tau = 0.25$, savings in computational complexity are large as compared to loss in prediction performance. However, if we further increase τ the loss in prediction performance as compared to the number of discarded features increases. Given the above, a threshold of $\tau = 0.25$ is thus considered to give the best compromise solution for MOS$_{MSFQM}^{(\tau)}$. The residual feature weights $w_{i,l}^{(0.25)}$ for a threshold of $\tau = 0.25$ are summarised in Table 18. In addition, the prediction performance indicators of the resulting metric MOS$_{MSFQM}^{(0.25)}$ are highlighted with bold font in Table 17 and the threshold $\tau = 0.25$ is illustrated in Fig. 36 by the horizontal dashed line.

In summary, the quality prediction performance of the MSFQM metric is improved in comparison to NHIQM and in fact, performs almost as well as the neural network based metric (see Chapter 6). The better performance of MSFQM compared to NHIQM comes at the cost of a higher complexity and larger

Table 17: Quality prediction performance indicators for $\Theta^{(\tau)}_{MSF}$ and $\text{MOS}^{(\tau)}_{MSFQM}$ when applying different thresholds τ.

			$\Theta^{(\tau)}_{MSF}$		$\text{MOS}^{(\tau)}_{MSFQM}$							
τ	DF	DL	$\rho_{P,T}$	$\rho_{P,V}$	$\rho_{P,T}$	$\rho_{P,V}$	$\rho_{S,T}$	$\rho_{S,V}$	$RMSE_T$	$RMSE_V$	$r_{0,T}$	$r_{0,V}$
0	4	0	0.893	0.893	0.926	0.933	0.914	0.947	8.819	10.078	0.483	0.6
0.1	15	0	0.892	0.899	0.926	0.936	0.912	0.952	8.834	10.033	0.483	0.55
0.15	16	0	0.892	0.9	0.926	0.936	0.912	0.952	8.836	10.03	0.483	0.55
0.2	18	0	0.892	0.903	0.924	0.936	0.913	0.952	8.93	10.119	0.483	0.5
0.25	**21**	**1**	**0.89**	**0.902**	**0.921**	**0.934**	**0.908**	**0.953**	**9.071**	**11.072**	**0.483**	**0.55**
0.3	22	1	0.878	0.901	0.917	0.921	0.898	0.959	9.329	11.812	0.533	0.55
0.35	23	1	0.878	0.898	0.917	0.919	0.898	0.959	9.335	11.931	0.533	0.6
0.65	24	1	0.879	0.896	0.917	0.917	0.896	0.955	9.32	12.016	0.533	0.6
0.75	25	1	0.861	0.904	0.902	0.901	0.883	0.944	10.08	11.489	0.533	0.5
0.8	26	1	0.858	0.883	0.901	0.89	0.879	0.926	10.138	12.19	0.55	0.7
0.85	27	1	0.598	0.685	0.692	0.783	0.759	0.741	17.027	13.438	0.833	0.6

Table 18: Remaining feature weights for $\Theta^{(0.25)}$.

	g_0	g_1	g_2	g_3	g_4	g_5
f_1	0.848	0	0.908	0	0	0
f_2	0	0	0	0	1	0
f_3	0.727	0	0.999	0.755	0	0
f_4	0.293	0.608	0.325	0	0	0
f_5	0	0	0	0	0	0

RR information, due to the Gaussian pyramid decomposition and nine instead of five extracted features. However, the combined MOO and thresholding procedure drastically reduced the number of features, as compared to a full set of features being extracted in each level, at only small cost of quality prediction performance.

8 Region-of-Interest Based Quality Assessment

The image quality metrics proposed in the earlier chapters all compute the quality scores based on the assumption that the content over the entire natural scene is of equal interest to the observer. Thus, it is implicitly assumed that distortions in different regions of the image contribute equally to an overall quality perception of the image, with respect to the interest of the content. It is, however, well known that natural scenes typically contain objects and regions that are of particularly high interest to an observer. Such regions-of-interest (ROI) include, for instance, humans and especially their faces, animals, and written text. One can therefore expect that distortions appearing in an ROI would have a larger impact on perceived quality degradations as compared to distortions appearing in the remainder of the image, the background (BG). This would be particularly true in the context of localised distortions caused by transmission errors, as we consider in the scope of this thesis, in comparison to global distortions resulting from source coding artifacts.

Given the above, we discuss in this chapter a framework to incorporate ROI awareness into existing image quality metrics. In this respect, the metrics are independently computed in the ROI and the BG to obtain quality measures for each region. The ROI and BG metrics are then subjected to a weighted pooling, resulting in an ROI aware quality metric. The framework does not require the code of an existing metric to be changed since the metrics are independently computed in their original form on both, the ROI and the BG. The MOO framework discussed in Section 5.1 is once again used to determine the optimal weights for ROI and BG metrics to improve the quality prediction accuracy and generalisation ability of the resulting metric. The benefits of applying the ROI framework is evaluated using NHIQM and two other image quality metrics, of which all do not take into account the impact of content saliency on the perception of structural distortions.

In analogy to MOS serving as a ground truth for image quality metric design, the design of the ROI framework needs to be based on a reliable ground truth with regards to the ROI in the test images. In this respect, we rule out automatic ROI detection algorithms, as they may cause possible ROI detection errors and thus, cause errors in the subsequent quality metric design. For this reason we conducted a subjective experiment for ROI identification, in which human observers were instructed to select ROI in the images they were presented. These subjectively selected ROI were then used as a basis for the ROI aware metric design.

In the following, we present the subjective ROI experiment that we conducted and analyse its outcomes. We then discuss the ROI awareness framework and evaluate the benefits of deploying it on three contemporary image quality metrics.

8.1 Subjective region-of-interest selections

We conducted a subjective ROI experiment which we refer to as experiment E3. In this experiment, human observers were asked to select ROI in the reference images that we used in the subjective quality experiments E1 and E2 (see Fig. 6). These hand-labelled ROI reveal insight as to which regions in the reference images are of particular interest to human observers. More importantly, they serve as a reliable ROI ground truth, similar to MOS being a quality ground truth, upon which to design ROI aware image quality metrics. In the following, the procedures of experiment E3 are discussed and the outcomes are analysed.

8.1.1 Details of experiment E3

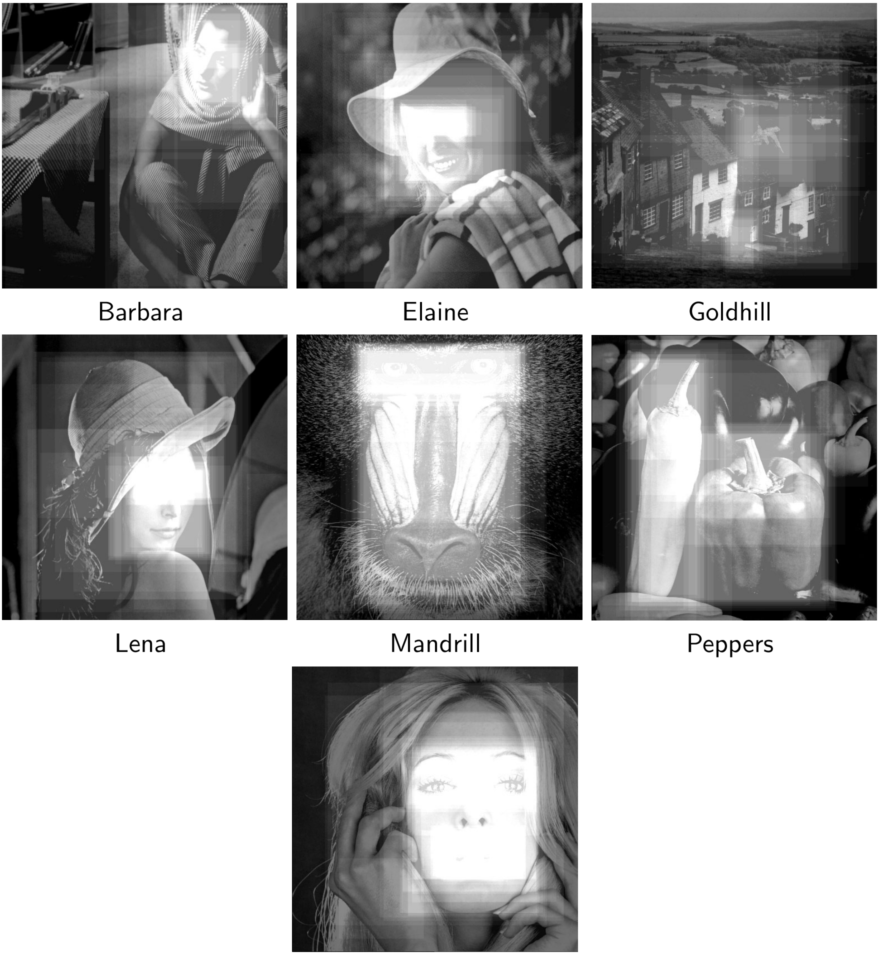
We conducted the subjective ROI experiment at BIT in Ronneby, Sweden. As with the quality experiments, E1 and E2, we had 30 non-expert viewers who participated, of which 17 were male and 13 were female. The viewers were presented the same set of reference images $\mathcal{I}_R$ that we used in experiments E1 and E2. The images were displayed on a 19" DELL screen at a viewing distance of 4 times the height of the images. The viewers' task was to select a region within each of the images that was of particular interest to them.

One training image was presented in order to explain the simple selection process and two stabilisation images were presented for the viewers to adapt to the selection process. The viewers were then presented the seven reference images in $\mathcal{I}_R$. We did not put any restrictions on the size of the ROI to be selected other than that the selected region needed to be a subset of the whole image. For simplicity, we considered only rectangular shaped ROI and allowed for only one ROI selection per image. We further allowed the viewers to re-select an ROI in case of dissatisfaction with the selected ROI. We did not impose any limits regarding the time needed for the ROI selection, however, given the simplicity of the ROI selection process most viewers were able to conduct the experiment within a few minutes.

The outcomes of the experiment enabled us to identify a subjectively determined ROI for each image in $\mathcal{I}_R$ and ultimately to deploy the ROI-based metric design framework, as proposed in Section 8.2. The experiment results are analysed in detail in the following sections.

8.1.2 ROI selections

The 30 ROI selections that we obtained for each reference image are visualised in Fig. 37. Here, all ROI selections have been added to the image as an intensity



Barbara

Elaine

Goldhill

Lena

Mandrill

Peppers

Tiffany

Figure 37: All 30 ROI selections for each of the reference images in $\mathcal{I}_R$ superimposed with an intensity shift.

shift and as such, a brighter area relates to more overlapping ROI and thus a higher interest in that particular region. In order to enhance the visualisation of the ROI, the images have been darkened before adding the ROI. For further reference, the coordinates of all ROI selections are also listed in Appendix B.

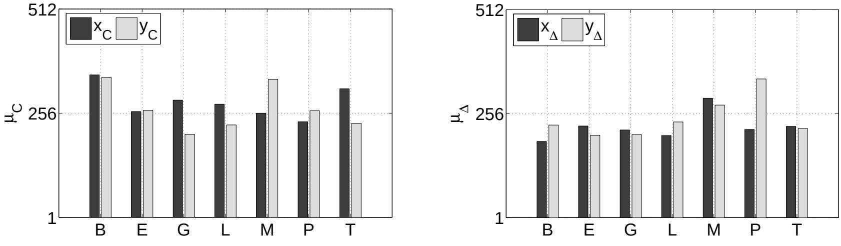
As one would expect, faces were of particular interest to the viewers and were thus primarily selected as the ROI. However, the size of the area in the image that is covered by the face seems to play an important role. If a whole person is shown in the image (for instance 'Barbara'), then the whole face is mostly chosen as the ROI. On the other hand, if most of the image is covered by the face (for instance 'Mandrill' or 'Tiffany'), then often details in the face are chosen rather than the whole face. In the case of 'Mandrill', such details mainly comprised of the eyes and the nose, whereas for 'Tiffany', along with the eyes the mouth was chosen most frequently.

In the case of a more complex scene, like 'Peppers', the agreement on an ROI between the viewers is by far less pronounced than in the case where a human or a human face is present. Here, different viewers have chosen different peppers as ROI or selected the three big peppers in the centre of the image. Most attention has actually been drawn to the two stems of the peppers, which may be due to their prominent appearance on the otherwise fairly uniform skins of the peppers. The disagreement between viewers is also apparent for a natural scene, such as 'Goldhill'. Only the man walking down the street seemed to be of interest to many viewers. Otherwise, varying single houses have been selected frequently as well as the whole block of houses.

8.1.3 Statistical analysis

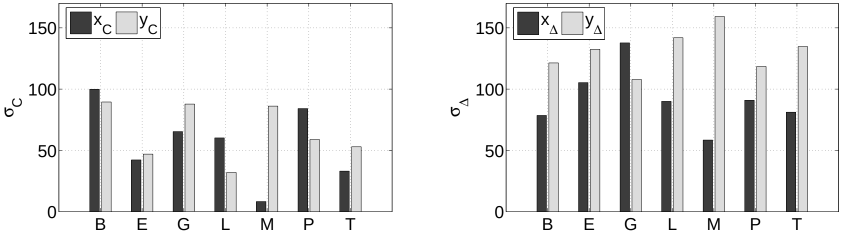
In order to gain more insight into the characteristics of the ROI selections we further analyse the ROI locations and ROI dimensions using simple statistics, such as the mean μ and the standard deviation σ of the ROI coordinates in horizontal and vertical direction. The results for the mean μ are summarised in Fig. 38 and for the standard deviation σ in Fig. 39. Here, x denotes the horizontal coordinate and y the vertical coordinate with the origin being in the bottom left corner of the image. Furthermore, x_C and y_C denote the ROI centre coordinates and x_Δ and y_Δ denote the ROI dimensions in x-direction and y-direction, respectively. The labels on the abscissa denote the first letters of the reference images in $\mathcal{I}_R$ (see Fig. 37).

In Fig. 38(a) it can be seen that the mean μ_C of the ROI centre coordinates, x_C and y_C, is around the image centre for most of the images. This may be somewhat expected since the salient region is often placed around the centre of a natural scene when, for instance, taking a photograph. The only exception here is the 'Barbara' image, for which the mean ROI is significantly shifted to the upper right corner towards the face. It is also worth noting that x_C for the



(a) (b)

Figure 38: Mean μ over all 30 ROI selections for: (a) centre coordinates and (b) horizontal (x-coordinate) and vertical (y-coordinate) dimensions.



(a) (b)

Figure 39: Standard deviation σ over all 30 ROI selections for: (a) centre coordinates and (b) horizontal (x-coordinate) and vertical (y-coordinate) dimensions.

image 'Mandrill' lies exactly in the horizontal centre of the image, which can be explained by the axis of symmetry of the 'Mandrill' face being centrally located in the horizontal direction.

Figure 38(b) reveals that the mean ROI dimensions, μ_Δ, for most images are very similar in both x- and y-direction. Interestingly, the 'Mandrill' image reveals much larger dimensions which is caused by many viewers selecting the whole face or the nose as ROI of considerable size. The large extent of the y-coordinate in the case of the 'Peppers' image is due to many selections of either all three big peppers or selections of the long pepper on the left.

The standard deviations σ_C of the ROI centre coordinates in Fig. 39(a) reveal information about the agreement of the viewers as to where the ROI is located, similar to CI with regard to MOS in subjective quality experiments. In this respect, a larger standard deviation, and thus a lower agreement, indicates that there may be either no dominant ROI or that there are multiple ROI present in the visual content. Given the above, the small values in case of 'Elaine', 'Lena', and 'Tiffany' further support earlier observations that faces are of strong interest to the viewers and that the agreement between viewers is high. On the other hand, larger standard deviations like for 'Goldhill' and 'Peppers' suggest that the identification of a dominant ROI is not as clear and thus, that the agreement between the viewers is lower. An exception is again given by the 'Barbara' image which comprises of a face but has, on the contrary, also the highest standard deviations. This may be due to the face being located in the periphery of the image and also due to other objects being present that some viewers found of interest, such as the object on the table to the left. With respect to the 'Mandrill' image it is interesting to point out the difference between the standard deviations in x- and y-directions. One can see that there is a strong agreement, that the ROI is located on the horizontal centre of the image, however, the agreement is low as to the vertical location of the ROI. This was also observed in the visual inspection of the ROI where many selections where found for the eyes, the nose, and the whole face, all of them being located on the horizontal centre but spread in the vertical direction.

Finally, comparing Fig. 39(b) to Fig. 39(a) reveals that the disagreement between viewers regarding the size of the ROI, quantified with the standard deviation of the ROI dimensions σ_Δ, seems to be large compared to the disagreement about location. It is further observed that for all images, apart from 'Goldhill', the disagreement is considerably higher in the vertical direction (y-coordinate) as compared to the horizontal direction (x-coordinate). This may be due to the viewers selecting either a whole body, a face, or parts of a face, where in all cases the width of the ROI selection is not as much affected as the height. This accounts in particular for images like 'Barbara', 'Lena', 'Mandrill', and 'Tiffany'.

8.1.4 Outlier elimination

In addition to the above observations, we found that for all seven reference images there were some ROI selections that were far away from the majority of the votes. In other words, the x- and/or y-coordinate of the centre of these ROI selections were numerically distant from the respective mean coordinates. We eliminated these, so called, outliers by adopting the criterion defined by the VQEG in [202] as follows:

$$|x_C - \mu_{x_C}| > 2 \cdot \sigma_{x_C} \qquad \text{or} \qquad |y_C - \mu_{y_C}| > 2 \cdot \sigma_{y_C}. \tag{79}$$

Table 19: Outlier ratios for the ROI selections in the reference images in $\mathcal{I}_R$.

Image	Barbara	Elaine	Goldhill	Lena	Mandrill	Peppers	Tiffany
r_0	$\frac{5}{30}$	$\frac{3}{30}$	$\frac{3}{30}$	$\frac{3}{30}$	$\frac{1}{30}$	$\frac{3}{30}$	$\frac{1}{30}$

As such, an ROI is considered to be an outlier if the distance of either x_C and/or y_C to the respective mean over all 30 selections is at least twice the corresponding standard deviation. Based on the number of eliminated outliers we define an outlier ratio for each of the images as

$$r_0 = \frac{R_0}{R} \tag{80}$$

where R_0 is the number of eliminated ROI selections and R is the number of all ROI selections.

The outlier ratios for all images are summarised in Table 19. One can see that the Barbara image exhibited the most outliers, which we believe is due to the location of the ROI in the periphery of the image. The least outliers can be observed for the 'Mandrill' and 'Tiffany' image, which are also the images with the face being present to a larger extent in comparison to the other face images. Hence, no other objects are present in the visual scene that may distract the viewers' attention away from the face.

8.1.5 Mean ROI

Similar to MOS from subjective quality experiments, we define mean ROI, ROI$_\mu$, for all seven reference images. Despite the variability of ROI selections in some of the images (see Section 8.1.2), we decided to only define one ROI$_\mu$ for each of the reference images. The reasons for this decision are threefold. Firstly, and most importantly, many of the ROI selections are overlapping or even include each other. For instance, in the case of the Tiffany image people mostly chose the eyes, the mouth, or the whole face. Thus, selecting the face as ROI includes both eyes and mouth. Similar observations were made for the other images. Secondly, in the context of wireless imaging we are aiming to keep the overhead and computational complexity low. Since a higher number of deployed ROI is directly related to an increased overhead, in terms of side information, and also an increased complexity, in terms of the number of computed metrics, we decided

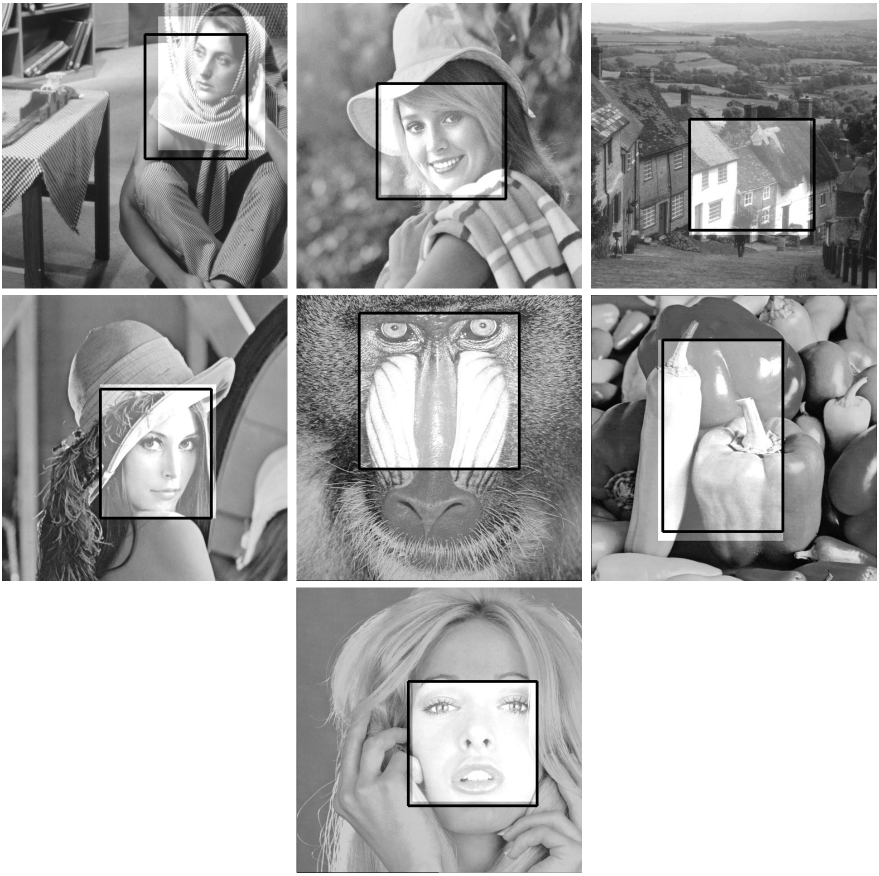


Figure 40: Mean ROI for the reference images in $\mathcal{I}_R$ (black frame: before outlier elimination; brightened area: after outlier elimination).

for only one ROI$_\mu$. Lastly, deploying only a single ROI$_\mu$ is in agreement with the subjective experiment in which we asked the viewers to select a single ROI.

Considering the above, we defined one ROI$_\mu$ for each image as the mean over all 30 ROI selections. In particular, the location of the ROI was computed as the mean over all ROI centre coordinates x_C and y_C. The size of the ROI was computed as the mean over all ROI dimensions x_Δ and y_Δ. The mean ROI

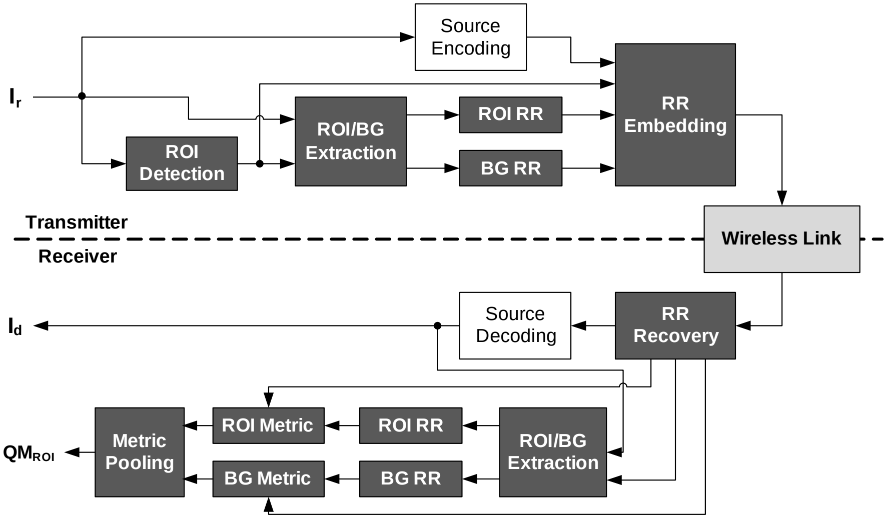


Figure 41: Reduced-reference quality assessment using the ROI aware image quality metric.

are shown in Fig. 40. Here, the black frames and the bright areas, respectively, indicate the ROI$_\mu$ before and after outlier elimination (see Section 8.1.4). One can see that the shift of the ROI$_\mu$ after outlier elimination is most prominent for the 'Barbara' image, which is in agreement with the largest number of outliers that this image exhibited (see Table 19). The ROI$_\mu$ after outlier elimination (bright area) are in the following used for the ROI aware image quality metric design.

8.2 Region-of-interest aware quality metric

An overview of the RR visual quality assessment framework (see Fig. 4) adapted to the deployment of the ROI aware quality metric is presented in Fig. 41. The first step at the transmitter is to detect the ROI in the image content. To facilitate online ROI detection, one may in a practical application deploy automated ROI detection algorithms to perform this task [138, 148]. However, in order to avoid ROI detection errors and subsequent errors in the metric design, we use here the mean ROI, ROI$_\mu$, from subjective experiment E3 as a ground truth for the ROI in the reference images. The RR information is then extracted from the ROI and BG of the original image. The ROI and BG RR information is then transmitted as side information along with the ROI locations. The respective RR information is extracted from the ROI and the BG at the receiver and the quality metric is independently computed on both, ROI and BG. In a final pooling stage, the ROI and BG metrics are combined to establish the ROI aware quality metric.

In the following, we discuss the different steps included in the ROI aware metric design and evaluate the performance of the framework on three image quality metrics; NHIQM [16], SSIM [14], and VIF [73].

8.2.1 Segmentation into ROI image, I_{ROI}, and BG image, I_{BG}

The mean ROI coordinates after outlier elimination were used to segment all reference and distorted images into ROI images, I_{ROI}, and BG images, I_{BG}. In particular, the ROI images were obtained by cutting out the area according to the mean ROI centre coordinates, μ_C, and the mean ROI dimensions, μ_Δ (see Fig. 38). The BG images then comprised of the remainder of the images with the pixels in the ROI set to zero.

8.2.2 ROI and background pooling

Let Φ be again our general definition of an image quality metric. Furthermore, let Φ_{ROI} be a metric computed on the ROI image, I_{ROI}, and Φ_{BG} be a metric computed on the BG image, I_{BG}. The ROI aware quality metric $\Phi^{(RA)}$ is then obtained as a weighted combination of the metrics computed in ROI and BG, Φ_{ROI} and Φ_{BG}. In particular, we deploy a variant of the Minkowski metric [210] in order to obtain the final metric Φ_{RA} as

$$\Phi^{(RA)}(\omega, \kappa, \nu) = [\omega \cdot \Phi_{ROI}^{\kappa} + (1-\omega) \cdot \Phi_{BG}^{\kappa}]^{\frac{1}{\nu}} \tag{81}$$

with $\omega \in [0,1]$ and $\kappa, \nu \in \mathbb{Z}^+$. For $\kappa = \nu$, the expression in (81) is also known as the weighted Minkowski metric. However, we have found that better quality prediction performance can be achieved by allowing for the Minkowski parameters κ and ν to have different values.

The weight ω regulates the contribution of Φ_{ROI} and Φ_{BG} to the overall quality metric $\Phi^{(RA)}$. With regards to our earlier conjecture that artifacts in the ROI may be perceived more annoying than in the background, one would expect the weight ω to have a value > 0.5, thus, giving the ROI metric a higher impact on the overall metric in comparison to the BG metric.

In the scope of this thesis, we derive the optimal weight $\omega^{(opt)}$ and the optimal Minkowski parameters $\kappa^{(opt)}$ and $\nu^{(opt)}$ for three image quality metrics, NHIQM,

Table 20: Optimal parameters of the ROI aware quality metrics $\Phi^{(RA)}$.

	$\omega^{(opt)}$	$\kappa^{(opt)}$	$\nu^{(opt)}$
$\Delta_{NHIQM}^{(RA)}$	0.593	3.142	4.066
SSIM$^{(RA)}$	0.823	4.062	0.534
VIF$^{(RA)}$	0.978	2.928	0.798

SSIM, and VIF, thus yielding $\Phi^{(RA)}(\omega, \kappa, \nu) \in \{\Delta_{NHIQM}^{(RA)}, \text{SSIM}^{(RA)}, \text{VIF}^{(RA)}\}$. The procedure to find the optimal parameters for these three metrics is discussed in the following section.

8.2.3 Optimisation of perceptual relevance weights

The MOO framework introduced in detail in Section 5.1 is deployed here once again to determine the optimal parameters $\omega^{(opt)}$, $\kappa^{(opt)}$, and $\nu^{(opt)}$. The MOO is conducted with respect to the same objectives O_A and O_G and with the Pearson linear correlation being computed between $\Phi^{(RA)}$ and the MOS from the subjective experiments E1 and E2 as

$$\rho_P = \frac{\sum_k (\Phi_k^{(RA)} - \overline{\Phi}^{(RA)})(\mathcal{M}_k - \overline{\mathcal{M}})}{\sqrt{\sum_k (\Phi_k^{(RA)} - \overline{\Phi}^{(RA)})^2} \sqrt{\sum_k (\mathcal{M}_k - \overline{\mathcal{M}})^2}}. \tag{82}$$

The goals are defined here as $O_A^* = -0.9$ and $O_G^* = 0.001$ and the magnitudes λ_A and λ_G are set to the absolute values of the goals.

The optimal weights $\omega^{(opt)}$, $\kappa^{(opt)}$, and $\nu^{(opt)}$ determined from the MOO are shown in Table 20. As discussed in Section 5.1, these parameter sets are trade-off solutions from a number of noninferior solutions and thus, the parameters are considered to be optimal in a Pareto sense. It is interesting to note that the weights $\omega^{(opt)}$ are larger than 0.5 for all three metrics, which confirms our earlier conjecture that the ROI metrics should receive a higher weight due to the artifacts in the ROI being more annoying than in the background. Also, one can see that the optimal parameters for SSIM$^{(RA)}$ and VIF$^{(RA)}$ are fairly similar, meaning, that both have a $\omega^{(opt)}$ at the higher end of the scale and a significantly larger value for $\kappa^{(opt)}$ as compared to $\nu^{(opt)}$. This is somewhat not unexpected since it has been shown [232] that both metrics have very strong relationships in their methodologies of objectively assessing perceived quality.

Table 21: Comparison of quality prediction performance between the original quality metrics Φ and the ROI aware quality metrics $\Phi^{(RA)}$.

Metric	Metric		Predicted MOS							
name	$\rho_{P,T}$	$\rho_{P,V}$	$\rho_{P,T}$	$\rho_{P,V}$	$\rho_{S,T}$	$\rho_{S,V}$	$RMSE_T$	$RMSE_V$	$r_{0,T}$	$r_{0,V}$
Δ_{NHIQM}	0.843	0.841	0.892	0.888	0.867	0.892	10.579	14.035	0.583	0.7
$\Delta_{NHIQM}^{(RA)}$	0.896	0.896	0.928	0.886	0.892	0.92	8.69	11.194	0.483	0.5
SSIM	0.582	0.434	0.697	0.628	0.558	0.347	16.733	18.498	0.7	0.9
SSIM$^{(RA)}$	0.732	0.623	0.734	0.614	0.657	0.486	15.847	17.575	0.667	0.9
VIF	0.713	0.727	0.789	0.788	0.813	0.729	14.486	13.892	0.817	0.7
VIF$^{(RA)}$	0.835	0.834	0.863	0.794	0.872	0.776	11.795	13.006	0.633	0.75

8.2.4 Quality prediction performance

The quality prediction performance of the ROI aware metrics is evaluated using the performance indicators introduced in Section 3.6.3. The results are presented in Table 21 for all three metrics and for the training and validation sets of images. In addition, predicted MOS are presented that have been derived for all ROI aware metrics. The results are compared to the quality prediction performance of the original metrics, as they were presented in Table 13.

From Table 21 one can see that the prediction accuracy of the ROI aware metrics, $\Delta_{NHIQM}^{(RA)}$, SSIM$^{(RA)}$, and VIF$^{(RA)}$, could be improved as compared to the original metrics Δ_{NHIQM}, SSIM, and VIF, on both the training (T) and validation (V) set. This is apparent not only in the Pearson linear correlation coefficient ρ_P but also in the RMSE. In particular, an improvement of about 5% for ρ_P can be observed for $\Delta_{NHIQM}^{(RA)}$ on both the training and validation set while maintaining the excellent generalisation ability of Δ_{NHIQM}. The improvement in ρ_P is even better for SSIM$^{(RA)}$ for which ρ_P could be increased by about 15-20%. Here, the generalisation of SSIM$^{(RA)}$ is not quite as good which may be due to SSIM already showing a lower correlation on the validation set, as compared to the training set. Finally, VIF$^{(RA)}$ shows both a significant improvement in ρ_P of about 12% and a well maintained generalisation.

Similar observations as for the Pearson correlation ρ_P can also be done for the Spearman correlation ρ_S, which is improved for all three metrics, $\Delta_{NHIQM}^{(RA)}$, SSIM$^{(RA)}$, and VIF$^{(RA)}$, on both the training and validation set. However, the improvement seems to be less prevalent as for the Pearson correlation and also, the generalisation ability is worse in the case of SSIM$^{(RA)}$ and VIF$^{(RA)}$. Both phenomena may be explained by the optimisation being performed with respect to the Pearson correlation rather than the Spearman correlation.

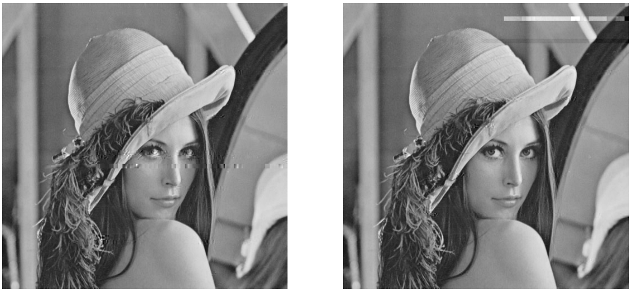
8.2.5 Illustrative examples

Given the results from the previous section, we have shown that the framework for ROI aware metric design was successfully deployed to three contemporary image quality metrics; Δ_{NHIQM}, SSIM, and VIF. To further illustrate the quality prediction performance improvement of the considered metrics we take a closer look at the prediction values of all three metrics for particular images, 'Lena' and 'Tiffany', that are shown in Fig. 42 and Fig. 43, respectively. In both figures, the images labelled with (a) contain mainly artifacts within the mean ROI, ROI$_\mu$, determined from the ROI experiment E3 (see Section 8.1.5) whereas the images labelled as (b) exhibit artifacts mainly in the BG. The images were selected taking into consideration a similar degree of distortion between the ROI distorted and the BG distorted images, thus, facilitating a better evaluation of the impact of the ROI on perceived quality. The related quality predictions of Δ_{NHIQM}, SSIM, and VIF are given in Table 22 and Table 23 for image 'Lena' and 'Tiffany', respectively. In the tables, the original quality metrics Φ are given as well as the ROI aware quality metrics $\Phi^{(RA)}$, along with all predicted MOS. It should be noted that for Δ_{NHIQM} and $\Delta_{NHIQM}^{(RA)}$ a higher value relates to lower quality, whereas for the other metrics and the predicted MOS, a higher value indicates higher quality.

In the ROI distorted 'Lena' image in Fig. 42(a) one can observe several distorted rows crossing her face around the eyes and the nose. On the other hand, in the BG distorted 'Lena' image in Fig. 42(b) there are artifacts located in the upper right corner, outside of the ROI. From Table 22 one can see that the original quality metrics Δ_{NHIQM}, SSIM, and VIF and their related predicted MOS judge the artifacts between the two images to have a similar impact on the overall quality of the image. This does not agree with the corresponding MOS, which exhibit a difference between the images, with the ROI distorted image being of about 10% lower perceived quality. These MOS are particularly interesting, since from a pure visibility point of view one would judge the artifacts in Fig. 42(b) to be more apparent, as compared to the artifacts in Fig. 42(a). However, the location of the artifacts in the eyes in Fig. 42(a) seems to outweigh this fact and therefore results

Table 22: Comparison of the ROI aware quality metrics, $\Phi^{(RA)}$, and the related original quality metrics, Φ, for image 'Lena'.

Metric	Artifact	Quality metric			Predicted MOS			MOS
type	location	Δ_{NHIQM}	SSIM	VIF	MOS$_{NHIQM}$	MOS$_{SSIM}$	MOS$_{VIF}$	
Φ	ROI	0.095	0.97	0.969	72.713	76.23	70.397	49.6
	BG	0.094	0.972	0.969	72.884	77.176	70.388	59.833
$\Phi^{(RA)}$	ROI	0.158	0.656	0.758	59.101	67.088	57.046	49.6
	BG	0.098	0.716	0.928	72.081	70.742	73.551	59.833



(a) (b)

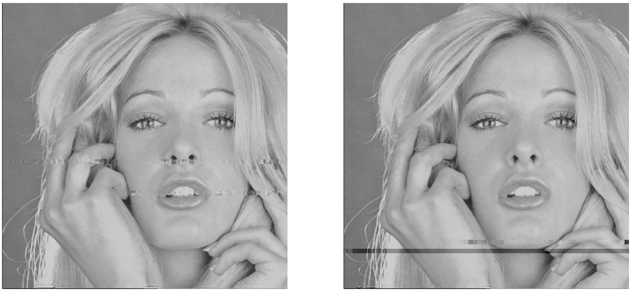
Figure 42: Distorted 'Lena' images, illustrating the annoyance of artifacts in the ROI as compared to the BG: (a) artifacts in ROI and (b) artifacts in BG.

in a stronger perceived quality degradation compared to Fig. 42(b). Unlike the original metrics, the ROI aware metrics $\Phi^{(RA)}$ and their related predicted MOS better account for this quality difference, as can be observed from the difference of these metrics between the ROI distorted image and the BG distorted image.

Similar observations as for the image 'Lena' can also be made for the image 'Tiffany'. In the ROI distorted image in Fig. 43(a), we find artifacts along the mouth and nose and some additional artifacts in the eyes. In the BG distorted image in Fig. 43(b) there are artifacts in the lower part of the image. Again, the visibility of artifacts is considered to be higher in Fig. 43(b) as compared to Fig. 43(a). This is also represented by all three original quality metrics in Table

Table 23: Comparison of the ROI aware quality metrics, $\Phi^{(RA)}$, and the related original quality metrics, Φ, for image 'Tiffany'.

Metric	Artifact	Quality metric			Predicted MOS			MOS
type	location	Δ_{NHIQM}	SSIM	VIF	MOS$_{NHIQM}$	MOS$_{SSIM}$	MOS$_{VIF}$	
Φ	ROI	0.084	0.971	0.956	72.101	76.726	67.72	47.27
	BG	0.116	0.97	0.939	66.576	76.362	64.42	55.33
$\Phi^{(RA)}$	ROI	0.183	0.683	0.62	54.331	68.695	46.374	47.27
	BG	0.096	0.912	0.972	72.537	83.856	78.533	55.33



(a) (b)

Figure 43: Distorted 'Tiffany' images, illustrating the annoyance of artifacts in the ROI as compared to the BG: (a) artifacts in ROI and (b) artifacts in BG.

23, which judge the quality of the ROI distorted image to be slightly better as compared to the BG distorted image. However, this is again in disagreement with the related MOS, as the observers judged the ROI distortions to result in a 8% larger quality degradation, as compared to the BG distortions. The human quality ratings are better captured by the ROI aware metrics $\Phi^{(RA)}$, as the judgment order of the original metrics is inverted to exhibit the same order as the MOS.

9 Eye Tracking and Subjective Image Quality Experiments

In the previous chapter, we considered ROI as a means to identify the objects and regions within a set of images that were of particular interest to a number of human observers. In this case, higher semantic processes are actively deployed by the observers to make an informed decision with respect to the perceived interest of objects in the scene. On the other hand, the attention of observers when viewing a natural scene is not only driven by the recognition of various objects (e. g. faces, houses, etc.) but also by more basic image features, such as colour, shape, size, orientation, and texture (see discussion in Section 1.5). The visual attention (VA) to these features is usually stimulus driven and relates to bottom-up attentional processes in the HVS. Thus, subjective experiments, such as the ROI experiment that we conducted, are not sufficient to identify the stimulus driven behaviour of human observers, as the observers were asked to actively perform the selection task.

The most common procedure to obtain data that reflects the VA of human observers is by means of eye tracking experiments [124]. In this chapter, we report about an eye tracking experiment that we conducted at the University of Western Sydney (UWS) in Campbelltown, Australia, to obtain subjective data related to VA of human observers in natural images. The experiment comprised of two parts, of which both served different purposes.

The first part was an eye tracking experiment in which the observers were shown a number of natural images from three different image quality databases. The experiment was conducted under task-free condition, meaning, that the participants were not instructed with any particular tasks but were asked to just view the images they were presented. The resulting gaze patterns recorded during this experiment reflect the saliency of the corresponding image content. Hence, they facilitate a better understanding of human viewing behaviour of natural image content and may serve to incorporate VA models into image quality metrics.

The second part of the experiment was conducted to gain a better understanding of the human viewing behaviour when judging image quality. This is of particular interest in the context of wireless imaging applications to identify the degree to which the complex distortion patterns alter the viewing behaviour. For this purpose, we recorded gaze patterns of human observers when judging the quality of the test images used in experiments E1 and E2. As the eye tracking was conducted under quality assessment task, the gaze patterns do not only reflect the content of the images but also inherently account for the viewing behaviour of the observers when performing the quality rating task. To gain also a deeper insight into the confidence of observers during quality assessment, we further collected confidence scores provided by the participants and recorded the response times needed to give a particular judgement. These quantities are considered to be valuable complements to CI as a measure of reliability of MOS.

The task-free experiment and the task-based experiment are referred to as experiments E4a and E4b, respectively. In the following sections, we first discuss the details that were common to both experiments, such as the laboratory environment, the eye tracker hardware, and the viewer panel. We then discuss the particulars for each experiment including the test material and the test procedures. This chapter thus serves to introduce the details of the experiments. An elaborate analysis of the results is provided in Chapter 10 and in Chapter 11.

9.1 Details common to experiments E4a and E4b

Both experiments E4a and E4b were conducted at the School of Computing and Mathematics of UWS. The experiment procedures were designed according to ITU-R Rec. BT.500-11 [19].

9.1.1 Laboratory environment

The experiments were conducted in a laboratory with low light conditions. A Samsung SyncMaster monitor of size 19" with a native screen resolution of 1280×1024 pixels was used for image presentation. The screen was placed in front of light grey blinds and any objects around the monitor that may have distracted the observers' attention were removed. The eye tracker was installed under the screen and the participants were seated at a distance of approximately 60 cm, corresponding to about four times the height of the presented images. A Snellen chart was used to test the visual acuity of each participant prior to the first session.

9.1.2 Eye tracking hardware

An EyeTech TM3 eye tracker [233] was used to record the gaze of the human observers during both experiments. A photo of the TM3 eye tracker is shown in Fig. 44. The TM3 consists of an infrared camera and two infrared light sources, one on either side of the camera. The accuracy with which the gaze is recorded lies within 1 degree of visual angle (dva). The eye tracker records gaze points (GP) at a rate of about 40-45 GP/s. A calibration of the TM3 for each person is done before every session using a 16 point calibration screen.

Figure 44: EyeTech TM3 eye tracker [233] used in experiments E4a and E4b.

9.1.3 Viewer panel

A total of 15 people participated in the experiments who were mainly staff and students from the Campbelltown campus of the University of Western Sydney. The age ranged from 20 to 60 years with an average age of 42 years. Nine participants were male and six were female. Twelve participants stated that they were not involved with image analysis in their professional and private activities. Three participants were or had been earlier somewhat involved with image analysis; one with face recognition, one with astronomical imaging, and one with image restoration.

9.2 Details of experiment E4a

Experiment E4a was conducted to obtain gaze patterns of a number of human observers when viewing natural images under a task-free condition. These gaze patterns are considered to be highly useful for image quality researchers to incorporate models of VA into their quality metrics. The details of the experiment are outlined in the following.

9.2.1 Test material

In this experiment, the observers were presented the reference images of three well known image quality databases; the IRCCyN/IVC database [36], the LIVE database [37], and the MICT database [35]. In addition to the publicly available test images in these databases, MOS from subjective quality experiments are also provided for each image. The additional eye tracking data from our experiment E4a then further facilitates to directly relate the quality perception (MOS) to the saliency of the images (gaze patterns).

The three databases contain a total of $10 + 29 + 14 = 53$ reference images, however, 11 images have been used both in the LIVE and MICT databases. We used these images only once and as such, a total of 42 images was presented to the participants. The three databases are introduced in the following and a summary of the databases is additionally provided in Table 24:

IRCCyN/IVC Database: The IRCCyN/IVC database [36] has been established by the Image and Video Communication (IVC) group of the Institut de Recherche en Communications et en Cybernétique (IRCCyN) in Nantes, France. Ten images of dimension 512×512 pixels were selected to create a total of 235 distorted images using JPEG coding, JPEG2000 coding, locally adaptive resolution coding, and blurring. Fifteen observers then rated the quality of the distorted images as compared to the reference images using the double stimulus impairment scale (DSIS) [19].

LIVE Database: The LIVE database [37] is provided by the Laboratory for Image & Video Engineering (LIVE) of the University of Texas at Austin, USA. Here, JPEG coding, JPEG2000 coding, Gaussian blur, white noise, and fast fading were applied to create a total of 779 distorted images from 29 reference images. The image widths are in the range of 480-768 pixels and the image heights are in the range of 438-720 pixels. Between 20-29 observers rated the quality of each image using a single stimulus (SS) assessment method.

MICT Database: The MICT database [35] has been made available by the Media Information and Communication Technology (MICT) Laboratory of the University of Toyama, Japan. The MICT database contains 168 distorted images obtained from 14 reference images using JPEG and JPEG2000 source encoding. The image widths and heights are, respectively, in the ranges 480-768 and 488-720 pixels. Sixteen observers rated the quality of the test images using the adjectival categorical judgement (ACJ) method [19].

9.2.2 Test procedure

The 42 images were presented to the participants in random order. Each image was shown for 12 s with a mid-grey screen shown between images for 3 s. The mid-grey screen contained a fixation point in the center which the participants were asked to focus on. As such, it was assured that the observation of each image started at the same location. Given the presentation times and the number of images, the length of each session was about 10 min.

Table 24: Overview of the IRCCyN/IVC, LIVE, and MICT databases.

Database	IRCCyN/IVC [36]	LIVE [37]	MICT [35]
Number of reference images	10	29	14
Number of test images	235	779	168
Image widths	512	480-768	480-768
Image heights	512	438-720	488-720
Number of observers/image	15	20-29	16
Assessment method	DSIS	SS	ACJ

During the whole experiment the gaze of the observers was recorded using the TM3 eye tracker. The participants were instructed to simply watch the images they were presented, without specifying any further task. As such, the gaze patterns recorded during this experiment serve as a ground truth regarding the saliency of the presented images.

9.2.3 Recorded data and post-processing

The TM3 tracks both eyes at the same time and records individual GP for each eye. An overall GP is then computed as the average between the two eyes. In addition, the TM3 records if an eye has been tracked in a particular time instance (eye status flag). If none of the two eyes could be tracked (for instance due to blinking) then the previous GP is recorded. Given the recording rate of the eye tracker and the presentation time of each image, we recorded about 480-540 GP per person and image. Thus, the total number of GP of all participants for a particular image adds up to approximately 7200-8100 GP/image. An example image visualising the total number of GP in an image is presented in Fig. 45.

The recorded data needs to be post-processed for two important reasons. Firstly, to eliminate GP that do not contribute to VA. From a technical viewpoint, such GP may include recordings for which the eye status flag indicated that none of the eyes could be tracked in a particular time instance. On the other hand, from a biological viewpoint, GP recorded during saccades need to be eliminated as VA is suppressed during these rapid eye movements.

Secondly, the GP need to be transformed into a more meaningful representation of the observers' attention. This can be comprehended when looking at

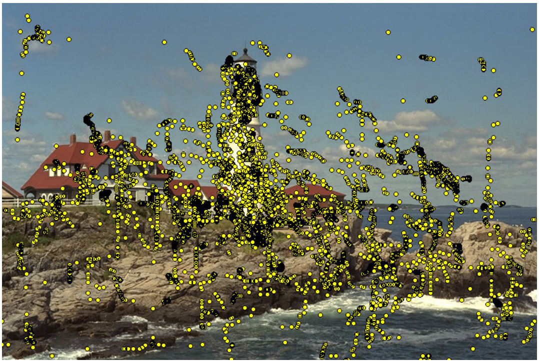


Figure 45: Visualisation of the GP of all 15 participants in experiment E4a.

Fig. 45, where the vast amount of GP does not easily reveal what the observers actually focused on. For this reason, the gaze patterns are typically converted into visual fixation patterns or saliency maps. The post-processing of the GP into these representations is discussed in Section 11.1 where the resulting saliency maps are consulted for further analysis.

9.3 Details of experiment E4b

Experiment E4b was conducted with the aim to gain a better understanding of human behaviour when judging image quality. This is particularly interesting in the context of wireless imaging distortions, as the range of distortions and the various distortion distributions constitute a difficult scenario for quality assessment. To achieve this goal, we recorded in addition to the quality scores also confidence scores and response times of the observers. As in experiment E4a, the TM3 eye tracker was used to record the gaze patterns of the observers to obtain valuable information with regards to the viewing behaviour during quality assessment.

9.3.1 Test material

In this experiment, the participants were shown the reference and distorted images from the image sets $\mathcal{I}_R$, $\mathcal{I}_1$, and $\mathcal{I}_2$, that were already used in experiments E1 and E2. For this purpose, two equal-sized sets $\mathcal{I}_A$ and $\mathcal{I}_B$ were created randomly from $\mathcal{I}_1$ and $\mathcal{I}_2$. In addition, the reference images $\mathcal{I}_R$ were randomly mixed into both $\mathcal{I}_A$ and $\mathcal{I}_B$, resulting in each of the sets containing a total of 47 images.

Table 25: Subjective rating scales for quality and confidence as used in experiment E4b.

Quality	Score	Confidence
Very Good	5	Very High
Good	4	High
Fair	3	Medium
Bad	2	Low
Very Bad	1	Very Low

9.3.2 Test procedure

The observers were presented the image sets $\mathcal{I}_A$ and $\mathcal{I}_B$ in random order in two consecutive sessions of about 10 min each. Each image was shown for approximately 8 s with a 5 s mid-grey screen presented between images. The TM3 eye tracker was used to record the gaze patterns of the participants throughout the experiment.

As we recorded the gaze patterns of the observers using the eye tracker, we conducted the quality assessment using a single stimulus method, thus having only one image presented on the screen at a time. The participants were asked to rate the image quality on a 5-point scale, with 5 being the highest quality. To minimise distraction of the participants during image viewing, and consequently to reduce unwanted alternation of the gaze patterns, the participants were instructed to do the rating during the grey screen presented between the images

In addition to the quality scores (QS) the participants were asked to provide confidence scores (CS) on a 5-point scale, as a measure of how difficult is was to judge the quality of a particular image. The observers were thus instructed to provide a higher CS if the corresponding quality judgement was considered to be easy, and vice versa. Both the quality scale and the confidence scale, as used in the experiment, are shown in Table 25. Here, the scales have intentionally been laid out to be very similar to ease the quality and confidence rating task for the observers.

As an additional and non-intrusive measure of confidence, the response times (RT) that the participant took to provide both the QS and the CS have been recorded by the experimenter. The RT was measured as the total time from the appearance of the grey screen after image presentation to the final judgement of

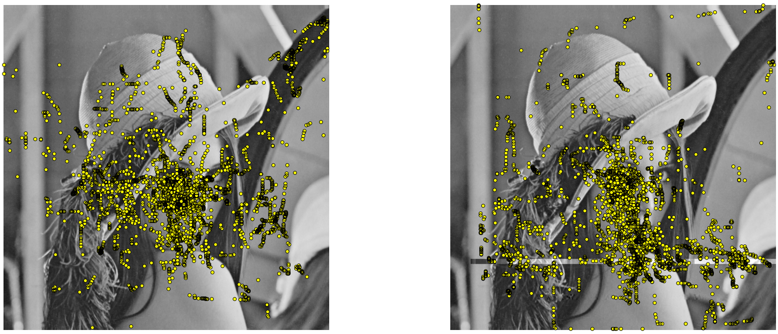


Figure 46: Visualisation of the GP of all 15 participants in experiment E4b for: (a) 'Lena' reference image and (b) 'Lena' distorted image.

both QS and CS by the observer. The participants were not made aware of the RT being recorded, as this would have possibly impacted on their rating behaviour.

In order for the participants to have an idea about the range of distortions in the test images and to adapt to the assessment procedure, a set of 7 training images was shown prior to the test images of the first session. The training images covered a wide range of artifact types and severities.

9.3.3 Recorded data

Given the number of observers, we recorded 15 QS, CS, and RT values for each image. The analysis of these quantities and in particular the relationships amongst them is discussed in detail in Chapter 10. Fifteen gaze patterns were further recorded for each image. Considering the recording rate of the eye tracker and the image presentation time, we recorded about 320-360 GP per person and image. Thus, the total number of GP for a particular image adds up over all participants to approximately 4800-5400 GP/image. Two example images visualising the total number of GP are presented in Fig. 46(a) and Fig. 46(b), respectively, for the 'Lena' reference image and for a distorted 'Lena' image. From the gaze patterns one can already observe a tendency of the GP to be shifted towards the distortion in the lower half of the image. However, as with the GP from experiment E4a, these GP need further processing into visual fixation patterns and saliency maps, which is explained in Section 11.1.

10 Observer Confidence During Image Quality Assessment

Rating the quality of images is not always an easy task for a human observer, as there can be many different types of distortions introduced into the image during the processes of capture, source coding, communication, and display. In particular, the error-prone transmission channel can be a source of a variety of different artifacts being present within the same image, thus, making the task of quality judgement significantly more difficult. This is further complicated by the often localised distortion distributions, in which case the observer may experience difficulty to judge overall image quality as a trade-off between distorted regions and otherwise undistorted regions of the image. In this respect, MOS obtained from subjective experiments have a varying level of reliability as to how well they represent the subjective perception of image quality as an average over a population of observers. Furthermore, MOS alone do not reflect the difficulty that an observer experienced when judging image quality.

In order to measure the reliability of a particular MOS, confidence intervals (CI) and related standard errors (SE) are usually computed to quantify the variation in quality rating between participants. A smaller CI reflects stronger agreement between the observers whereas a larger CI suggests a stronger disagreement between the observers with respect to the perceived quality of the rated images. However, there are different sources of variation that are reflected by CI and SE of which we consider three of the most influential ones to be the following:

S1: **Preference**: The perception of visual quality is subjective and as such, can be very different from observer to observer. Thus, there is usually a disagreement between the ratings provided by different observers related to their preference of different artifacts. This is particularly given in the context of quality assessment between images containing various distortion types and distributions. For instance, some observers may find blocking artifacts to be more annoying compared to blur artifacts, but other observers may perceive blur to be more annoying. Therefore, the personal preference of different observers regarding the distortions contained in the images is one source of variation contained in CI and SE.

S2: **Detection**: Depending on the strength of the distortions and the masking effects of the underlying visual content, some observers may detect distortions that others do not detect. As such, the observers that detected the distortions are likely to provide a lower quality score compared to the observers that consider the image content to be undistorted. This factor of variation in the CI and SE is strongly related to attentional mechanisms, as different viewers look at different things and thus have different likelihood of detecting distortions depending on the location of their appearance.

S3: **Confidence**: The difficulty that observers face when judging image quality depends on many factors including the distortion types, distortion distributions, distortion strengths, and the interaction of the distortions with the underlying image content. Given these factors, the difficulty of judging the quality of an image may vary strongly between images. The difficulty in quality judgement that observers experience is reflected in the confidence with which they provide a particular quality score and is considered here to be a source of variation in the CI and SE.

In this chapter, we are particularly interested in evaluating the confidence with which an observer rates the quality of an image and how the confidence is related to the corresponding quality ratings and the SE. The evaluation is based on the confidence scores (CS) and response times (RT) in relation to the quality scores (QS) which were all obtained from the subjective experiment E4b, as explained in Chapter 9. The RT has been widely used in psychophysics [234] and is considered here as an alternative, non-intrusive measure of confidence, as it may be inconvenient to require too much information from a participant during a subjective experiment.

The analysis in this chapter is undertaken considering three hypotheses, which are based on intuitive expectations regarding the relationship between the different quantities obtained in experiment E4b. To be more precise, we hypothesise that:

H1. It is easier to rate an image if its quality is either very good or very bad while images of medium quality are harder to judge. As a measure of difficulty when judging image quality we consider a CS given by a human observer.

H2. The confidence of a human observer when rating the quality of an image is strongly related to the RT of the quality rating. As such, we expect a longer RT for images that are harder to judge.

H3. Observer confidence can be predicted with reasonable accuracy based on the given QS in combination with the RT measured. Such a confidence prediction may be used as a measure of reliability of a particular MOS in addition to CI and SE.

The goals of this chapter are thus twofold. Firstly, we aim to establish relationships of CS and RT with QS and their corresponding SE, obtained from the subjective image quality experiment E4b (see Chapter 9). This serves to identify the degree to which the disagreement between observers is related to their confidence in relation to the quality judgements given and might provide valuable insight into the perception of the complex distortion patterns observed in wireless imaging applications. Secondly, we aim to model the prediction of mean CS based on QS, RT, and SE, as a non-intrusive measure of observer confidence to complement the usually computed CI and SE.

In relation to the three sources (S1, S2, S3) that we identified to have a major influence in the CI, the disagreement between observers can then be broken down into the impact of observer 'preference' and distortion 'detection' on one hand and observer 'confidence' on the other hand. Future studies could take this one step further by also assessing the individual impact of 'preference' and 'detection', for instance, by deploying appropriate surveys and rating scales in the experiment, respectively. However, this is outside the scope of this thesis.

In the following sections, we first analyse the different scores (QS, CS, RT) individually and then discuss a detailed analysis of their interrelation. Different models for mean CS prediction are then established based on the QS, RT, and SE. It should be noted that we focus in this chapter on the quality assessment outcomes of experiment E4b in terms of QS, CS, and RT, disregarding the eye tracking data. The related gaze patterns are analysed in Chapter 11.

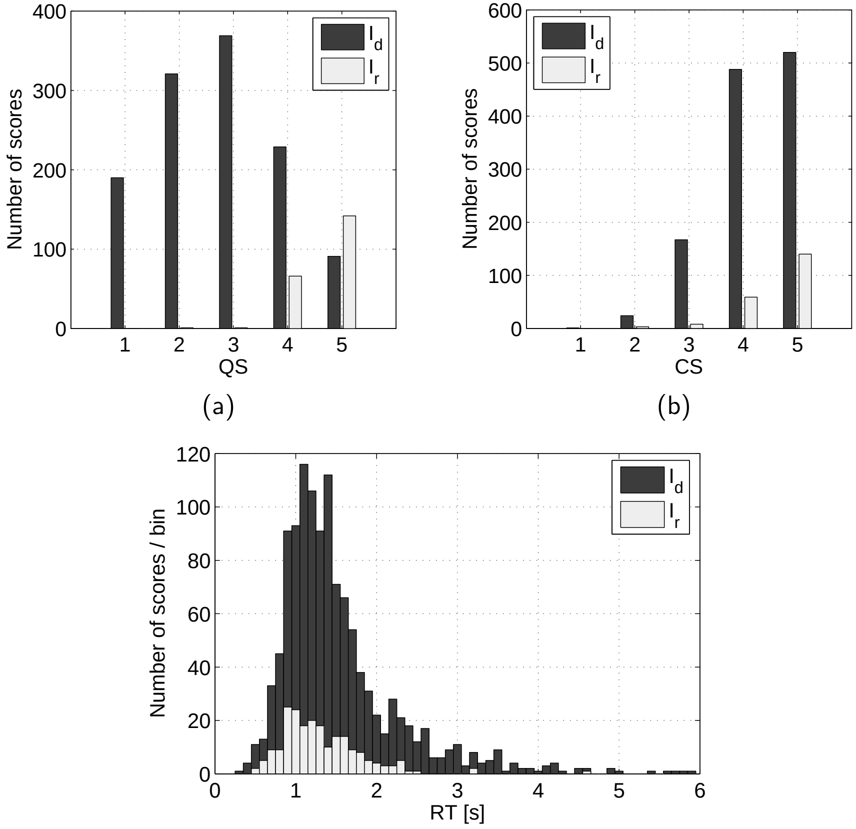
10.1 Analysis of quality scores, confidence scores, and response times

In this section, we analyse the QS, CS, and RT independently from each other. Given the 94 images that have been presented to the 15 viewers in experiment E4b, a total of 1410 values were available for each of the three quantities. Where applicable, we conducted the analysis independently for the image sets $\mathcal{I}_r$ and $\mathcal{I}_d$ to allow for comparison between the reference images and distorted images, respectively.

10.1.1 Distributions of subjective scores

The distributions of the QS, CS, and RT are presented in Fig. 47(a)-(c), respectively. The value of each quantity is given along the abscissa and the number of each value is shown on the ordinate.

Considering the two sets of reference images, $\mathcal{I}_r$, and distorted images, $\mathcal{I}_d$, in its entirety, one can see in Fig. 47(a) that the whole spectrum of QS values from 1 to 5 was covered by the participants, with a tendency for values towards the



(a)

(b)

(c)

Figure 47: Distribution of the quantities for the reference and distorted images obtained in experiment E4b: (a) QS, (b) CS, and (c) RT.

middle of the scale. This indicates that the test images covered a wide range of distortions relating to strongly varying perceived quality. Regarding the reference images only; these cover mainly the upper range of quality, as would be expected.

A different scoring behaviour can be observed in Fig. 47(b) for the CS values, where mainly the upper range of the scale has been covered. This applies both to the reference images and the distorted images. In fact, only 1 out of the total

Table 26: Statistical analysis of QS, CS, and RT for distorted images, $\mathcal{I}_d$, and reference images, $\mathcal{I}_r$.

Image type	Measured quantity	Statistical measure			
		μ	σ	β	γ
$\mathcal{I}_d$	QS	2.758	1.157	0.160	2.223
	CS	4.252	0.773	-0.770	3.052
	RT [s]	1.534	0.754	2.058	8.939
$\mathcal{I}_r$	QS	4.662	0.513	-1.318	5.024
	CS	4.600	0.636	-1.669	5.918
	RT [s]	1.318	0.511	1.907	10.724

1410 CS has been given as CS=1.

The continuous RT are presented as a histogram in Fig. 47(c). Here, each bin represents the number of RT that fall within a 100 ms range. It can be seen that the peaks of this distribution fall within 1-1.5 s for the distorted images. For the reference images the tallest bins lay around 0.8-1 s. Apart from a few exceptions, the RT for the reference images tend to go up to 2.5 s whereas the distribution of RT for the distorted images stretches beyond 4.5 s. This indicates that the participants could provide faster decisions for undistorted reference images that were considered to be of perfect quality.

10.1.2 Statistical analysis

A statistical analysis of QS, CS, and RT is provided in Table 26 to reveal further insight into the results obtained for these three quantities. In particular, the mean μ, the standard deviation σ, the skewness β, and the kurtosis γ (see Section 2.3.1) are presented for all three quantities and independently for the reference images, $\mathcal{I}_r$, and the distorted images, $\mathcal{I}_d$.

The statistics in the table confirm earlier observations that the CS are on average higher and the RT are on average lower for the reference images, as compared to the distorted images. The lower standard deviation of all three quantities on the reference images indicates that the observers were in higher agreement as compared to the distorted images. All distributions apart from QS on the distorted images are strongly skewed to either higher or smaller values, which is revealed by the positive and negative values of the parameter β. To be precise, the RT are asymmetrically spread towards higher values whereas the CS are asymmetrically spread towards lower values. Finally, the kurtosis γ shows that only the QS and CS for the distorted images experience a close to normal distribution, whereas all other quantities exhibit a leptokurtic distribution resulting in a more acute peak around the mean, as compared to a normal distribution. The RT experience especially high kurtosis values caused by the few very long RT as compared to the considerably lower mean RT.

10.1.3 Consistency over time

Assuming that the observers gain, to some degree, experience with quality evaluation with the number of images they have rated, one could expect that the CS and the corresponding RT may, respectively, increase and decrease during the progression of the subjective experiment. On the other hand, as we assured that a wide range of distortions was covered in both sessions and considering that we presented training images covering the range of distortions, it is desirable that the average QS should not vary too much between the sessions.

Given the above, we compare the averages of QS, CS, and RT between the first and the second session to see whether there are any changes in the rating behaviour. The averages over all observers and images of QS, CS, and RT are shown for both sessions in Fig. 48. It can be seen that, as anticipated, the QS does not change dramatically for both the reference images $\mathcal{I}_r$ and the distorted images $\mathcal{I}_d$. On the contrary to our assumptions, however, the CS and RT averages also do not change much between the two sessions. The RT decreases slightly in the second session but the CS remain almost the same. This indicates that the confidence of the observers remained stable throughout the experiment and that the training session was sufficient in developing a level of confidence already in the first session.

10.2 Interrelation between quality scores, confidence scores, and response times

In this section, we analyse the relationship between the QS, CS, and RT. For this purpose, we define the means over all participants for each of the images. In particular, the mean quality scores (MQS) are denoted as μ_{QS}, the mean confidence scores (MCS) are denoted as μ_{CS}, and the mean response times (MRT) are denoted as μ_{RT}. It should be noted that the acronym MQS is used here instead

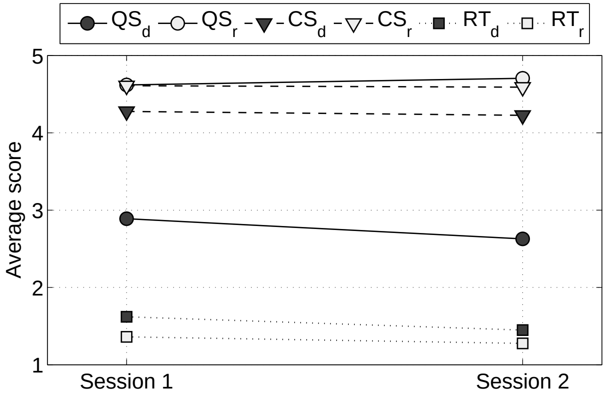


Figure 48: Comparison of average QS, CS, and RT between the two sessions.

of MOS to distinguish these scores from the MCS, which are also opinion-based scores and could thus be abbreviated with MOS.

10.2.1 QS-CS pairs

We hypothesised that it may be easier for a human observer to judge the quality of images at either end of the quality scale and that it may be harder to judge quality in the middle range of qualities (see H1). As such, one would expect high CS at either end of the quality scale. This hypothesis is related to the narrower CI at either end of the quality scale, as observed in the analysis of the MOS of experiments E1 and E2 (see Section 2.3.2 and in particular Fig. 8). However, in this earlier analysis it was not clear as to how much these narrower CI are actually related to the observers' confidence or are just a result of the limits of the quality scale.

More light is shed onto this issue by a combined analysis of the QS and CS obtained in subjective experiment E4b. The number of particular combinations of QS and CS as given by the participants are shown in Fig. 49. One can see that for QS at both the high end of the scale (QS $= 5$) and the low end of the scale (QS $= 1$), the confidence of the majority of human observers has been very high. This very high confidence drops towards the middle of the quality scale. However, one can see that the lower values of CS ($\leq$ 4) are predominant in the middle of the quality scale. These observations confirm hypothesis H1 and further indicate

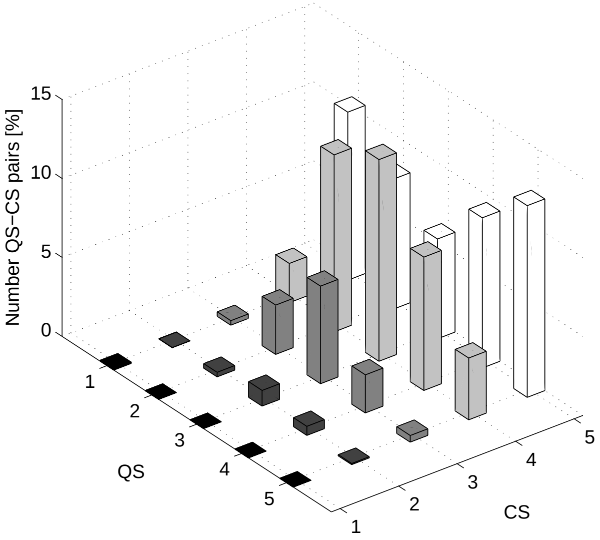


Figure 49: Number of occurrences of pairs of QS and CS.

that the narrower CI are not just due to the scale limits but may also be related to the observers confidence during quality assessment.

10.2.2 Average RT for QS and CS

We further hypothesised that RT may be longer for images that are harder to judge since the participant might require more time to make a decision (see H2). This may in turn be inversely related to the CS, meaning, a higher confidence should result in a quicker response. Thus RT may provide an indirect measure of observer confidence.

The average RT over all participants and all images are shown in Fig. 50 for both CS and QS. In alignment with our hypothesis one can see that the RT increases with the CS decreasing from 5 to 3. However, for lower CS the RT seems to drop, which is especially given for CS = 1. This seems at first contradictory, as one would expect a longer RT for a lower CS. It should be noted here though that there was only a few ratings CS=2 and in fact only one single rating CS=1. As such, it is hard to evaluate the validity of these values and further experimental data might be needed to achieve stronger statistical power of the results and

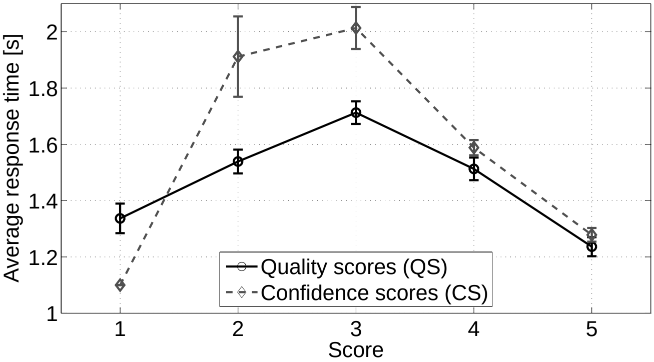


Figure 50: Average RT (with standard error of the mean) over all participants and images relating to particular QS and CS.

hence, be able to draw stronger conclusions. As it is, the rating CS=1 could possibly constitute an outlier.

From Fig. 50 one can also observe that the RT are increasing towards the middle of the quality scale, which is in alignment with the decreasing CS towards the middle of the quality scale (see Fig. 49). This suggests that RT is indeed related to CS and as such, reveals information about the confidence of an observer during image quality assessment.

10.2.3 Regression analysis

The relationship between the mean scores, MQS, MCS, and MRT, is further evaluated using regression analysis. We additionally relate these three scores to the SE of the MQS, $\sigma_{\mathcal{M}_n}$, which is computed as follows

$$\sigma_{\mathcal{M}_n} = \frac{\sigma_{\mathcal{M}}}{\sqrt{n}} \tag{83}$$

with $\sigma_{\mathcal{M}}$ being the standard deviation of the MQS and n being the number of observers over which the MQS was computed, here $n = 15$.

Similar to the discussions in Sections 2.3.2 and 3.5.6, we considered the classes of polynomial, exponential, power, and logistic functions to establish relationships between the different quantities. It turned out that first and second order polynomials most suitably represented these relationships. The resulting fitting curves

Table 27: Fitting functions with corresponding parameters and goodness of fit measures for MQS, MCS, MRT, and SE.

Variable	Fitting curve	Parameters	R^2	RMSE	SSE
MCS vs MQS	$p_2x^2+p_1x+p_0$	$p_2=0.199$ $p_1=-1.183$ $p_0=5.809$	0.681	0.170	2.641
MRT vs MQS	$p_2x^2+p_1x+p_0$	$p_2=-0.134$ $p_1=0.827$ $p_0=0.395$	0.512	0.159	2.299
SE vs MQS	$p_2x^2+p_1x+p_0$	$p_2=-0.026$ $p_1=0.155$ $p_0=-0.039$	0.539	0.030	0.083
MCS vs MRT	$p_2x^2+p_1x+p_0$	$p_2=0.832$ $p_1=-3.423$ $p_0=7.527$	0.516	0.201	4.005
SE vs MCS	p_1x+p_0	$p_1=-0.077$ $p_0=0.488$	0.269	0.038	0.132
SE vs MRT	p_1x+p_0	$p_1=0.086$ $p_0=0.030$	0.195	0.040	0.145

between all possible pairs of MQS, MCS, MRT, and SE are shown in Fig. 51(a)-(f). The corresponding prediction functions and goodness of fit measures are presented in Table 27.

It can be seen from Fig. 51(a) that the MCS are higher at either end of the MQS scale, confirming the observations on the individual CS scores in relation to QS (see Section 10.2.1). Both MRT and SE over MQS exhibit similarly shaped fitting curves with the lowest values at either end of the MQS scale, as can be seen in Fig. 51(b) and Fig. 51(c), respectively. It is interesting to point out that MCS, MRT, and SE are best fitted with MQS, using a quadratic polynomial function with a strong symmetry of the corresponding fitting curves around the center of the quality scale (MQS=3). The squared correlation coefficient R^2 indicates that MCS exhibits the strongest relation to MQS, followed by SE and MRT.

As for the other relationships; it can be seen from Fig. 51(d) and the cor-

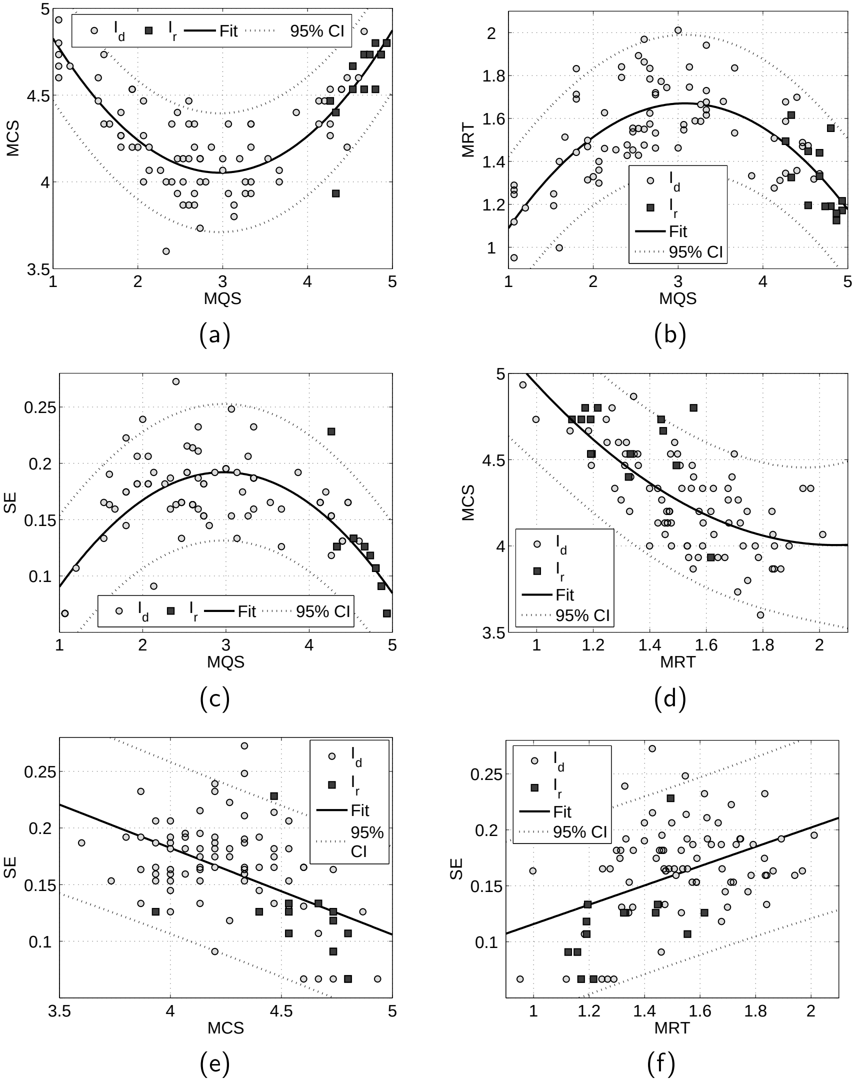


Figure 51: Fitting curves and 95% CI for distorted images $\mathcal{I}_d$ and reference images $\mathcal{I}_r$ for: (a) MCS over MQS, (b) MRT over MQS, (c) SE over MQS, (d) MCS over MRT, (e) SE over MCS, and (f) SE over MRT.

responding goodness of fit measures that MCS and MRT also exhibit a strong relationship, whereas the relation of both MCS and MRT to SE is rather weak. In fact, both MCS and MRT seem to be rather uncorrelated to SE, which is a strong indication that the other sources of variation in SE, as discussed earlier at the beginning of Chapter 10, have a strong impact on the magnitude of a particular SE of the MQS. It is also worth noting the symmetry of the linear fitting curves in Fig. 51(e) and Fig. 51(f), revealing the inversely oriented (though weak) relationships of MCS and MRT with SE.

10.2.4 Correlation analysis and bootstrap estimation of the standard error

The above findings show that there is a strong relationship between MQS, MCS, and MRT and also between MQS and SE. In fact, MCS, MRT, and SE are not directly related to MQS but rather to the distance of MQS to the middle of the quality scale $m_{QS} = 3$. To be precise, MCS increases with the distance of MQS from $m_{QS} = 3$ to the scale limits, whereas MRT and SE decrease towards the scale limits. For further analysis we therefore define a delta-MQS (DQS) measure as follows

$$\mu_{QS}^{\Delta} = |\mu_{QS} - m_{QS}|. \tag{84}$$

We then compute the Pearson linear correlation coefficient, ρ_P, and the Spearman rank order correlation coefficient, ρ_S, for all combinations of μ_{QS}^{Δ}, μ_{CS}, μ_{RT}, and ε_s to establish a full overview of the interdependencies.

All correlations are presented in Table 28 where negative values indicate that the corresponding quantities are inversely related to each other. The correlation coefficients give further evidence of the strong interdependencies between DQS, MCS, and MRT. This is particularly given for DQS and MCS with correlations above 0.8. These observations are true for both the Pearson correlation ρ_P and the Spearman correlation ρ_S. The SE on the other hand experience lower correlations with all three measures, but in particular with MCS and MRT.

We conduct a bootstrap analysis [235] to estimate the standard errors of the correlation coefficients presented in Table 28. In general, bootstrap methods are used for assessing uncertainty in parameter estimation and for empirical estimation of sampling distributions, when the form of the population of which the samples were drawn is unknown. Let $X = [x_1, x_2, \ldots, x_n]$ be a sample of size n, drawn from a sample space Ω, and let ϱ_b be a statistic of interest for which we want to estimate the standard error. The idea is then to treat X as if it represents the entire sample space Ω and use a Monte Carlo algorithm [236] to evaluate the statistic of interest in three steps:

Table 28: Pearson linear correlation coefficient ρ_P (above diagonal) and Spearman rank order correlation coefficient ρ_S (below diagonal).

	DQS	MCS	MRT	SE
DQS	1.000	0.825	-0.714	-0.620
MCS	0.809	1.000	-0.697	-0.519
MRT	-0.715	-0.698	1.000	0.441
SE	-0.521	-0.483	0.393	1.000

1. Create a bootstrapped sample $\hat{X} = [\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n]$ by drawing n values with replacement from X, meaning, that each value in $\hat{X}$ is drawn from the whole sample X as $\hat{x}_i \in [x_1, x_2, \ldots, x_n], i \in [1, n]$. Thus, some values x_i may be present in $\hat{X}$ several times whereas others may be entirely absent. This process is repeated B times, resulting in B bootstrapped samples $\hat{X}_b, b \in [1, B]$.

2. Evaluate the statistic of interest $\varrho_b(\hat{X}_b)$ for each bootstrapped sample $\hat{X}_b$.

3. Compute the standard deviation over all ϱ_b as a bootstrap estimate of the standard error as follows

$$\sigma_{BS} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\varrho_b(\hat{X}_b) - \mu_{\varrho_b})^2} \tag{85}$$

with μ_{ϱ_b} being the mean over $\varrho_b(\hat{X}_b)$ of all bootstrapped samples $\hat{X}_b$.

The accuracy of the bootstrap estimate improves with the number of bootstrapped samples B. In [237], the number of bootstrapped samples chosen was $B = 1000$, to obtain a reasonably accurate bootstrap estimate of the statistic of interest. However, as computational capacities have significantly improved over the past two decades, we chose $B = 100000$ to further improve the accuracy of our bootstrap estimate of the standard error.

The statistics of interest are the Pearson linear correlation coefficients, ρ_P, and the Spearman rank order correlation coefficient, ρ_S. For each of the six possible combinations of DQS, MCS, MRT, and SE we randomly create the $B = 100000$ bootstrapped samples on which both ρ_P and ρ_S are computed. The estimated

Table 29: Estimated standard errors for Pearson linear correlation coefficient ρ_P (above diagonal) and Spearman rank order correlation coefficient ρ_S (below diagonal).

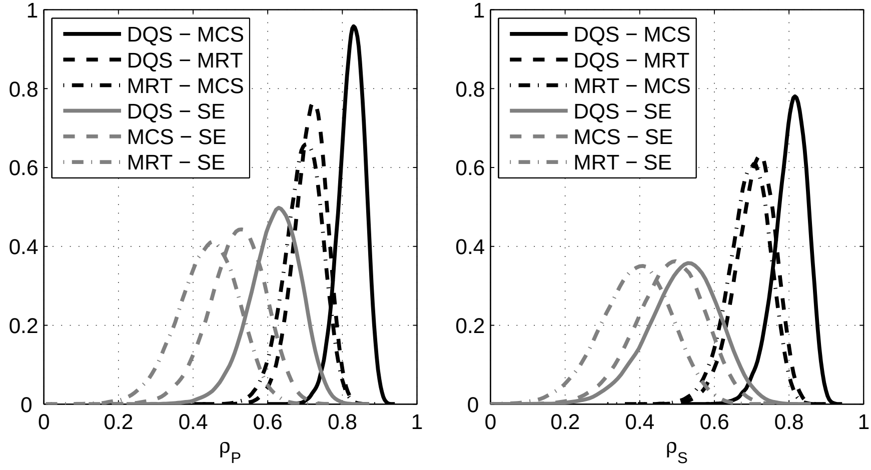
	DQS	MCS	MRT	SE
DQS	0	0.035	0.045	0.069
MCS	0.045	0	0.05	0.076
MRT	0.052	0.056	0	0.081
SE	0.094	0.092	0.095	0

standard errors of the correlations over all bootstrapped samples are presented in Table 29. The corresponding normalised density estimates [238] are shown in Fig. 52(a) and Fig. 52(b) for ρ_P and ρ_S, respectively.

One can see from Table 29, that the standard errors for the correlations between DQS-MCS, DQS-MRT, and MCS-MRT are considerably smaller, as compared to the standard errors for the correlations of DQS-SE, MCS-SE, and MRT-SE. This is also reflected in the narrower but taller density estimates in Fig. 52. In fact, the correlations between DQS-MCS, DQS-MRT, and MCS-MRT are almost exclusively far above 0.5, whereas the correlations between DQS-SE, MCS-SE, and MRT-SE exhibit a considerable amount of correlations well below 0.5. Given these standard error estimates, the high correlations between DQS-MCS, DQS-MRT, and MCS-MRT can indeed be considered to be of high certainty, unlike the correlations of the three measure DQS, MCS, and MRT with SE. This is further evidence that DQS, MCS, and MRT exhibit a stronger interrelation as compared to the relation of each of these quantities to the SE.

10.3 Observer confidence prediction

From the analysis in the previous sections it is apparent that MCS is strongly related to both DQS and MRT. Even though both DQS and MRT already provide a reasonable indication of an observers confidence when rating image quality, one may suspect that a combination of DQS and MRT could result in a further improvement of confidence prediction (see H3). Although SE exhibits only a minor interdependence with MCS, as compared to DQS and MRT, consideration of SE in a combinatorial prediction model may provide further improvement of confidence prediction. In this section, we thus aim on modelling the prediction of



(a)

(b)

Figure 52: Normalised distributions over 100000 random bootstrap samples for: (a) Pearson linear correlation coefficient ρ_P and (b) Spearman rank order correlation coefficient ρ_S.

observer confidence based on DQS, MRT, and SE.

10.3.1 Combinatorial prediction model

Similar to (81), we use a variant of the Minkowski metric to combine the different factors DQS, MRT, and SE into a confidence prediction model. In this respect, we look into two different classes of models. The first class, denoted as M_j, $j \in \{1, 2, 3\}$, incorporates the factors directly into the combinatorial prediction model, whereas with the second class, M_j^*, the factors are mapped to MCS before being incorporated into the combinatorial model. For this purpose, polynomial mapping functions are used, similar to the ones presented in Table 27.

For each of the two model classes, M_j and M_j^*, we consider three different combinations of DQS, MRT, and SE. The first combination is based only on DQS and MRT, keeping in mind that SE has a rather weak interrelation with MCS. Secondly, we define models based on only DQS and SE, given that these two quantities are readily available after an image quality experiment and no RT have to be recorded. Finally, all three factors DQS, MRT, and SE are incorporated into a model. The respective models of these three combinations are denoted as M_1, M_2, and M_3 for the first model class and as M_1^*, M_2^*, and M_3^* for the second model class.

Given the above, the resulting six confidence prediction models that are taken into account in this thesis are given as follows:

$$\begin{aligned}
\mathsf{M}_1: &\quad \mu_{CS}^{(M_1)}(\omega_1,p) &= \left[\omega_1\cdot(\mu_{QS}^{\Delta})^p + (1-\omega_1)\cdot\left(\frac{1}{\mu_{RT}}\right)^p\right]^{\frac{1}{p}} \\
\mathsf{M}_1^*: &\quad \mu_{CS}^{(M_1^*)}(\omega_1,p) &= \left[\omega_1\cdot(\mu_{CS}^{(QS)})^p + (1-\omega_1)\cdot(\mu_{CS}^{(RT)})^p\right]^{\frac{1}{p}} \\
\mathsf{M}_2: &\quad \mu_{CS}^{(M_2)}(\omega_1,p) &= \left[\omega_1\cdot(\mu_{QS}^{\Delta})^p + (1-\omega_1)\cdot\left(\frac{1}{\sigma_{\mathcal{M}_n}}\right)^p\right]^{\frac{1}{p}} \\
\mathsf{M}_2^*: &\quad \mu_{CS}^{(M_2^*)}(\omega_1,p) &= \left[\omega_1\cdot(\mu_{CS}^{(QS)})^p + (1-\omega_1)\cdot(\mu_{CS}^{(SE)})^p\right]^{\frac{1}{p}} \\
\mathsf{M}_3: &\quad \mu_{CS}^{(M_3)}(\omega_1,\omega_2,\omega_3,p) &= \left[\omega_1\cdot(\mu_{QS})^p + \omega_2\cdot\left(\frac{1}{\mu_{RT}}\right)^p + \omega_3\cdot\left(\frac{1}{\sigma_{\mathcal{M}_n}}\right)^p\right]^{\frac{1}{p}} \\
\mathsf{M}_3^*: &\quad \mu_{CS}^{(M_3^*)}(\omega_1,\omega_2,\omega_3,p) &= \left[\omega_1\cdot(\mu_{CS}^{(QS)})^p + \omega_2\cdot(\mu_{CS}^{(RT)})^p + \omega_3\cdot(\mu_{CS}^{(SE)})^p\right]^{\frac{1}{p}}
\end{aligned} \tag{86}$$

where $p \in \mathbb{Z}^+$ is the Minkowski parameter [209] and $\omega_i \in [0,1], i \in \{1,2,3\}$, are the relevance weights that control the contribution of the different factors to the overall prediction model. The optimal parameters p and ω_i are obtained by exhaustive search in the parameter space.

The functions $\mu_{CS}^{(QS)}$, $\mu_{CS}^{(RT)}$, and $\mu_{CS}^{(SE)}$ as part of the models M_1^*, M_2^*, and M_3^* are represented by the following polynomial mapping functions

$$\begin{aligned}
\mu_{CS}^{(QS)} &= p_1\cdot\mu_{QS}^{\Delta} + p_0 \\
\mu_{CS}^{(RT)} &= p_2\cdot\mu_{RT}^2 + p_1\cdot\mu_{RT} + p_0 \\
\mu_{CS}^{(SE)} &= p_1\cdot\sigma_{\mathcal{M}_n} + p_0
\end{aligned} \tag{87}$$

where the parameters p_0, p_1, and p_2 are determined through linear regression.

10.3.2 Model performance evaluation using cross-validation

We use k-fold cross-validation (CV) [239] to train and validate each of the six models presented in (86). Cross-validation is a resampling strategy used to validate the performance of prediction models by random sub-sampling of the available data. To be more precise, the original data is randomly subdivided into k sets. In each CV step, $k-1$ sets are then used for model training and the remaining set

Table 30: Confidence prediction performance indicators for all six models.

Model number		M_1	M_1^*	M_2	M_2^*	M_3	M_3^*
Training	ρ_P	0.845	0.844	0.825	0.825	0.84	0.844
	ρ_S	0.828	0.826	0.809	0.808	0.823	0.826
	$RMSE$	0.159	0.159	0.168	0.168	0.161	0.159
	r_0	0.489	0.461	0.544	0.57	0.474	0.461
Validation	ρ_P	0.863	0.868	0.862	0.862	0.868	0.867
	ρ_S	0.815	0.822	0.83	0.834	0.813	0.826
	$RMSE$	0.152	0.153	0.161	0.161	0.154	0.153
	r_0	0.464	0.461	0.544	0.558	0.528	0.461
Test	ρ_P	0.84	0.844	0.821	0.822	0.836	0.843
	ρ_S	0.858	0.859	0.838	0.838	0.84	0.859
	$RMSE$	0.177	0.175	0.186	0.186	0.179	0.176
	r_0	0.579	0.579	0.684	0.684	0.684	0.579

is used for model validation. This procedure is repeated k times with each of the k sets being used exactly once for the model validation. The results over the k folds are then averaged to determine a performance estimation of the prediction model.

In addition to the training and validation sets, we also leave out a number of images for a final performance test of the model, based on averaged model parameters determined from the CV. For this purpose, we split the set of 94 images used in experiment E4b into 10 approximately equal sized sets of 9 to 10 images. Eight of these sets are then randomly selected and used for an 8-fold CV, whereas the remaining two sets are combined into a test set. We then perform the 8-fold CV for the six models and analyse their performance on both training set and validation set. The performance of the prediction models in relation to the MCS is evaluated using the indicators described in Section 3.6.3. As a result, for each of the 8 folds, we obtain a set of model parameters and a set of confidence prediction performance indicators. These parameters and performance indicators are then averaged over all 8 folds. The averaged model parameters are then used to compute the final model which is evaluated on the test set.

The confidence prediction performance indicators are presented in Table 30

Table 31: Average model parameters over the 8 folds determined on the training sets in the cross-validation.

Model number	M_1	M_1^*	M_2	M_2^*	M_3	M_3^*
p	2.939	15.123	0.486	13.846	0.864	15.624
w_1	0.196	0.689	0.988	0.98	0.643	0.813
w_2	-	-	-	-	1	0.366
w_3	-	-	-	-	0.01	0.01
p_2	-	-	-	0.831	-	-
p_1	-	0.484	-	-3.432	-	-3.464
p_0	-	3.8	-	7.541	-	4.853

for all six models and for training, validation, and test sets. The corresponding model parameters are presented in Table 31 where p and ω_i relate to the models in (86) and p_0, p_1, and p_2 relate to the mapping functions in (87).

By comparing the prediction performance of the models between training, validation, and test sets, one can see that all models generalise very well. This applies, in particular, to the validation set where the performance indicators are very competitive with the corresponding performance indicators on the training set. The confidence prediction performance on the test set is very comparable for the correlation coefficients but slightly worse for the RMSE and the outlier ratio, r_0. However, generally one can conclude that all models are able to generalise well to unknown images.

Comparison of the confidence prediction performance indicators between the different models reveals that generally the models based on DQS and MRT (M_1, M_1^*) and the models based on DQS, MRT, and SE (M_3, M_3^*) are superior to the models based on DQS and SE (M_2, M_2^*). In fact, the correlation coefficients of the models M_2 and M_2^* on the test set are very similar to the correlations between DQS and MCS presented in Table 28, indicating that the SE does not contribute to an overall improved confidence prediction performance. This can also be comprehended by the negligibly small weight that SE receives in both models M_2 and M_2^*, as shown in Table 31. The models M_3 and M_3^* (incorporating SE) provide very similar performance to the models M_1 and M_1^* (disregarding SE), which is additional evidence of the negligible contribution of SE. On the other hand, the improved correlation coefficients of the models M_1 and M_1^* in comparison to the correlation between DQS and MCS (see Table 28) indicate a

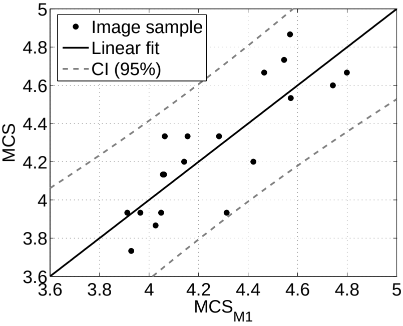


Figure 53: MCS versus predicted MCS, MCS$_{M1}$, using prediction model M1.

positive impact of the MRT being included into the confidence prediction model.

It can further be observed that the performance between the two model classes M$_j$ and M$_j^*$ is very similar, which applies to all three combinations of DQS, MRT, and SE. This suggests that both model classes are equally suitable to combine the different factors into an overall prediction model. Given that M$_j$ generally needs less parameters, and is thus less complex, as compared to M$_j^*$, one may in fact consider M$_j$ as the preferred class of models for the factor pooling.

Finally, with respect to the factor weights ω_i it should be noted, that they do not only reflect the importance that each factor carries with respect to the overall prediction model. The weights also inherently compensate for the magnitude differences between the different factors DQS, MRT, and SE as they are represented on different scales. This is particularly true for the model class M$_j$ where all factors are included directly and without mapping to a common scale. With model class M$_j^*$, on the other hand, all factors are mapped onto the MCS scale and thus, the weights relating to M$_j^*$ better reflect the relative importance of the different factors on the overall prediction model.

Given the above, we consider M$_1$ as the preferred model for observer confidence prediction. The reason being, its equal to superior prediction performance and lower complexity in comparison to the other models. A scatter plot of the MCS over the predicted MCS using model M$_1$, MCS$_{M1}$, is presented for the test set in Fig. 53, along with a linear fit and 95% CI.

11 Task-Free and Task-Based Visual Attention in Natural Images

In the previous chapter, we evaluated in detail the relationship between QS, CS, and RT from the experiment E4b. In this chapter, we analyse the gaze patterns obtained in both experiments E4a and E4b. The aim of this analysis is to gain a better understanding of human viewing behaviour when observing natural image content, both under task-free condition and under quality assessment task.

In this respect, we first discuss the post-processing of the GP into visual fixation patterns (VFP) and saliency maps (SM). The SM of experiment E4a represent the visual saliency of the image content, since the gaze patterns were recorded under task-free condition. The SM of experiment E4b, on the other hand, not only reflect the content saliency but also the viewing strategy deployed by the observers when judging the image quality.

The analysis of the SM from experiment E4a focuses on the variation of the viewing behaviour of the different observers. The motivation for this being that SM are widely used as a ground truth for VA modelling, similar to MOS which constitute a ground truth for quality models. However, little is known regarding the variation of gaze patterns between different observers that are used as a basis for the creation of the SM. To shed some light on this issue, we analyse in this chapter the SM of the individual observers in relation to each other and in relation to the image content. The discussion in this chapter is thus not restricted to the image communication context but is thought to be a more general contribution towards more reliable SM as a ground truth for VA modelling, which can subsequently be integrated into quality metric design.

The analysis of the gaze patterns obtained in experiment E4b serves to better understand human viewing behaviour when assessing quality of images containing complex distortion patterns. In particular, we evaluate the relative impact of the image content and the distortions on the SM of the human observers. It is revealed, that the selected ROI from experiment E3 (see Section 8.1) are strongly connected to the SM from experiment E4b. This is in agreement with the findings by Elazary et al. [240] and Masciocchi et al. [154], who showed that SM from VA models predict objects of interest well above chance. However, the SM from our experiment were obtained under quality assessment task when viewing images with distortions well in the suprathreshold regime. The outcomes reveal that the quality of images is mainly assessed in the ROI, even in the presence of strong distortions.

It should be noted, that the analysis and discussions provided in the following sections are by no means considered to be exhaustive. The results presented are instead intended to provide some insight into human viewing behaviour during both, task-free and task-based (quality assessment task) viewing of natural images. We further discuss some open issues that have in our opinion not been sufficiently addressed by the image quality research community when incorporating visual saliency into image quality models.

11.1 Processing of gaze patterns

The gaze patterns recorded in the subjective experiments E4a and E4b are post-processed into VFP and SM to obtain a more meaningful representation of the attentional behaviour of the observers when viewing the presented images. For this purpose, the gaze patterns are first converted into fixations by disregarding those GP that have been recorded during saccades and GP that have been recorded while none of the two eyes were tracked. The resulting VFP for each of the images then represents the locations and durations of the observers' focus of attention (FoA).

The human retina is highly space variant in processing and sampling of visual information. The accuracy is highest in the central point of focus, the fovea, and decreases strongly with increasing eccentricity to the fovea. Therefore, visual information is not only captured by the fovea but also by the photoreceptors surrounding it. To account for this gradually decreasing sampling accuracy, the fixations are further filtered using a Gaussian kernel resulting in a final SM. The processing of GP into VFP and SM is explained in detail in the following sections.

11.1.1 Creation of visual fixation patterns

A pseudo code for the creation of VFP from GP is provided in Algorithm 1 [241]. Here, the GP for a particular viewer and image are scanned in sequential order. The GP are assigned to clusters $\mathcal{C}_j$ according to a pre-defined threshold τ_{clus}. For this purpose, the mean $\mu(j)$ over all GP in the current cluster is computed, including the new $GP(i)$ at a particular time instance i. If the distance of $GP(i)$ to the mean $\mu(j)$ is below the threshold τ_{clus}, then $GP(i)$ is added to the current cluster $\mathcal{C}_j$. If the distance is above the threshold, the current cluster $\mathcal{C}_j$ is saved, the counter j is increased by one, and $GP(i)$ is added to the next cluster $\mathcal{C}_j$. After the clustering process, each cluster is considered to be a fixation $\mathcal{F}_n$ if it contains at least a pre-defined number, F_{min}, of GP. The value τ_{clus} was chosen with respect to the size of the presented image on the screen and the viewing

Algorithm 1 Pseudo code for creation of visual fixation patterns [241].

```
define cluster threshold $\tau_{clus}$
define minimum number of fixations $F_{min}$
set counter $j = 1$
create first cluster $\mathcal{C}_j$
for $i = 1$ to number of GP do
    compute mean $\mu(j)$ of $GP(i)$ plus all GP in $\mathcal{C}_j$
    compute Euclidean distance $\delta(i)$ of $GP(i)$ to mean $\mu(j)$
    if $\delta(i) < \tau_{clus}$ then
        enter $GP(i)$ into cluster $\mathcal{C}_j$
    else
        save current cluster $\mathcal{C}_j$
        increase counter $j = j + 1$
        create new cluster $\mathcal{C}_j$
        enter $GP(i)$ into $\mathcal{C}_j$
    end if
end for
save current cluster $\mathcal{C}_j$
for $k = 1$ to $j$ do
    compute number $N_{GP}$ of GP in $\mathcal{C}_k$
    if $N_{GP} \geq F_{min}$ then
        $\mathcal{F}_n = \mathcal{C}_k$
    end if
end for
```

distance as $\tau_{clus} = 20$. The value for F_{min} was chosen as $F_{min} = 4$, representing a commonly used lower threshold of about 100ms, above which the clustered GP are considered to be a fixation.

11.1.2 Creation of saliency maps

The pseudo code for the creation of SM from the VFP is given in Algorithm 2. Here, we first initialise the SM, I_{SM}, and enter the fixations by means of single-pixel peaks. The amplitude of the peaks is in correspondence with the fixation lengths which are in turn given by the number of GP that each fixation is based on. The SM is then convolved with a Gaussian filter kernel ϕ_G, as illustrated in Fig. 54, to obtain $I_{SM,\phi}$. We chose maximum filter dimensions of $x_{max} = y_{max} = 105$ pixels and a standard deviation of $\sigma = x_{max}/3 = 35$

Algorithm 2 Pseudo code for creation of saliency maps.

```
initialise the saliency map $I_{SM}$ with zeros
for $p = 1$ to number of participants do
  add fixations $\mathcal{F}(p)$ to $I_{SM}$
end for
create a Gaussian filter kernel $\phi_G$
convolve $I_{SM}$ with $\phi_G \rightarrow I_{SM,\phi}$
normalise $I_{SM,\phi}$ into the range [0...1] $\rightarrow \tilde{I}_{SM,\phi}$
multiply the image with $\tilde{I}_{SM,\phi}$ for visualisation
```

pixels. The part above the grey threshold constitutes the area of the filter kernel that covers the corresponding pixels in the image which are processed with high acuity by the fovea. This threshold assumes a size of the fovea of 2 dva and further depends on the screen resolution of 1280×1024 pixels and the viewing distance of 60 cm. The final SM is then created by normalising $I_{SM,\phi}$ into a range of $[0 \ldots 1]$ with higher values indicating more salient pixels. For visualisation of the saliency, the corresponding image is multiplied pixel-by-pixel with the SM, resulting in salient regions to receive more brightness compared to the remainder of the image.

11.2 Inter-observer saliency variation in task-free image viewing

As with any data recorded in experiments that involve human observers, a certain level of variability is expected between the gaze patterns, and thus the resulting SM, of the participants. The degree to which the SM vary between the participants depends strongly on the saliency of the image content. Hence, images that contain regions of strong saliency, and therefore attract much of the viewers' attention, may have more consistent SM between the viewers, whereas SM related to images without strong salient regions may experience larger discrepancies. Generally, however, SM are treated as being equally relevant and reliable as a ground truth for VA modelling, disregarding the inter-observer variability and the content of the image. In this respect, it is of interest to quantify the variability between the SM of the observers and thus, obtain some insight regarding the reliability of the associated SM for different image content.

The above discussion is in line with related procedures conducted in quality assessment where, for instance, the variability of the given quality scores between

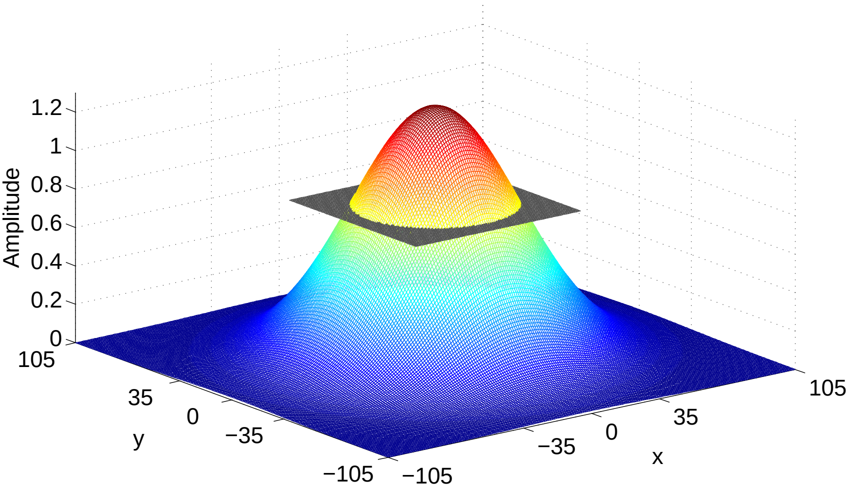


Figure 54: Gaussian filter kernel with $\sigma = 35$ pixels. The area above the grey threshold indicates the pixels that are captured by the fovea of the human eye.

observers is quantified using CI and the related SE. Subjective scores that are far away from the majority of the votes are often eliminated as outliers. Similar considerations could be taken into account with regards to gaze patterns obtained from eye tracking experiments to obtain SM that are more robust and reliable.

In this section, we therefore analyse the inter-observer saliency variations based on the gaze patterns and associated SM obtained from experiment E4a. We are particularly interested in analysing the variation of the saliency between observers and its dependance on the visual saliency of the content. This provides valuable insight into the reliability of SM created for particular image content. Ultimately, it may serve to define criteria for outlier detection and removal of observers to obtain more reliable SM to be used as a ground truth for the design of VA models. Through the analysis in this section we aim on providing evidence towards answering questions such as: How reliable is a particular set of saliency data? Are there big variations between observers? Do the variations depend on the image content? Can one identify outlier SM and exclude these observers?

To further advance VA modelling research in relation to image quality assessment, we made the gaze patterns from eye tracking experiment E4a publicly

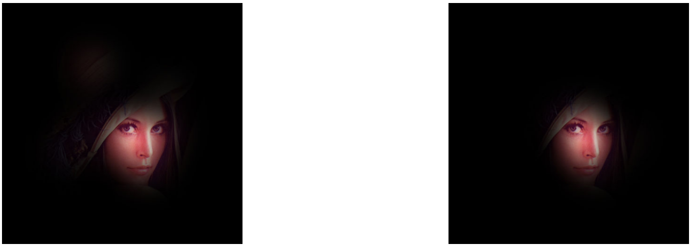


Figure 55: Example of highly correlated saliency maps with $\rho_P^{(i)}(O_m, O_n) = 0.968$ for image 'lenat'.

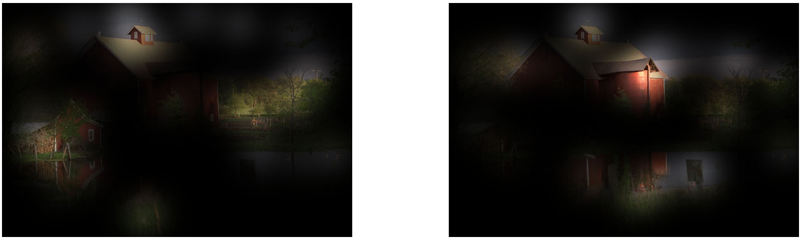


Figure 56: Example of nearly uncorrelated saliency maps with $\rho_P^{(i)}(O_m, O_n) = 0.000321$ for image 'house/kp22'.

available in the Visual Attention for Image Quality (VAIQ) database. Details about the VAIQ database can be found in Appendix C. Here, the reader can also find a visualisation of the SM overlayed on the natural image content. The original names from the quality databases are provided below each of the figures. The SM in Appendix C will also be referred to in the following analysis.

11.2.1 Cross-correlation analysis

As a basis for the following analysis, we created SM for each observer and image, resulting in a total of $15 \times 42 = 630$ SM. A pixel-based cross-correlation analysis [242] of the SM is then conducted to determine the variability between the SM with respect to the different observers and the different content, using the Pearson linear correlation coefficient as follows:

$$\rho_P^{(i)}(O_m,O_n)=\frac{\sum\limits_x\sum\limits_y\left(SM_{xy}^{(i,m)}-\overline{SM}_{xy}^{(i,m)}\right)\left(SM_{xy}^{(i,n)}-\overline{SM}_{xy}^{(i,n)}\right)}{\sqrt{\sum\limits_x\sum\limits_y\left(SM_{xy}^{(i,m)}-\overline{SM}_{xy}^{(i,m)}\right)^2}\sqrt{\sum\limits_x\sum\limits_y\left(SM_{xy}^{(i,n)}-\overline{SM}_{xy}^{(i,n)}\right)^2}}. \tag{88}$$

Here, $SM^{(i,m)}$ and $SM^{(i,n)}$ are, respectively, the SM of the observers O_m and O_n for the i^{th} image, $x \in [1,X]$ and $y \in [1,Y]$ are, respectively, the horizontal and vertical pixel coordinates, and $\overline{SM}^{(i,m)}$ and $\overline{SM}^{(i,n)}$ denote the respective mean pixel values. The correlation coefficient is computed in the range from -1 to 1, with a larger value corresponding to higher similarity between the SM. Examples of highly correlated and nearly uncorrelated SM of two observers are shown in Fig. 55 and Fig. 56, respectively.

11.2.2 Distribution of cross-correlations

We start by analysing the cross-correlations $\rho_P^{(i)}(O_m,O_n)$ between all $N_O = 15$ observers and for all $N_I = 42$ images to gain a general idea about the distribution of $\rho_P^{(i)}(O_m,O_n)$, disregarding particular viewers and image content. Given the number of observers, we computed 105 cross-correlations for each image, resulting in a total of 4410 cross-correlations over all 42 images. A histogram of the cross-correlations is shown in Fig. 57. It can be seen, that by large the majority of the cross-correlations are positive, indicating general agreement between the SM of the observers. In fact, the highest cross-correlation between two observers was computed as $\rho_P^{(i)}(O_m,O_n) = 0.968$ for the image 'lenat' (see Fig. 55). On the other hand, there is also a considerable amount of $\rho_P^{(i)}(O_m,O_n)$ around zero and in the negative range of the scale, indicating low agreement between the corresponding SM.

To formalise the distribution of all cross-correlations we conducted a curve fitting of the histogram considering different fitting functions. A mixture of two Gaussian distributions turned out to provide the best goodness of fit. The fitting curve is shown in Fig. 57 and the fitting function is given as

$$y(x) = 67.04 \cdot e^{-\left(\frac{x-90.77}{40.73}\right)^2} - 16.53 \cdot e^{-\left(\frac{x-138.5}{10.41}\right)^2} \tag{89}$$

with a squared correlation coefficient R^2 = 0.963 and a root mean squared error RMSE = 4.876. The maximum of the distribution is located at a correlation coefficient $\rho_P^{(i)}(O_m,O_n) = 0.51$.

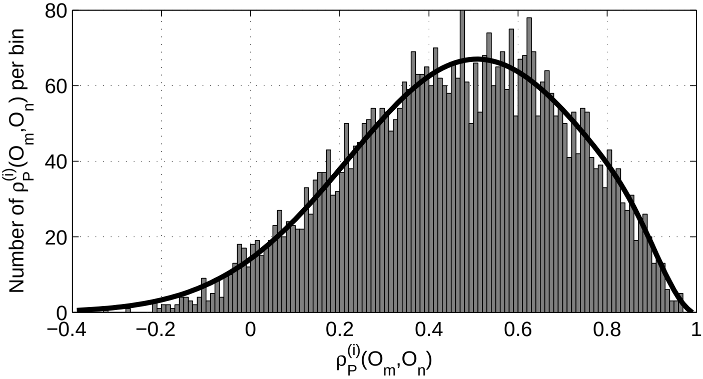


Figure 57: Histogram and fitting curve of cross-correlations $\rho_P^{(i)}(O_m, O_n)$ over all observers and all images.

11.2.3 Observer related analysis

In case of subjective quality assessment it is common practice to determine subjective quality scores that are far away from the majority of the votes, to classify them as outliers, and to exclude them from the final MOS. Analogously, one may also want to determine gaze patterns or SM that are significantly different from the majority of the SM and as such, represent viewing behaviour of an observer that is drastically different from the other observers. The difference can, of course, be due to many reasons that are not necessarily easy to determine. For instance, the observer may have not payed close attention to the experiment or he/she just had a different interest in the image content, compared to the majority of the observers.

We identify the agreement between all observers by averaging the cross-correlations between two observers, O_m and O_n, over all 42 images as

$$\overline{\rho_{P,O}}(O_m, O_n) = \frac{1}{N_I} \sum_{i=1}^{N_I} \rho_P^{(i)}(O_m, O_n). \tag{90}$$

The averaged correlations are presented in Fig. 58 with red indicating the highest correlation and dark blue relating to the lowest correlation. One can see that observers 1, 3, and 10 exhibit comparably lower correlations to the majority of the other observers, showing low agreement with the larger population of observers.

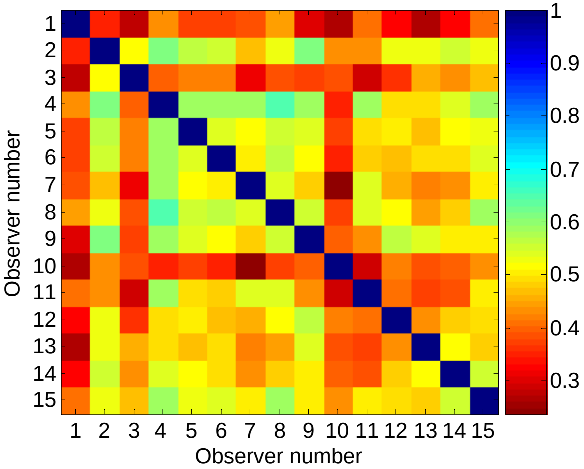


Figure 58: Mean cross-correlations, $\overline{\rho_{P,O}}(O_m, O_n)$, over all 42 images.

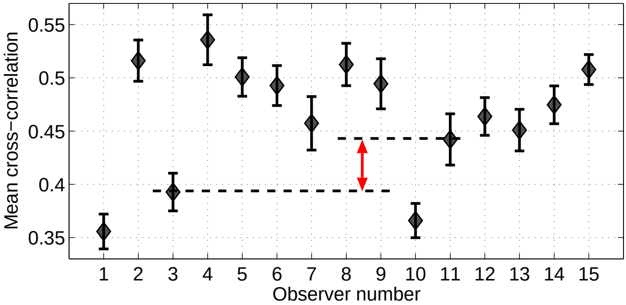


Figure 59: Mean cross-correlations over 42 images and 14 observers.

One could suspect that observers 1, 3, and 10 may correlate well with each other, however, the cross-correlations between these observers are also very low.

To further illustrate the low agreement of these three observers we additionally compute the marginal distribution of the correlations in Fig. 58 by averaging the cross-correlations for a particular observer over all other 14 observers. The result is shown in Fig. 59. Here it is even more apparent that observers 1, 3, and 10 have a significantly lower correlation with the other observers, indicated by the gap between observer 3 and observer 11 (highlighted by the red arrow). Therefore, one could regard observers 1, 3, and 10 as outliers and consider them for exclusion from the final SM creation. More specific criteria would of course be needed upon which to base such a decision about outlier SM, similar to the ones defined by VQEG for MOS [202].

11.2.4 Content related analysis

In this section, we identify the degree to which the content of the image impacts on the agreement between the SM of the different observers. For this purpose we average all cross-correlations for a particular image as

$$\overline{\rho_{P,I}}(O_m, O_n) = \frac{2}{N_O(N_O-1)} \sum_{m=1}^{N_O-1} \sum_{n=m+1}^{N_O} \rho_P^{(i)}(O_m, O_n). \tag{91}$$

The results are presented in Fig. 60 in order of decreasing $\overline{\rho_{P,I}}(O_m, O_n)$. It is interesting to note that there is a strong dependence of the cross-correlations on the content of the images. In fact, the average cross-correlations range from as high as 0.682 for the 'barba' image to as low as 0.225 for the 'bikes' image. There does not appear to be any strong drop of the correlations but rather a gradual decrease over all images, indicating that there is a large variety of content and saliency covered in the images used in the experiment. It is further interesting to point out that the standard error over all correlations does not appear to be very different between the image contents.

A few more findings are briefly highlighted in the following. Firstly, all images that exhibit faces (human and animal) and human beings are found in the upper half of the correlations, thus showing stronger agreement between the SM of the observers. This is rather expected as humans and their faces are well known to attract attention. One exception is the image 'bikes' which also contains humans but has the lowest average correlation in the considered set of images. This is thought to be due to two reasons. Firstly, the humans are wearing helmets, thus hiding the faces which would otherwise attract attention. Secondly, there are several humans present, thus representing multiple salient regions. The latter phenomenon is also apparent in the image 'rapids', which contains multiple humans but experiences only an average correlation. On the other hand, the image 'mandr' contains only one face but also experiences an average correlation. This is found to be due to the area of the image that the face covers; as it is very large, different people look at different parts of the face. Finally, a very high correlation

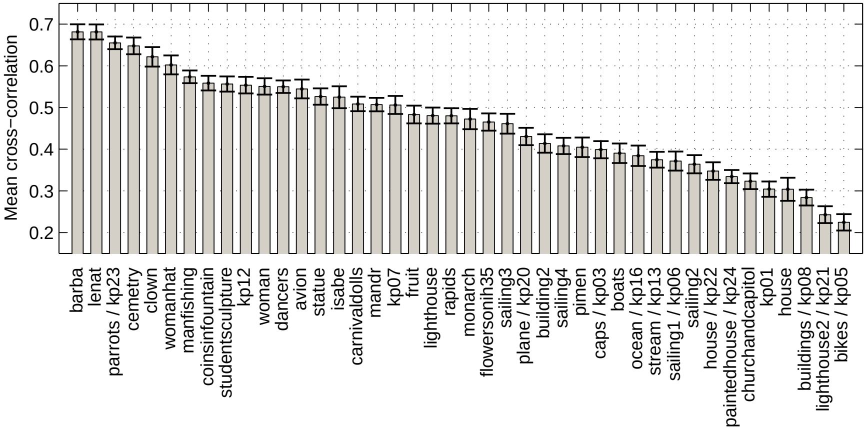


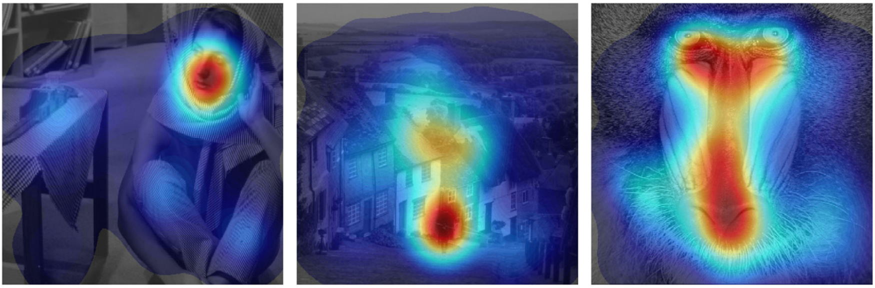
Figure 60: Mean cross-correlations, $\overline{\rho_{P,I}}(O_m, O_n)$, over all 15 observers.

was computed for the image 'cemetry'. This image does not contain any humans or faces, but instead contains a plaque with text on it, which drew most of the viewers' attention.

To summarise these observations, images that have regions or objects that are known to be salient, such as humans, faces, animals, and text, generally result in higher correlations of the SM. However, if there are multiple salient regions present in the image, or if the salient region is too small or too large, then the agreement between observers drops which can be observed in an decline of the cross-correlations.

11.3 Visual attention during image quality assessment

Unlike the SM created from experiment E4a (see Section 11.2), the SM from experiment E4b do not directly reflect the saliency of the visual content, as the gaze patterns were recorded under quality assessment task. Hence, the gaze patterns inherently reflect the search strategies that human observers deploy when analysing an image to provide a final judgement of the overall quality. The resulting SM are thus strongly based on top-down attentional viewing behaviour in addition to the bottom-up attention. This is further increased as the content of the images presented in experiment E4b has been repeated numerous times and as such, the observers became familiar with the visual scenes they were shown.



Barbara Goldhill Mandrill

Figure 61: Example images visualising the heat maps created from the gaze patterns of experiment E4b.

Given the above, we analyse in this section the human viewing behaviour when assessing image quality. We are particularly interested in the relative importance of image content and distortions on the gaze patterns, and the resulting SM. In this respect, we aim to gain insight into what type, strength, and distribution of distortions are attended to perform the quality assessment task. We are further interested in evaluating whether the region of analysis changes between different distorted images of the same content, and how much the changes are content and distortion dependent. We reveal that the ROI obtained from experiment E3 (see Section 8.1) are strongly connected to the SM obtained in experiment E4b. This result is largely independent of the distortion types, strengths, and distributions.

For illustration, some SM are presented in Fig. 61 for the reference images 'Barbara', 'Goldhill', and 'Mandrill'. The SM were visualised as heat maps (HM) and superimposed over the image content. A red area in the HM indicates highest saliency whereas a dark blue area indicates lowest saliency, with the intermediate saliency levels in between. The grey areas represent regions that have not been attended. The SM for all reference images and distorted images obtained in experiment E4b are presented in Appendix D. In particular, the HM for the reference images, $\mathcal{I}_r$, are shown in Fig. 93 and Fig. 94 whereas the HM for the distorted images, $\mathcal{I}_d$, are presented in Fig. 95-101.

11.3.1 Consistency of viewing behaviour between the two sessions

To determine the impact of the distortions on the viewing behaviour of the participants we need a reference SM to which the SM of the distorted images can

Table 32: Pearson linear correlation coefficient, ρ_P, between the saliency maps of the reference images from the first and second session.

	Barbara	Elaine	Goldhill	Lena	Mandrill	Peppers	Tiffany
ρ_P	0.973	0.978	0.946	0.952	0.914	0.912	0.966

be compared to. For this purpose we consult the SM computed on the seven reference images $\mathcal{I}_r$. As each reference image has been presented twice during the experiment, once in the first session and once in the second session, we can also evaluate the consistency of the observers' viewing behaviour when being presented the same image twice.

The HM for the reference images are shown in Fig. 93 and Fig. 94 of Appendix D, with the left column containing the SM created from the gaze patterns of the first session and the middle column presenting the SM created from the gaze patterns of the second session. The SM in the rightmost column have been created using the gaze patterns from both sessions. In addition, the Pearson linear correlation coefficients, ρ_P, between the SM from the first and second session are given in Table 32 for all reference images.

One can clearly see from Fig. 93 and Fig. 94, that for all seven contents of the reference images, the SM between the first and the second session are very similar. This observation is supported by the very high correlation coefficients, which are all well above 0.9. In fact, the images 'Barbara', 'Elaine', 'Lena', and 'Tiffany' even exhibit correlations above 0.95 between the SM. These images contain humans and their faces which have already been shown in Section 8.1 to be of high interest to the observers. It can be seen here that these images with more dominant ROI also exhibit a more consistent viewing behaviour of the observers, as compared to the images with less dominant ROI, such as 'Peppers', or with multiple ROI, such as 'Mandrill'.

Particularly worth noting is the high correlation of the SM for the image 'Barbara'. This image experienced a wider spread of the ROI selections, which was assumed to be due to the off-center location of Barbara's face and the other salient objects in the scene. However, the viewing behaviour during image quality assessment is very stable with the highest saliency in the face, followed by salient regions on the legs and the object on the table.

In relation to the image 'Goldhill' it is interesting to see the strong saliency on the man walking down the street. This strong focus is even more pronounced in the SM of the second session as compared to the first session, which may suggest that more people would have detected the man by then. In any case, this is a strong indication that people are searching for distortions in objects and regions that they are particularly interested in.

The lowest correlations are exhibited by the images 'Mandrill' and 'Peppers'. In the case of the former image, this is thought to be due to the face, and thus the ROI, covering the entire visual scene. Hence, there is no distinct fixation point as with the images that contain smaller ROI. Furthermore, 'Mandrill' contains several ROI in terms of the eyes, the mouth, and the large nose. On the other hand, the image 'Peppers' does not contain any strong ROI, which explains the more spread and inconsistent SM.

11.3.2 Visual attention to structural distortions

The viewing behaviour can be considered to be fairly consistent on the reference images with the highest saliency coinciding well with the ROI from experiment E3. As such, changes in the SM can be related to changes in the image content in terms of distortions. To evaluate the search strategies of the observers when assessing the distorted images, we thus analyse the HM of the distorted images, $\mathcal{I}_d$, which are presented in Fig. 95-101 of Appendix D. The images are sorted with respect to the decreasing MOS of experiments E1 and E2. In particular, images 1-40 represent the images used in experiment E1, sorted with decreasing MOS. Images 41-80 represent the images used in experiment E2, also sorted with decreasing MOS. The reason for keeping the order as in experiments E1 and E2 being, that the reader can relate the visual examination of the HM to the feature metrics presented in Section 3.1, and in particular Fig. 11.

We quantify the difference of the SM on the distorted images compared to the corresponding SM on the reference images with the correlation coefficient ρ_P. Here, we use the reference SM that were computed using the gaze patterns from both sessions (the right column in Fig. 93 and Fig. 94 of Appendix D). The correlations for all distorted images are presented in Fig. 62. One can see that for all seven reference image contents there is a wide spread of correlations. In general, however, there seems to be a tendency of the images with higher viewing consistency on the reference images (see Section 11.3.1) to also have a higher consistency with respect to the distorted images. To shed some more light on this, we computed the average, μ_ρ, and standard deviation, σ_ρ, over all correlations related to a particular image content, which are summarised in Table 33. Here one can clearly see that 'Elaine', 'Lena', and 'Tiffany' have a comparably higher correlation than 'Goldhill', 'Mandrill', and 'Peppers', which is

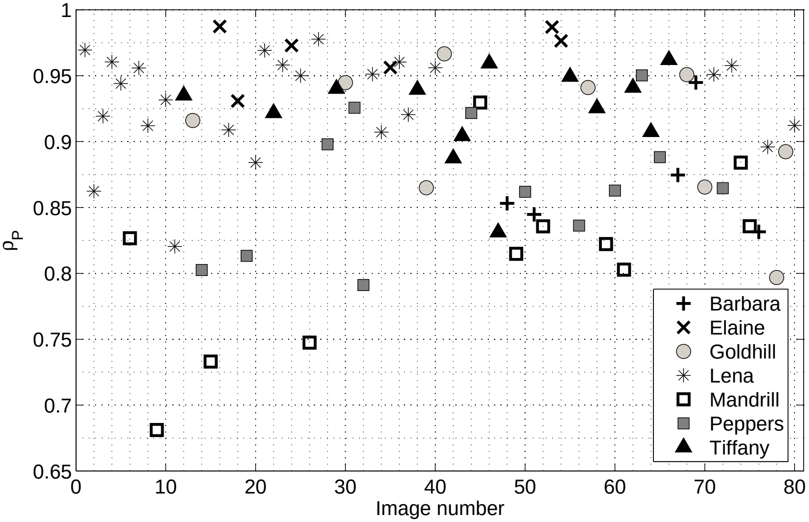


Figure 62: Pearson linear correlation coefficient, ρ_P, between the saliency maps of all distorted images and their corresponding reference images.

in alignment with the results presented on viewing consistency in Section 11.3.1. The only exception being here, once again, the image 'Barbara', which has shown high viewing consistency on the reference images but exhibits a lower correlation on the distorted images. This indicates that the distortions shifted the FoA to other regions of the image, in addition to the highly salient face, for the quality assessment to be performed.

We attempted to further determine quantitative relationships between the distortions contained in an image and the related SM computed from the gaze patterns of all observers. However, the relationship between these two factors seems highly complex and not as intuitive as one might expect. For instance, one could suspect, that strongly distorted images more significantly change the SM as compared to weakly distorted images, as the content of the underlying image is more severely altered. This would mean that the correlations presented in Fig. 62 would gradually decrease with the image numbers increasing from 1 to 40 and from 41 to 80, as this relates to stronger distortions and lower related MOS, which is obviously not the case.

Table 33: Average, μ_ρ, and standard deviation, σ_ρ, of the Pearson linear correlation coefficient, ρ_P, between the saliency maps over all distorted images and their corresponding reference images.

	Barbara	Elaine	Goldhill	Lena	Mandrill	Peppers	Tiffany
μ_ρ	0.87	0.969	0.904	0.931	0.81	0.868	0.923
σ_ρ	0.045	0.022	0.055	0.038	0.069	0.051	0.035

Given the above, we conducted a qualitative analysis of the HM presented in Fig. 95-101 of Appendix D in relation to the distortions in the images. The major observations can be summarised as follows:

- **Distortion location:** The observers generally tend to analyse the quality within the average ROI obtained from experiment E3. This is particularly true when distortions are present both inside and outside the ROI. This phenomenon can be observed in images 33 and 71 which exhibit strong artifacts at the bottom of the image but the observers still analyse the quality in Lena's face, as there are also subtle distortions present. If on the other hand distortions are absent from the ROI and only present in the BG, then the search range is shifted to other parts of the image outside the ROI (e. g. images 20 and 47).

- **Distortion distribution:** Global distortions generally do not alter the SM as much as local distortions do. For instance, the images 37, 38, and 40 are strongly distorted on a global level, but the impact on the SM is only small, as indicated by the high correlation coefficients. On the other hand, local distortions are more likely to change the gaze patterns, in particular if they are located outside the ROI (e. g. images 19, 20, and 47). These observations are in agreement with the findings reported in [243].

- **Distortion strength:** As far as we can observe, the distortion strength has surprisingly little impact on the alternation of the SM, as compared to the distortion location and distribution. Strong distortions that are locally distributed tend to change the SM (e. g. images 10, 17, 19, and 20) but so do weak locally distributed distortions (e. g. images 2, 9, 11, and 14). In fact, it seems that subtle distortions often change the SM more than strong distortions which can be attributed to the fact that they need to be attended comparably longer for thorough analysis. This is clearly present in image 9 which received the lowest of all correlations. Here, very subtle ringing artifacts are present on the side lobes of Mandrill's nose that are thoroughly investigated by the observers, as indicated by the HM.

- **Distortion type:** The different distortion types in relation to their strengths seem to have a considerable impact on the viewing behaviour. Blocking artifacts usually need to be fairly strong for the observers to attend them. For instance, the strong line of blocking artifacts in image 20 draws much attention whereas the weaker blocking artifacts in image 25 change the SM only marginally. On the other hand, ringing artifacts can considerably change the SM (e. g. images 2, 9 and 14) even if they are only very subtle. One reason for this may be related to ringing artifacts being considered to be more 'unusual' as compared to blocking artifacts that most observers would have been exposed to before. Also, the more complex structure of the ringing artifacts may lead to a more thorough analysis. Block intensity shifts are typically analysed at the border between the two different intensities (e. g. images 68, 70, 72, and 77) rather than in either of the intensity shifted areas.

To further analyse what distortions have caused the most severe changes in the SM we list, in the following, the images that have received the lowest correlations for each of the image contents and briefly summarise the artifacts contained in the images:

- **Barbara - Image 76:**

  extreme mix of blocking, ringing, noise, and block intensity shift artifacts

- **Elaine - Image 18:**

  ringing artifacts around Elaine's head and body

- **Goldhill - Image 78:**

  extreme blocking artifacts globally distributed

- **Lena - Image 11:**

  subtle ringing artifacts around the shoulder and in the feather boa

- **Mandrill - Image 9:**

  subtle ringing artifacts in the side lobes of Mandrill's nose

- **Peppers - Image 32:**

  blocking artifacts locally distributed in the lower half of the image

- **Tiffany - Image 47:**

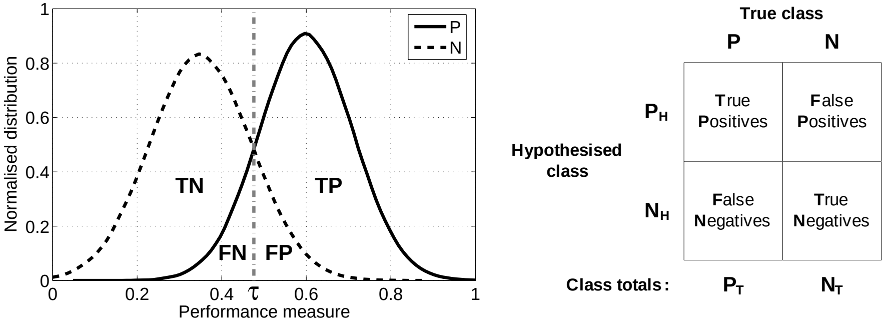
  subtle ringing artifacts in the hair and around the hand

In summary, the variety of different artifact locations, distributions, types, and strengths highlights again the complex interaction of the distortions in the images and the viewing behaviour of the participants when judging the image quality. Both, subtle artifacts (close to the near-threshold regime) and strong artifacts (well in the suprathreshold regime) can have an impact on the gaze patterns, with the subtle distortions appearing to dominate over the strong distortions. Generally, local distortions seem to change the gaze patterns more severe than global distortions, unless the global distortions are very strong. Finally, blocking artifacts, and especially ringing artifacts, seem to strongly attract attention and are major sources of changes in the gaze patterns.

11.3.3 Overview of Receiver Operating Characteristic analysis

The qualitative analysis from the previous section has indicated that human observers tend to analyse image quality in regions that are of high interest to them. To provide further evidence for this phenomenon, we evaluate the SM of the distorted images in relation to the ROI from experiment E3 by means of Receiver Operating Characteristic (ROC) analysis [244]. ROC analysis is normally used for binary classification of a performance measure into one of two classes. Here, we make an unconventional use of ROC analysis to quantify the level of saliency that is present in the ROI and the BG of the distorted images. In the following, we briefly discuss ROC analysis.

ROC analysis is typically used to make a binary classification into a predicted or hypothesised class, based on an instance of a measured quantity or performance indicator. For instance, in medical science ROC analysis is often used to make a decision whether a patient has a particular disease or not, based on results of medical tests that were performed. In this case, the actual presence of the disease represents a true positive class, whereas the absence of the disease represents a true negative class. The two classes cannot necessarily be perfectly separated, meaning, that based on a particular measure one cannot always perfectly decide whether to classify as positive or negative. This is illustrated in Fig. 63 where the two distributions of the true positive (P) and true negative (N) classes are



		True class	
		P	N
Hypothesised class	P_H	True Positives	False Positives
	N_H	False Negatives	True Negatives
Class totals:		P_T	N_T

Figure 63: Example distributions and the corresponding confusion matrix.

overlapping. Performance measures above a threshold τ are classified as belonging to the positive class, whereas performance measures below the threshold τ are classified to belong to the negative class. This classification is, however, not always correct as the positive class spreads below the threshold and the negative class spreads above the threshold. Thus, there is a certain percentage of misclassifications with four possible outcomes of the classification:

- **True positive (TP):** the instance is *positive* and it is classified as *positive*
- **True negative (TN):** the instance is *negative* and it is classified as *negative*
- **False positive (FP):** the instance is *negative* and it is classified as *positive*
- **False negative (FN):** the instance is *positive* and it is classified as *negative*

The respective regions are denoted under the distributions in Fig. 63 and the relative magnitudes of these four outcomes are typically listed in a confusion matrix, as shown on the right in Fig. 63. Here, P and N represent, respectively, the true positive and true negative classes with a total of P_T and N_T instances. Correspondingly, P_H and N_H represent the hypothesised (or predicted) classes based on the binary classification. A number of performance metrics can be computed from the confusion matrix of which, in the context of ROC analysis, the two most relevant ones are the true positive rate (TPR) and the false positive rate (FPR). The TPR is given as the ratio of correctly classified positive instances, TP, to the total number of positive instance, P_T, as follows:

$$\mathsf{TPR} = \frac{\mathsf{TP}}{\mathsf{P}_T}. \tag{92}$$

Similarly, the FPR is the ratio of negative instances that were falsely classified to be positive, FP, to the total number of positive instance, P_T, and is given by

$$\mathsf{FPR} = \frac{\mathsf{FP}}{\mathsf{P}_T}. \tag{93}$$

The relative magnitudes of TPR and FPR are regulated by the decision threshold τ, where larger τ result in lower TPR but also in lower FPR. On the other hand, a lower τ changes the classification outcome in favour of higher TPR but also results in more misclassifications in terms of higher FPR. In the context of ROC analysis, this trade-off between TPR and FPR is represented in the ROC curve, where TPR is plotted over FPR for all possible magnitudes of τ covering the extent of the positive and negative distributions.

To illustrate the interdependence of the class distributions and the related ROC curves, three different pairs of distributions and ROC curves are shown in Fig. 64. From Fig. 64(a) it can be observed, that the ROC curve rises steeply if the distributions are separated nicely. This relates to a strong increase of the TPR at only little cost of FPR when moving the decision threshold from large values to lower values. Figure 64(b) shows that the ROC curve lowers towards the diagonal if the distributions are overlapping more, resulting in a higher FPR for a given TPR. If the positive class is in fact centered around lower values than the negative class, as is the case in Fig. 64(c), then the corresponding ROC curve falls below the diagonal.

The diagonal in the ROC space represents a classifier that randomly guesses a class, as TPR and FPR are equal for all values along the diagonal. Thus, to perform a more informed classification it is desirable to move as far away as possible from the diagonal towards the upper left corner. In fact, the upper left corner itself represents perfect classification, as the related distributions are exclusively separated and thus, no FP and FN classifications occur.

The area under the ROC curve (AUC) is typically computed as a measure of classification performance. As such, a ROC curve that rises faster and runs above the diagonal usually exhibits a larger AUC, as compared to slow rising ROC curves that are closer to or even under the diagonal. The AUC for the examples in Fig. 64 are shown in the respective ROC spaces. The AUC for the diagonal is computed to be 0.5, relating to 50% of the entire ROC space.

11.3.4 Interrelation analysis of saliency maps and ROI using ROC

In the context of evaluating the SM in relation to the ROI, we use the ROC analysis to quantify the relative amount of pixels that are present in either the

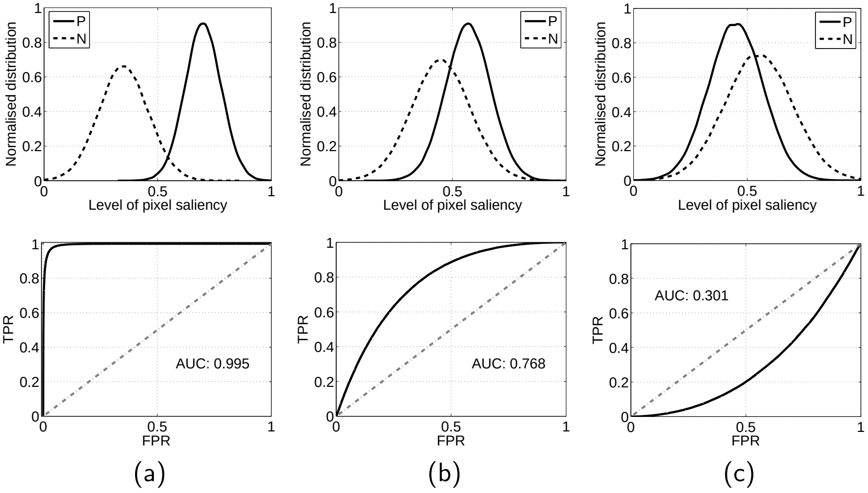
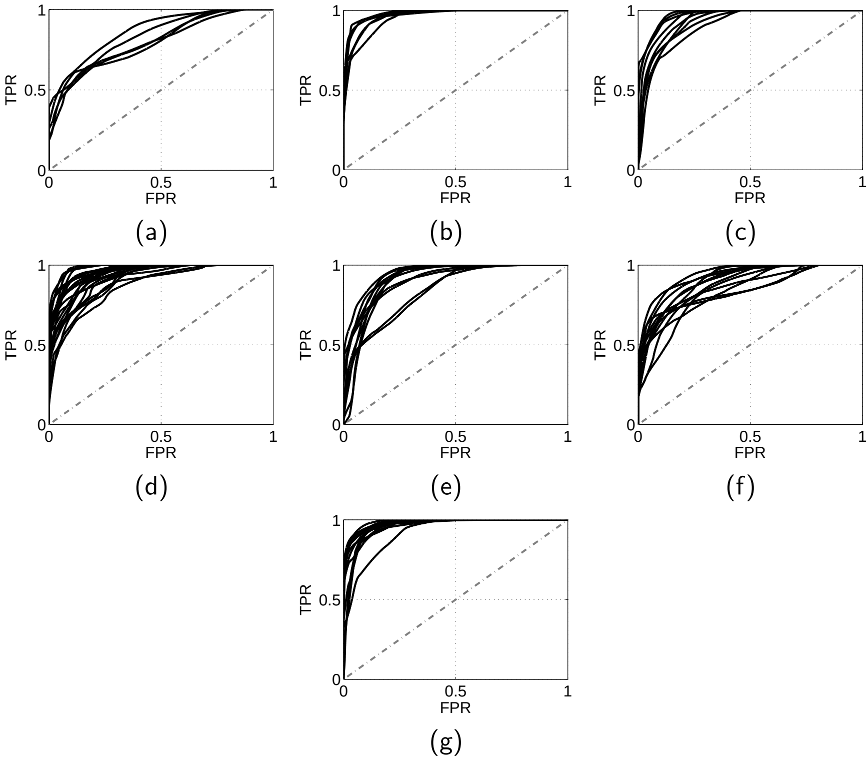


Figure 64: Illustrative example to demonstrate the effect of two distributions on the shape of the ROC curve and the magnitude of the corresponding area under the ROC curve (AUC).

ROI or the BG, given a particular saliency. Considering our earlier observations that the viewers analyse the quality mainly within the ROI, one would expect the presence of high-magnitude pixels of the SM to dominate in the ROI as compared to the BG.

In relation to the convention used in the previous section, we define the SM pixels located in the ROI to belong to the true positive class and the pixels in the BG to belong to the true negative class. The saliency level of the pixels is our performance measure. We hypothesise that the performance measure should be generally higher for the positive class (ROI) as compared to the negative class (BG), which relates to the cases presented in Fig. 64(a) and Fig. 64(b).

We create the ROC curve for each image by adapting the threshold τ with respect to the normalised magnitudes of the SM in the range from 0 to 1. By doing so, we obtain for each level of saliency (τ) the relative amount of pixels that are located in the ROI and the BG. If the pixels in the ROI would have exclusively higher saliency than the pixels in the BG, then the ROC curve would go through the top left corner of the ROC space. Even though this is not expected to be the case, given our earlier conjecture of higher saliency in the ROI one would expect



(a) (b) (c) (d) (e) (f) (g)

Figure 65: ROC curves between the SM and ROI for all distorted images of: (a) 'Barbara', (b) 'Elaine', (c) 'Goldhill', (d) 'Lena', (e) 'Mandrill', (f) 'Peppers', and (g) 'Tiffany'.

the ROC curve to be considerably higher than the diagonal of the ROC space. The related AUC would hence be expected to be well above 0.5.

The ROC curves for all distorted images are presented in Fig. 65(a)-(g), separated with regards to their respective seven reference images. For all image contents, the mean, μ_{AUC}, and standard deviation, σ_{AUC}, of the AUC computed over all respective ROC curves are presented in Table 34. It can be seen from both the ROC curves and the AUC statistics that indeed the highest saliency levels were located within the ROI. This is true for all distorted images without any exception, as all ROC curves are located clearly above the diagonal and the

Table 34: Mean, μ_{AUC}, and standard deviation, σ_{AUC}, of the area under the ROC curve (AUC) over all distorted images.

	Barbara	Elaine	Goldhill	Lena	Mandrill	Peppers	Tiffany
μ_{AUC}	0.818	0.97	0.945	0.946	0.907	0.884	0.97
σ_{AUC}	0.029	0.01	0.025	0.031	0.038	0.039	0.018

mean AUC are well beyond 0.5, in fact, even well beyond 0.8.

The highest mean AUC were computed for the images 'Elaine' and 'Tiffany', closely followed by 'Goldhill' and 'Lena'. The high AUC for 'Goldhill' is particularly interesting, as this complex scene has had a wide spread of ROI selections with only the man walking the streets somewhat standing out. However, as can be seen from the HM in Appendix D, the man has actually been attended a considerable amount by the observers to perform the quality assessment. Not unexpectedly, the 'Barbara' image has received the lowest mean AUC. This can be comprehended by consulting the HM, which show that the legs and the object on the table were frequently attended for quality assessment, both being outside the ROI.

11.3.5 Initial versus late viewing behaviour

Unlike with the selective ROI there is a temporal factor captured with the recorded gaze patterns, meaning, that information is available as to in which order the different regions were attended by the viewers. In this respect it is of interest to analyse whether the evidence provided in the previous section applies for both, SM that are based on early fixations and SM based on later fixations during the 8 s duration of image presentation. Through visual inspection of the fixations we found that there is indeed a distinct difference between the early and the later fixations of the viewers. In particular, it is apparent that the early fixations are mostly located within or around the ROI, whereas a considerable amount of later fixations is located outside the ROI. This suggests, that the viewers start their quality assessment task in the regions that are of interest to them and once evaluated, move on to other regions in the image.

This phenomenon is illustrated in Fig. 66, where the top row shows the first fixation and the bottom row shows the last fixation (combined from both sessions) of every observer for the three reference images 'Barbara', 'Goldhill', and 'Tiffany'. The size of the circle relates to the length of the fixation. It can clearly be seen that the last fixations tend to be more widely spread as compared to the first

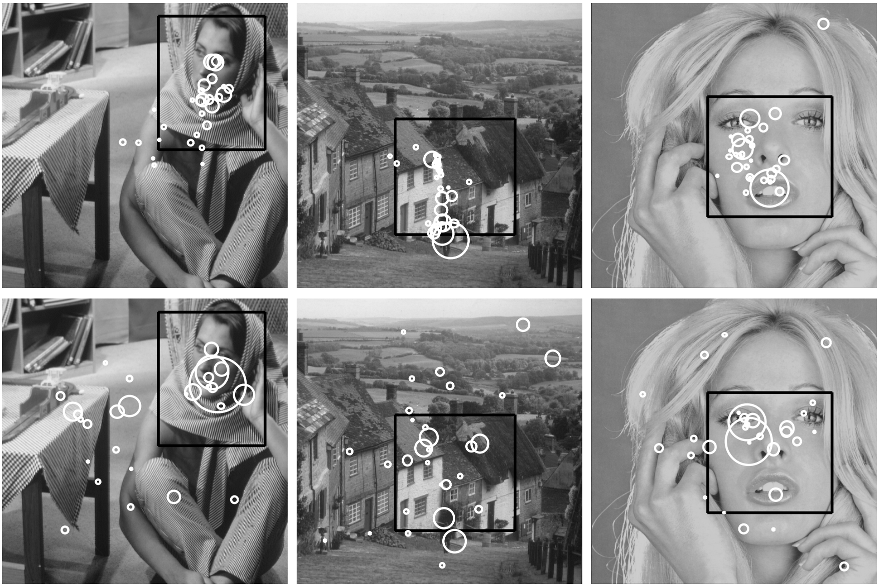
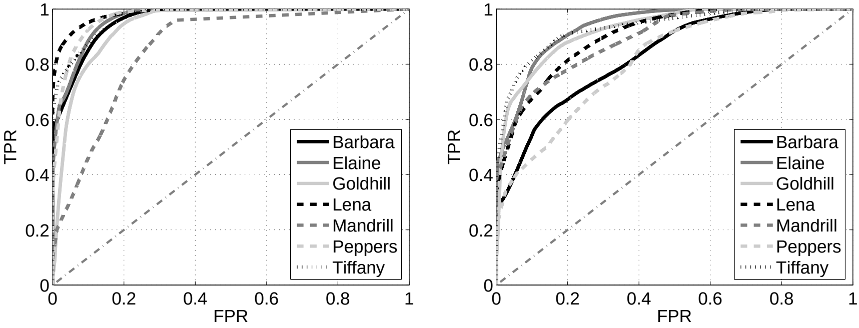


Figure 66: Comparison of the first fixation of every observer (top row) with the last fixation of every observer (bottom row) for the images 'Barbara', 'Goldhill', and 'Tiffany'. The black rectangle marks the mean ROI from experiment E3.

fixations, although still a considerable amount of fixations remains in the ROI. Similar observations have been made for other images.

To quantify these observations, we utilise again the ROC curves and the related AUC. The ROC curves corresponding to the three reference images in Fig. 66, along with the ROC curves of the other reference images, are shown in Fig. 67. The curves reveal that for the SM based on the first fixations, the saliency in the ROI is indeed comparably higher than in the SM based on the last fixations. This is true for all reference images but the 'Mandrill' image, for which the ROC curve of the SM based on the first fixation is considerably lower.

The AUC presented in Table 35 provide further evidence for these observations. Here, the AUC are presented as averages over the reference images, $\mathcal{I}_r$, and as averages over all distorted images, $\mathcal{I}_d$, related to a particular content. The AUC show that the saliency is generally higher in the ROI for the first fixations, F_1, as compared to the last fixations, F_L. However, the AUC also reveal that this



(a) (b)

Figure 67: ROC curves between the SM and ROI for all reference images with the SM being created from: (a) the first or (b) the last fixation of every observer.

Table 35: Area under the ROC curve for the ROI selections and the saliency maps created from the first fixation, F_1, and the last fixation, F_L, of each observer.

		Barbara	Elaine	Goldhill	Lena	Mandrill	Peppers	Tiffany
$\mathcal{I}_r$	F_1	0.962	0.967	0.94	0.985	0.858	0.97	0.971
	F_L	0.832	0.936	0.926	0.904	0.893	0.812	0.935
$\mathcal{I}_d$	F_1	0.943	0.98	0.945	0.974	0.871	0.915	0.962
	F_L	0.812	0.959	0.934	0.926	0.852	0.834	0.929

is not necessarily given for the 'Mandrill' image, as for the reference images the AUC is in fact higher for the SM based on the last fixation. This suggests, that the viewers did not turn to the BG for quality analysis during their later fixations. One possible reason could be, that the BG of the 'Mandrill' image contains the highly textured fur of the Mandrill, which makes quality assessment comparably harder to the more uniform regions around the nose.

Similar findings as the ones discussed in this section were reported in [155] where gaze patterns from a task-free eye tracking experiment were evaluated in relation to importance maps. It was concluded that the observers attended regions of high importance with their early fixations, whereas regions of lower importance were attended later during image viewing.

12 Eye Tracking and Video Impairment Assessment

The methods exploited in the previous chapters indicated that consideration of ROI and VA can be beneficial for quality assessment in the context of transmission errors. This is mainly motivated by the complex distortion patterns, and especially the localised distortions, caused by image transmission, as compared to the global distortions caused by source coding. The content saliency of the distortion region thus plays a more prominent role and a local weighting of artifacts becomes of great interest. However, the gain through VA and saliency models for image applications is expected to be limited as the content does not change and thus, suprathreshold artifacts would be detected eventually during image viewing.

In video applications, on the other hand, the content continuously changes and thus, the attention of the viewer is shifted in a much more dynamic way, as compared to images. Furthermore, distortions caused by transmission errors occur not only local in space but also local in time, meaning, the distortions may appear for a particular duration and disappear again. The notion of 'surprise' established by Itti et al. [245] states that humans gaze towards temporally novel events in a video stream in a highly significant manner. Thus, if the distortions are suprathreshold then the likelihood of detection is increased by the dynamic appearance and disappearance of the distortions. This is further determined, amongst other factors, by the relation of the saliency of the video content and the distortion duration. Distortions appearing in a region of higher saliency may be more likely detected, and thus perceived as annoying, as compared to distortions in a non-salient region. Similarly, longer durations may be more likely detected than shorter durations and may hence be perceived as more annoying.

To investigate the impact of content saliency and distortion duration on perceived annoyance of localised packet loss distortions, we conducted a combined video quality and eye tracking experiment, which is explained in detail in this chapter. The outcomes are analysed in Chapter 13 and provide valuable insight into the relation between loss duration and the content saliency of the distorted region and their impact on the overall perceived annoyance of packet loss distortions. The results are further utilised in a saliency awareness framework, which is discussed in Chapter 14, to improve quality prediction performance of existing video quality metrics. The creation of test sequences and the procedures of the experiment are outlined in the following sections.

12.1 Creation of distorted video sequences

The video sequences to be used in the experiment were selected with respect to two criteria; the saliency of the content and the spatial and temporal characteristics of the content. All video sequences were then encoded and transmission errors were introduced using a packet loss simulator. The details of the test sequences generation are explained in the following sections.

12.1.1 Identification of content saliency

The primary concern of this experiment was to identify the impact of content saliency on the perceived annoyance of packet loss distortions. Hence, a reliable ground truth was to be used to identify the saliency in the reference sequences. For this reason, we utilised gaze patterns from a previously conducted eye tracking experiment [157] in which 30 video sequences in standard definition (SD) format were presented to 37 participants. The gaze patterns were recorded using a dual-Purkinje eye tracker from Cambridge Research Systems [246]. In this eye tracking experiment, the sequences were presented under task-free condition and as such, the recorded gaze patterns represent the saliency of the visual content.

The gaze patterns were post-processed to eliminate saccades, leaving the fixations and smooth pursuit eye movements that contribute to VA. A Gaussian filter was then deployed to create the final SM for all frames of each sequence based on the visual fixations of all 37 observers. We visually inspected these SM to identify frames that contain regions of particular high saliency.

12.1.2 Source encoding and creation of loss patterns

The video sequences were encoded in H.264/AVC format [191, 214] using the JM 16.1 reference software [247]. As we are interested in evaluating the perceptual impact of transmission errors rather than source coding distortions, we encoded the sequences in high quality with a fixed quantisation parameter of QP=28. The fixed QP further minimises quality differences between the various sequences unlike, for instance, a constant bit rate would do. The sequences were encoded in High profile with an IBBPBBP... GOP structure of two different lengths; 30 frames (GOP30) and 10 frames (GOP10). The frame rate was set to 25 fps and as a result, the two GOP lengths correspond to 1.2 s and 0.4 s, respectively.

We utilised an adapted version of the Joint Video Team (JVT) loss simulator [248] to introduce packet loss into the H.264/AVC bit stream. An overview of the packet loss insertion is illustrated in Fig. 68 for both the GOP30 and GOP10 coded video sequences. The packet loss was introduced into a single I frame in

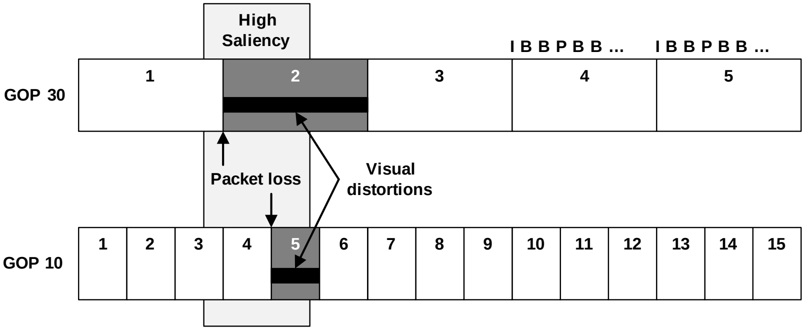


Figure 68: Packet loss insertion into I frames containing highly salient regions for GOP30 and GOP10 encoded video sequences.

each sequence, resulting in error propagation until the next I frame, due to the inter-frame prediction of the P and B frames. Thus, the two different GOP lengths (30/10 frames) relate to the maximum lengths of error propagation (1.2/0.4 s). To have better control regarding the location and extent of the corresponding spatial loss patterns we chose a fixed number of 45 macro blocks (MB) per slice during source encoding. Given that SD video has dimensions 720×576 pixels, corresponding to 45×36 MB, each slice represents exactly one row of MB.

To identify the impact of saliency on the perception of the distortions, we introduced packet loss into the sequences such that the corresponding visual distortions appear either in a salient region or in a non-salient region, based on the salient frames as identified in Section 12.1.1. In particular, we created test sequences with packet loss introduced in 5 slices, spatially centered around the most salient region in an I frame. We then created a corresponding sequence with 5 slices of distortions introduced into a non-salient region of the same I frame. The extent of the loss pattern was intentionally kept constant to allow for a better comparison between distortions in the salient region and the non-salient region.

We created such two sequences for both the GOP30 and GOP10 coded videos, resulting in a total of four distorted sequences for each reference sequence SEQ$_R$. The subsets of distorted sequences will in the following be referred to as SEQ$_{S,0.4}$, SEQ$_{N,0.4}$, SEQ$_{S,1.2}$, and SEQ$_{N,1.2}$, where 0.4 relates to error propagation length for GOP10 and, accordingly, 1.2 relates to GOP30. The indices S and N refer to the distortions being inserted either into the salient or the non-salient regions, respectively.

All sequences were shortened to 150 frames, corresponding to 6 s duration.

During the creation of the test sequences it was assured that no distorted frames were present in the first second and the last second of the video and also not immediately before or after scene cuts.

12.1.3 Spatial and temporal content classification

In order to keep the experiment at a reasonable length and thus, to lower the strain on the participants, we selected 20 out of the 30 reference sequences to be used in the experiment along with the corresponding distorted sequences. To cover a wide range of different visual content, both with respect to its spatial and temporal characteristics, we further classified the reference sequences using spatial information (SI) and temporal information (TI) indicators [20]. The SI indicator measures the spatial information over a number of N frames F_n. For this purpose, each frame is filtered using the Sobel operator $Sobel(\cdot)$ and the standard deviation σ is computed for the filtered frames. The SI indicator is then the maximum standard deviation over the N frames and is given as

$$SI = \max_N(\sigma(Sobel(F_n))). \tag{94}$$

The TI indicator is based on the motion changes in the video which is measured as the luminance pixel difference between two consecutive frames as

$$M_n(i,j) = F_n(i,j) - F_{n-1}(i,j) \tag{95}$$

where i and j denote the row and column number in the n^{th} frame, respectively. The standard deviation σ is then computed for each difference frame. Similar to the SI indicator, the TI indicator then is the maximum standard deviation over N frames as

$$TI = \max_N(\sigma(M_n(i,j))). \tag{96}$$

As both SI and TI indicators may change significantly throughout the duration of a sequence, we used for the content classification only the 30 distorted frames of the GOP30 coded video sequences. The SI and TI for all sequences are shown in Fig. 69. The selection of the test videos was done with respect to covering a wide range of SI and TI indicators. In Fig. 69, the numbered dots represent the 20 sequences chosen for the experiment whereas the remaining 10 sequences were not included in the test set.

Example frames for each of the 20 sequences SEQ$^{(i)}$, $i \in \{1, 2, \ldots, 20\}$, used in the experiment are shown in Fig. 70 and Fig. 71. To be precise, the I frame is presented in which the packet loss was introduced to create the sequences

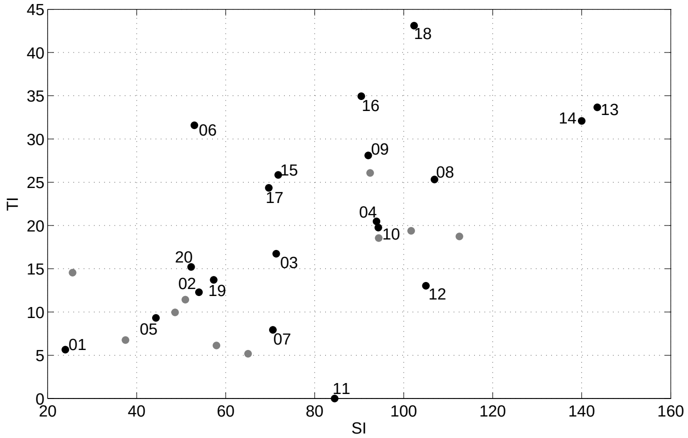
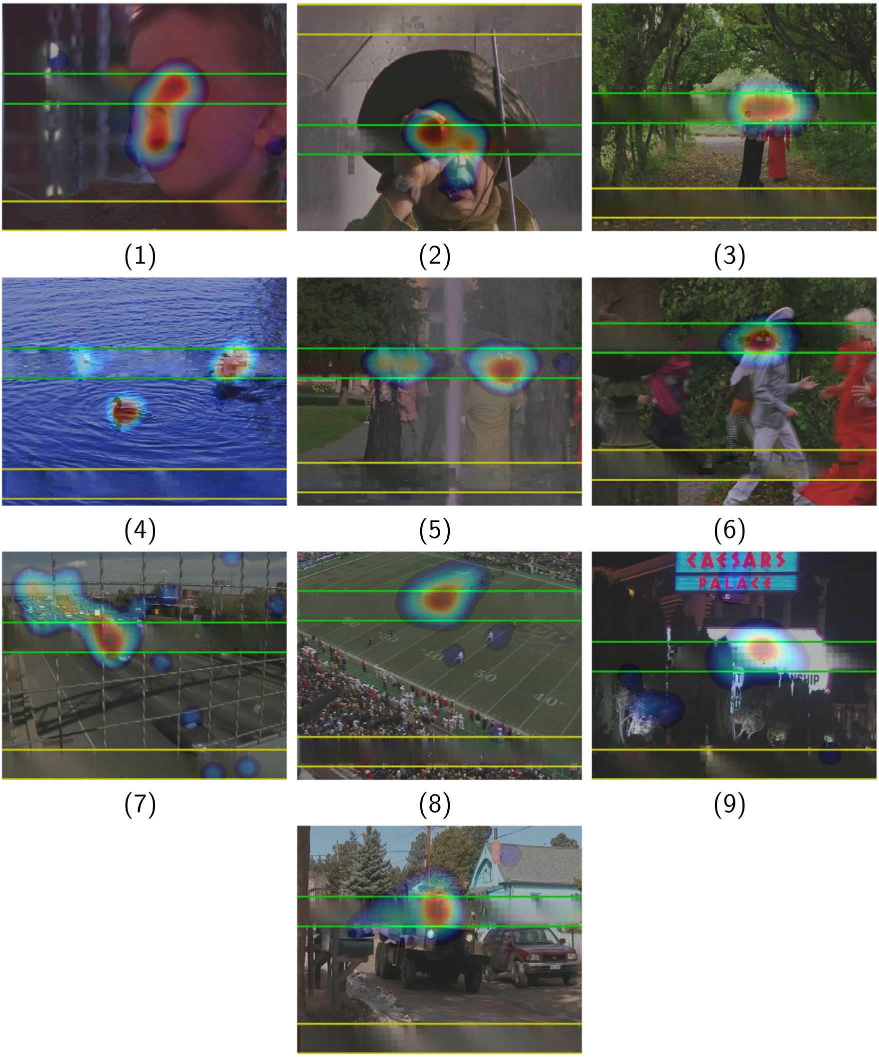


Figure 69: Spatial information (SI) and temporal information (TI) indicators [20] for all 30 sequences. The numbered dots represent the 20 sequences that have been selected for the experiment.

SEQ$_{S,0.4}$ and SEQ$_{N,0.4}$. For visualisation purposes in this thesis, the distortions of both the salient region and the non-salient region are presented within the same frame. The salient distortion region is additionally highlighted with green lines and the non-salient distortion region is highlighted with yellow lines. The saliency information from the task-free eye tracking experiment [157] on the reference images is additionally visualised using HM.

12.2 Details of experiment E5

The combined video quality and eye tracking experiment was conducted at the Image and Video Communication (IVC) department at the University of Nantes, Nantes, France, and is in the following referred to as experiment E5. The experiment procedures were designed according to ITU Rec. BT.500 [19] and will be discussed in the following sections.



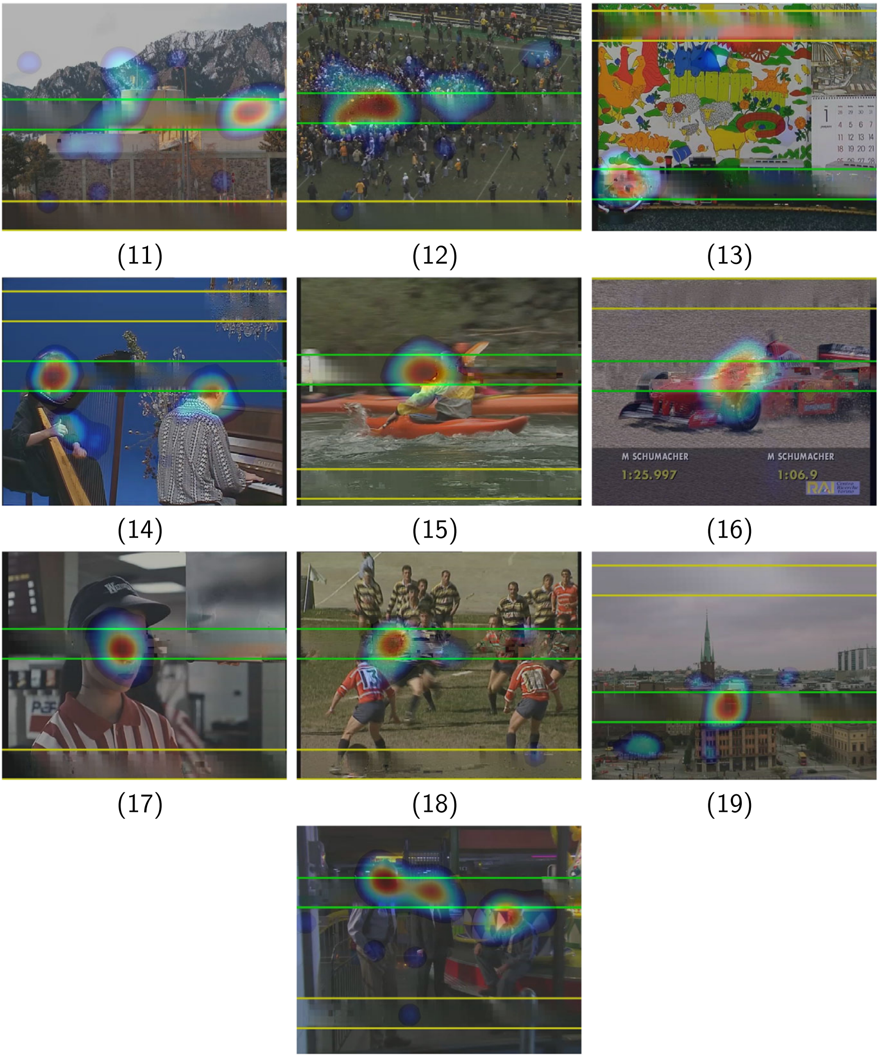
(1) (2) (3)

(4) (5) (6)

(7) (8) (9)

(10)

Figure 70: Example frames for video sequences 1-10, visualising both the salient (green lines) and non-salient (yellow lines) distortion regions and also the saliency information using heat maps.



(11) (12) (13)

(14) (15) (16)

(17) (18) (19)

(20)

Figure 71: Example frames for video sequences 11-20, visualising both the salient (green lines) and non-salient (yellow lines) distortion regions and also the saliency information using heat maps.

12.2.1 Laboratory setup

The laboratory in which the experiment took place was set up with grey covers on all walls and was illuminated with low light levels. The video sequences were presented on an LVM-401W full high definition (HD) screen by TVlogic with a size of 40" and a native resolution of 1920×1080 pixels. A mid-grey background was added to the SD test sequences to be displayed on the HD screen. The Video Clarity ClearView [249] video server was used for real-time playback of the sequences. The observers were seated at a distance of about 150 cm, corresponding to six times the height of the displayed video sequences.

12.2.2 Eye tracking hardware

The iView X$^{\text{TM}}$ Hi-Speed eye tracker by SensoMotoric Instruments (SMI) [250] was used to record the gaze patterns of the human observers during the experiment. The iView X$^{\text{TM}}$ Hi-Speed consists of a sturdy tower with a chin rest and a head rest. The gaze is recorded using an infrared camera and the pupils are illuminated using two infrared light sources. The recording rate of the iView X$^{\text{TM}}$ Hi-Speed is 500 GP/s. The gaze tracking accuracy is given by the manufacturer to be in the range of 0.25-0.5 dva. A photo of the laboratory setup with the eye tracker in the front is shown in Fig. 72.

12.2.3 Viewer panel

A total of 30 non-expert viewers participated in the experiment out of which 10 were female and 20 were male. The participants were mainly students and staff of the University of Nantes with the ages ranging between 15 to 39 years and an average age of about 23 years. Prior to each experiment, the visual accuracy of the participants was tested using a Snellen chart and any colour deficiencies were ruled out using the Ishihara test.

12.2.4 Experiment procedures

The participants were presented the 100 test sequences (20 reference sequences SEQ$_R$ plus 80 distorted sequences SEQ$_{S,0.4}$, SEQ$_{N,0.4}$, SEQ$_{S,1.2}$, and SEQ$_{N,1.2}$) in a pseudo random order, with a distance between the same content of at least 5 presentations. The sequences were presented using a single stimulus method, meaning, that the distorted sequences were presented without their corresponding reference sequence. The reference sequences were randomly mixed with the set of

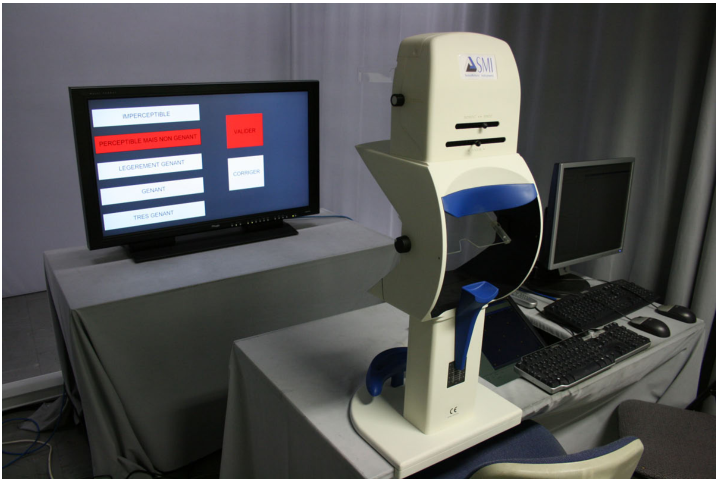


Figure 72: Laboratory setup with the SMI eye tracker in the foreground.

distorted sequences and the participants were not told if the currently presented sequence contained distortions or not.

Before the test sequences the participants were shown 6 training sequences in a fixed order for them to adapt to the impairment rating system and to get a feeling for the distortions that can be expected in the test sequences. For this purpose, training sequences were selected from the remaining 10 sequences (see Sec. 12.1.3 and Fig. 69) that covered the range of distortions in the test sequences.

The 5-grade impairment scale [19] was used to assess the annoyance of the distortions in the sequences. Here, the observers were asked to assign one of the following adjectival ratings to each of the sequences: 'Imperceptible (5)', 'Perceptible, but not annoying (4)', 'Slightly annoying (3)', 'Annoying (2)', and 'Very annoying (1)'. In this experiment, we chose the impairment scale over the quality scale, which is also defined in [19], as it has the advantage that the rating 'Imperceptible' directly allows to identify whether participants actually detected the distortions in the sequences or not. This knowledge is of interest for the subsequent analysis of the experiment outcomes, but can also be useful to improve the prediction performance of packet loss visibility models [251] or of video quality metrics [252].

The rating of a sequence was conducted after the presentation of the sequence

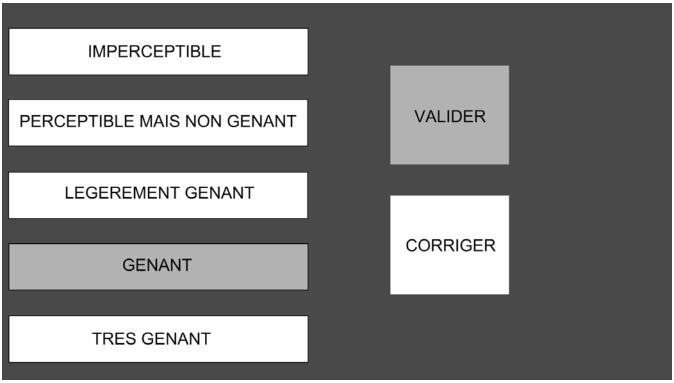


Figure 73: Five-grade impairment scale as utilised in experiment E5 [19].

was finished. The impairment scale, as shown in Fig. 73 (with the corresponding French labels), was displayed over the whole area of the screen with all scoring fields initially being white. The participants then had to look for at least one second at the score they wanted to assign to the previously presented sequence. The corresponding field then turned red as a feedback that the score was selected. To confirm the selection the participants then had to look at the field 'Valider', which then also turned red, and the presentation of the following sequence was initiated. The participants also had a chance to correct the rating if they accidently selected the wrong field or if they were not satisfied with their choice. In this case, they needed to look at the 'Corriger' field instead, which turned all fields white again, allowing for a new rating to be conducted. No restrictions regarding the number of corrections and no time limits were imposed on the quality rating procedure.

Given the length of 6 s per video sequence and the time for the impairment rating between the sequences, each experiment lasted about 30-40 min. To avoid fatigue of the viewers' eyes, we included a break after about half of the sequences was presented.

12.2.5 Recorded data and post-processing

For each video sequence, we recorded 30 impairment scores and 30 gaze patterns using the eye tracker. Given the length of 6 s and the recording rate of the eye tracker of 500 GP/s, about 3000 GP were recorded per person and video sequence. Consequently, about 90000 GP were recorded for each sequence as a total over all viewers. As with the gaze patterns from the eye tracking experiments E4a and E4b, these GP have to be post-processed into VFP and SM. This process is substantially more challenging for video as it is for images since the visual scene is changing constantly and thus, GP recorded in different locations do not necessarily relate to a change of FoA but can instead be caused by smooth pursuit eye movements following a moving object through the scene. The post-processing of the GP is discussed in Section 13.2.1.

13 Impact of Content Saliency on Packet Loss Distortion Perception

The subjective impairment ratings and the eye tracking data obtained from experiment E5 are in the following analysed in detail with the aim to establish a better understanding of the perceived annoyance of the packet loss distortions. In this respect, we are particularly interested in evaluating the impact of the video content, the saliency of the distortion region, and the distortion duration on the overall perceived annoyance.

We first focus on the analysis of the impairment ratings given by the participants and evaluate whether distortions in a salient region have been perceived to be more annoying, as compared to distortions in a non-salient region. It is shown that there is indeed a highly significant impact of the content saliency on the perceived distortion annoyance. This is true for different distortion durations and for a wide variety of different video sequence content. The distortion duration, on the other hand, is revealed to have a comparably lower impact than the content saliency.

The recorded gaze patterns are evaluated with respect to the amount of attention that the distortion regions have received, with the goal to identify whether particular distortions have actually been focused on or not. It is revealed that the FoA is indeed shifted, in particular in the sequences that are distorted in the non-salient region, SEQ$_{N,0.4}$ and SEQ$_{N,1.2}$. However, the shift is strongly dependent on the video content and the distortion visibility.

The results from experiment E5 and the insights provided in this chapter are considered to be highly valuable to gain a better understanding about the perceived annoyance of packet loss distortions in relation to content saliency. They are further beneficial for the development of VQM that take into account content saliency in their quality estimation. In fact, in Chapter 14 we discuss some simple models that are deployed to make existing VQM saliency aware, showing strong improvement in prediction performance.

13.1 Perceived annoyance of packet loss distortions

The subjective impairment ratings from experiment E5 are analysed in detail in the following. In this respect, we focus on identifying the relative influence of the content saliency in the packet loss distortion region, the distortion duration, and the video sequence content, on the overall perceived annoyance.

13.1.1 Distortion class specific MOS differences

The 30 subjective impairment ratings for each sequence are averaged into MOS. Corresponding to the subsets of sequences, SEQ$_{R}$, SEQ$_{S,0.4}$, SEQ$_{N,0.4}$, SEQ$_{S,1.2}$, and SEQ$_{N,1.2}$ (see Sec. 12.1.2), we define subsets of MOS as MOS$_{R}$, MOS$_{S,0.4}$, MOS$_{N,0.4}$, MOS$_{S,1.2}$, and MOS$_{N,1.2}$, respectively. To evaluate the impact of content saliency and distortion duration on the overall annoyance, we further define MOS differences, Δ_{MOS}, as follows

$$\begin{aligned}\Delta_{\text{MOS},0.4} &= \text{MOS}_{N,0.4} - \text{MOS}_{S,0.4} && (97)\\ \Delta_{\text{MOS},1.2} &= \text{MOS}_{N,1.2} - \text{MOS}_{S,1.2} && (98)\\ \Delta_{\text{MOS},S} &= \text{MOS}_{S,0.4} - \text{MOS}_{S,1.2} && (99)\\ \Delta_{\text{MOS},N} &= \text{MOS}_{N,0.4} - \text{MOS}_{N,1.2}. && (100)\end{aligned}$$

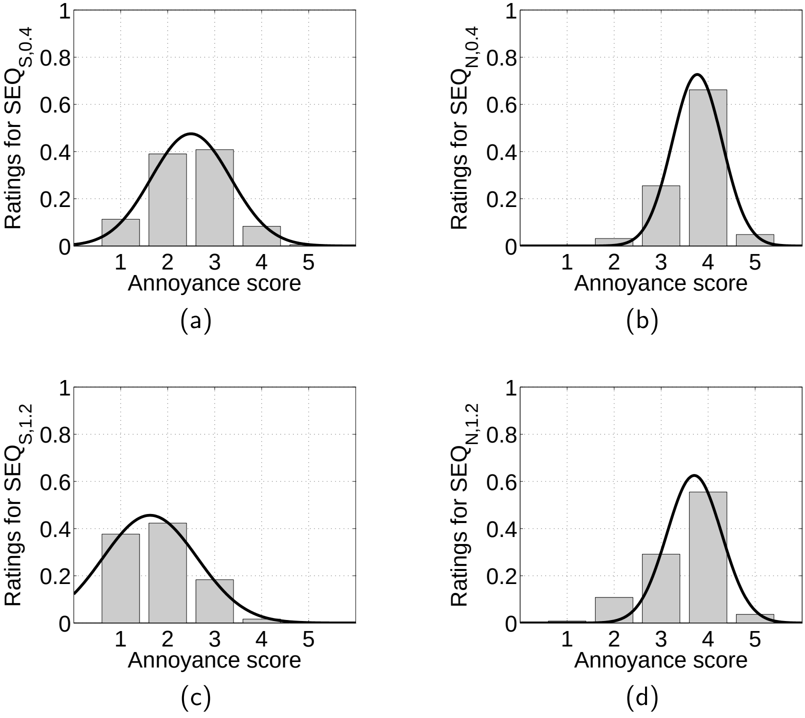
Here, for instance, $\Delta_{\text{MOS},0.4}$ represents the MOS difference between the salient (S) and the non-salient (N) region in case of short distortion propagation of 0.4 s. Similarly, $\Delta_{\text{MOS},S}$ represents the MOS difference between short (0.4 s) and long (1.2 s) distortion propagation in case of distortions in the salient region.

13.1.2 Distribution of impairment ratings

Given the 30 participants and the 100 video sequences, a total of 3000 impairment ratings were collected during the experiment. As such, 600 ratings were given for each subset of sequences. The normalised distribution of the ratings for the four subsets of distorted sequences is presented in Fig. 74 (the subset of reference sequences, SEQ$_{R}$, has been left out as almost exclusively all ratings were equal to 5). Here, the number of ratings for each annoyance score have been normalised with respect to the total number of 600 ratings within each subset.

Figure 74 shows a strong tendency that the salient region distorted sequences, SEQ$_{S,0.4}$ and SEQ$_{S,1.2}$, received generally lower ratings as compared to the non-salient region distorted sequences, SEQ$_{N,0.4}$ and SEQ$_{N,1.2}$. It can also be observed that the ratings for SEQ$_{S,0.4}$ and SEQ$_{S,1.2}$ are generally more spread as compared to SEQ$_{N,0.4}$ and SEQ$_{N,1.2}$, which observe high peaks at an annoyance score of 4. These observations indicate, that the ratings are more similar between the sequence subsets that contain distortions in the same region (salient or non-salient) than between the sequence subsets that contain distortions of the same duration (long or short).

To further illustrate the above observations, we have conducted a curve fitting



(a)

(b)

(c)

(d)

Figure 74: Normalised distributions of the total number of ratings for the four distorted sequence subsets: (a) SEQ$_{S,0.4}$, (b) SEQ$_{N,0.4}$, (c) SEQ$_{S,1.2}$, and (d) SEQ$_{N,1.2}$.

of the score distributions using a Gaussian fitting function as

$$y(x) = p_1 \cdot e^{-\left(\frac{x-p_2}{p_3}\right)^2}. \tag{101}$$

The fitting function parameters as well as the goodness of fit measures are summarised in Table 36 for all four distorted sequence subsets. Here, the parameter p_1 determines the height of the distribution maximum, the parameter p_2 represents the corresponding value on the annoyance score scale, and the parameter p_3 is related to the width of the Gaussian fitting curve. These parameters provide quantitative evidence that the ratings of SEQ$_{S,0.4}$ and SEQ$_{S,1.2}$ are similarly

Table 36: Gaussian curve fitting for the total number of ratings within all distorted sequence subsets.

Subset	Fitting parameters			Goodness of fit	
	p_1	p_2	p_3	R^2	RMSE
SEQ$_{S,0.4}$	0.476	2.495	1.195	0.996	0.016
SEQ$_{N,0.4}$	0.727	3.769	0.753	0.997	0.021
SEQ$_{S,1.2}$	0.456	1.625	1.418	0.998	0.009
SEQ$_{N,1.2}$	0.625	3.706	0.824	0.95	0.072

distributed, as are the ratings of SEQ$_{N,0.4}$ and SEQ$_{N,1.2}$. The corresponding goodness of fit measures, the root mean squared error (RMSE) and the squared correlation coefficient R^2, further show that the ratings of all four subsets can be accurately fitted using a Gaussian distribution.

13.1.3 Impairment ratings averaged over distortion classes

The results from the previous section indicate that the sequences with distortions in the salient region generally receive lower ratings as compared to the sequences with distortions in the non-salient region. Further evidence of this observation is given by the MOS computed for each of the five subsets and averaged over all 20 different contents, which is presented in Table 37. It can be seen that, expectedly, SEQ$_R$ received the highest MOS, followed by SEQ$_{N,0.4}$, SEQ$_{N,1.2}$, SEQ$_{S,0.4}$, and SEQ$_{S,1.2}$. Thus, as an average over a large number of different contents, the distortions in the non-salient region were perceived as far less annoying in comparison to the salient region. It is particularly worth noting that MOS$_{N,1.2}$ received an average MOS that is 1.02 higher than MOS$_{S,0.4}$, even though the distortion in the non-salient region is three times longer than the distortion in the salient region.

On the other hand, the distortion duration seems to play only a minor role as compared to the saliency of the location. This is particularly true for distortions in the non-salient region, where the small difference of 0.22 between MOS$_{N,0.4}$ and MOS$_{N,1.2}$ indicates only little impact of distortion duration on perceived annoyance. The larger difference of 0.64 between MOS$_{S,0.4}$ and MOS$_{S,1.2}$ suggests that the duration plays a more prominent role in the case of distortions appearing in the salient region.

Table 37: MOS averaged over all sequences within the five subsets of sequences.

MOS$_R$	MOS$_{N,0.4}$	MOS$_{N,1.2}$	MOS$_{S,0.4}$	MOS$_{S,1.2}$
4.97	3.72	3.5	2.48	1.84

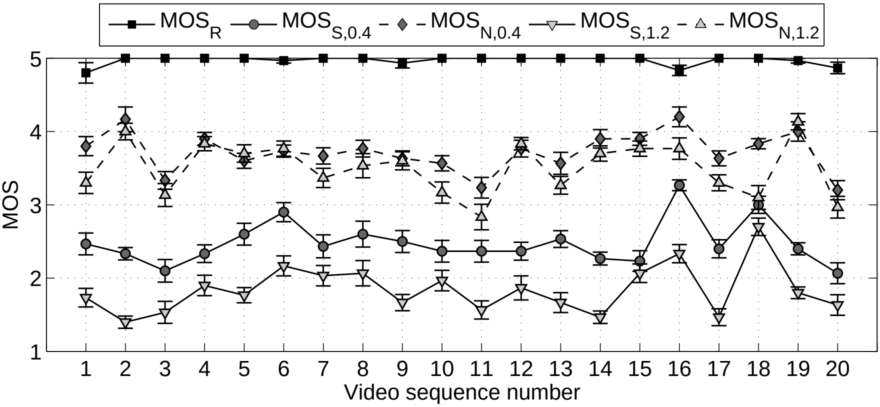


Figure 75: MOS and standard errors for all 20 contents of all sequence subsets (SEQ$_R$, SEQ$_{S,0.4}$, SEQ$_{N,0.4}$, SEQ$_{S,1.2}$, SEQ$_{N,1.2}$).

13.1.4 Dependency on natural video content

To identify whether the hierarchy of MOS presented in Table 37 is valid for different content, the MOS for all 20 sequence contents in the 5 subsets are presented in Fig. 75. It can be observed that the hierarchy of MOS between the subsets is almost exclusively the same as in Table 37 for all video sequences. This is a strong indication that the higher annoyance of distortions in the salient region as compared to the lower annoyance of distortions in the non-salient region is valid for a broad range of different video contents with strongly varying spatial and temporal characteristics (see SI and TI indicators in Fig. 69).

The Δ_{MOS} presented in Fig. 76 reflect the difference in annoyance both with respect to content saliency and distortion duration. It can be seen, that for almost all sequences the difference in MOS is considerably larger for $\Delta_{\text{MOS},0.4}$ and $\Delta_{\text{MOS},1.2}$ as compared to $\Delta_{\text{MOS},S}$ and $\Delta_{\text{MOS},N}$. These results support the above observations that the observers distinguished annoyance levels more pronounced

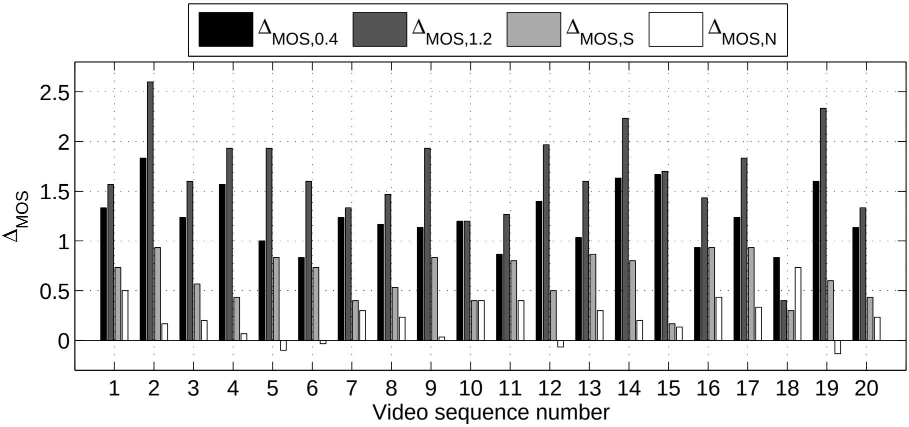


Figure 76: MOS differences for all 20 contents of the distorted sequence subsets (SEQ$_{S,0.4}$, SEQ$_{N,0.4}$, SEQ$_{S,1.2}$, SEQ$_{N,1.2}$).

with respect to the content saliency of the distorted region (salient or non-salient) as compared to distortion duration (long or short) and give further evidence that this is true for a large variety of different content. Figure 76 also shows that the difference between salient region and non-salient region is usually more pronounced in case of long distortions, $\Delta_{\mathrm{MOS},1.2}$, as compared to short distortions, $\Delta_{\mathrm{MOS},0.4}$. Similarly, the distinction between long and short distortions is observed to be more pronounced in the salient region, $\Delta_{\mathrm{MOS},S}$, as compared to the non-salient region, $\Delta_{\mathrm{MOS},N}$. In particular the small values of $\Delta_{\mathrm{MOS},N}$ indicate that the annoyance of distortions in the non-salient region varies only very little with respect to the duration. Similar observations were made on the MOS averaged over the whole subset and indeed, this appears to be true for a broad range of sequence contents.

The above observations apply for all sequences but one, sequence 18. This sequence contains a close up of a rugby game with extremely high motion over large parts of the frames, which is also apparent by it having the highest TI indicator out of all sequences (see Fig. 69). Here the distinction between long and short distortions was in fact stronger in the non-salient region. This may due to stronger masking effects caused by the extremely high motion in the salient region, as compared to the lower motion in the background. As such, the distortions in the salient region were not perceived as severe, which also explains the fairly high

MOS scores. This may also be the reason for the distinction between salient region and non-salient region being more pronounced for the short distortion duration, unlike for all other sequences, where it is more pronounced for the long distortion duration.

Videos containing very high motion have already been found by Wang et al. [253] to have a considerable impact on the parameters of their packet loss distortion model. Furthermore, Lin et al. [113] found that packet loss visibility strongly depends on camera motion. These results and the findings presented in this thesis indicate that high motion video sequences in the presence of packet loss distortions may need particularly careful treatment.

13.1.5 Analysis of variance with respect to the distortion classes

From the analysis of the MOS in this section we can summarise, so far, two observations. Firstly, the content saliency of the distortion region seems to have a stronger impact on the MOS as compared to the distortion duration. This is particularly apparent in the MOS averaged over the four distortion classes, as presented in Table 37. Secondly, the hierarchy of MOS with respect to the different distortion classes is almost exclusively the same for all 20 different video contents (see Fig. 75). The relative difference between the MOS of the distortion classes, however, varies between the different video contents, as can be observed from Fig. 76. It is thus of interest to more precisely quantify the impact of the three factors (content, saliency, distortion duration) on the MOS, and also to determine the interaction between the different factors.

For this purpose, we conduct an analysis of variance (ANOVA) [242] of the three factors with respect to the MOS values. The general idea of ANOVA is that the considered factors, also referred to as independent variables, affect the MOS, the dependent variable, in a particular way. The degree to which each of the independent variables affects the dependent variable is determined by comparing the variances between each factor level to the variances of the samples within each factor level. In our case, we have three different factors, the video content, the saliency of the distortion region, and the distortion duration, which are in the following referred to as $\mathcal{F}_C$, $\mathcal{F}_S$, and $\mathcal{F}_D$, respectively. Here, factor $\mathcal{F}_C$ has twenty levels represented by the twenty different contents of the reference video sequences. Factor $\mathcal{F}_S$ has two levels (salient/non-salient), as does factor $\mathcal{F}_D$ (0.4 s/1.2 s). The effect of these three factors on the MOS are referred to as the main effects, whereas the mutual effects of these factors on the MOS are referred to as the interaction effects. Given the three factors, we conduct a three-way ANOVA that evaluates both the main effects of the three factors and also the interaction between them.

The ANOVA is essentially based on several hypotheses with respect to the main effects and the interaction effects. In particular, the null-hypothesis for the main effects states that the means between the different levels of a factor are drawn from the same population. This would relate to a comparably low variance between the factor levels as compared to the variance between the samples of each factor level. As such, the difference between the different levels of a factor would be considered to be not significant. Similarly, the null-hypothesis for the interaction effects states that there is no interaction between different factors.

Rejecting any of the above hypotheses means that the respective main factor or interaction has a significant effect on the MOS. Whether or not to reject a null-hypothesis is determined using the F-test [254] and the related F value and the probability p. The null-hypothesis is rejected when the value of F exceeds a critical value, where the critical value corresponds to the limits of the confidence interval (CI) of the sample distribution within a factor level. The CI is typically chosen to be 95% and thus, a F value being equal to the critical value relates to a probability of $p = 0.05$ that the F value has been obtained, given that the null-hypothesis is true. In other words, given that two factor means are indeed from the same population, there is only a 5% chance that an F value as extreme as the critical value would have been obtained. Therefore, the null-hypothesis is typically rejected for F values larger than the critical value, relating to probabilities $p < 0.05$ of false rejection (Type I error).

The results of the three-way ANOVA are presented in Table 38. Here, the sum of squares (SS), the degrees of freedom (DoF), the mean squares (MS), the F value, and the probability p are given for the main effects and the first order interactions. From the main effects one can observe that indeed all three factors are highly significant as their respective p values are well below the threshold of $p = 0.05$ for rejection of the null-hypothesis. The magnitudes of the different p values further reveal that the saliency factor $\mathcal{F}_S$ has most impact on the MOS, followed by the distortion duration, $\mathcal{F}_D$, and the video content, $\mathcal{F}_C$. Regarding the interactions, it can be observed that there is a significant interaction between the content and the saliency, $\mathcal{F}_C \times \mathcal{F}_S$, and also between the saliency and the distortion duration, $\mathcal{F}_S \times \mathcal{F}_D$. There seems, however, to be no significant interaction between the content and the distortion duration, $\mathcal{F}_C \times \mathcal{F}_D$.

These results confirm the observations that the content saliency indeed has the strongest impact on the perceived annoyance of the distortions. The content of the video sequences, however, also plays a significant role with regards to the degree to which the saliency impacts on the distortion perception.

The results found here disagree with the findings by Moore et al. [255], who

Table 38: Three-way ANOVA table showing the main effects and interactions of the three factors, content $\mathcal{F}_C$, saliency $\mathcal{F}_S$, and distortion duration $\mathcal{F}_D$, in relation to the MOS values.

Effects	Factors	SS	DoF	MS	F	p
Main	$\mathcal{F}_C$	4.04	19	0.213	7.62	$2.404 \cdot 10^{-5}$
	$\mathcal{F}_S$	42.244	1	42.244	1513.76	0
	$\mathcal{F}_D$	3.641	1	3.641	130.47	$5.934 \cdot 10^{-10}$
Interaction	$\mathcal{F}_C \times \mathcal{F}_S$	2.442	19	0.129	4.61	$8.292 \cdot 10^{-4}$
	$\mathcal{F}_C \times \mathcal{F}_D$	0.446	19	0.024	0.84	0.645
	$\mathcal{F}_S \times \mathcal{F}_D$	0.882	1	0.882	31.61	$2.021 \cdot 10^{-5}$
Error		0.53	19	0.028	-	-
Total		54.224	79	-	-	-

reported only a minor impact of the importance in a video sequence on the overall distortion annoyance. In their work, however, the importance was not based on visual saliency, but rather on segmentation of the video frames into nine equally sized rectangles which were subsequently rated by human observers regarding their importance. Hence, the methodological differences may lead to different conclusions, as in our work the SM used are mainly based on bottom-up attentional processes whereas the importance ratings in [255] are mainly based on top-down processes. Furthermore, the FoA recorded through eye tracking, as we used it, shifts more dynamically with the content of the video sequences, unlike the static importance rectangles used in [255].

13.1.6 Detection of distortions

In the previous sections the perceived annoyance of the packet loss distortions has been analysed in detail. In this section, we evaluate whether distortions have been detected at all and how the distortion detection relates to the video content, the saliency, and the distortion duration. The 'Imperceptible' rating given in the impairment scale provides, in this context, valuable information whether distortions have actually been detected by the observer or not. As such, the 'Imperceptible' rating (impairment ratings equal to 5) corresponds to distortions that were not

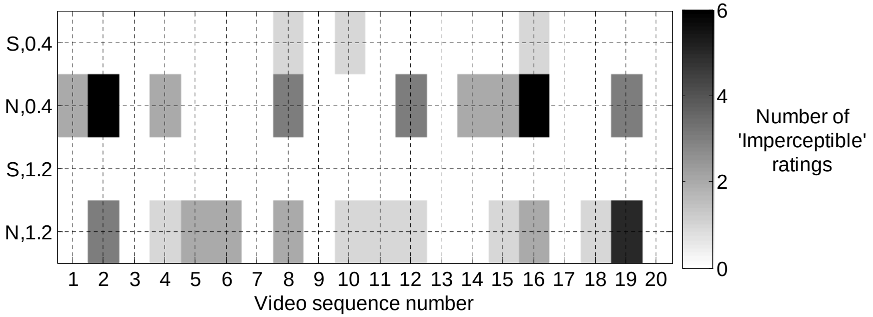


Figure 77: Number of 'Imperceptible' ratings for all 20 contents of the distorted sequence subsets (SEQ$_{S,0.4}$, SEQ$_{N,0.4}$, SEQ$_{S,1.2}$, SEQ$_{N,1.2}$).

detected in the sequences, whereas the other possible ratings (impairment ratings smaller than 5) relate to the degree of annoyance of detected distortions.

The number of 'Imperceptible' ratings are visualised in Fig. 77 for all 20 video contents of the distorted sequence subsets. It can be seen that many 'Imperceptible' ratings have been given for SEQ$_{N,0.4}$ (29 ratings) and SEQ$_{N,1.2}$ (22 ratings), whereas only few have been given for SEQ$_{S,0.4}$ (3 ratings) and in fact none for SEQ$_{S,1.2}$. This is thought to be due to mainly two reasons. Firstly, as the attention is usually on the salient region, the observer is more likely to miss distortions in non-salient regions. Secondly, salient regions typically exhibit features that facilitate stronger visualisation of distortions, such as high local contrast, and distinguished shapes and colours. Non-salient regions are often composed of image parts that are more uniform, such as a sky or a water surface.

It can be further observed from Fig. 77 that for sequences 2, 16, and 19 there was a particularly high number of 'Imperceptible' ratings in the sequences that are distorted in the non-salient region. These three sequences exhibit fairly uniform non-salient regions and in addition, the attention of the observers is strongly focused on the salient regions in all three sequences, as indicated by the HM in Fig. 70 and Fig. 71.

13.2 Visual attention to localised packet loss distortions

The gaze patterns recorded during the task-free eye tracking experiment (see Section 12.1.1) served to identify the content saliency of the video sequences.

On the other hand, the gaze patterns recorded during experiment E5 do not directly reflect the saliency of the visual content, as the experiment was conducted under quality assessment task. As such, the gaze patterns and the related SM reflect the viewing behaviour of the participants while evaluating the quality of the sequences.

In the following, we evaluate the gaze patterns recorded in experiment E5, to gain some more insight into the viewing behaviour of human observers when assessing the quality of packet loss distorted video sequences. Similar to the analysis of the image eye tracking data in Chapter 11, we first create SM for each video frame. We then perform a ROC analysis (see Section 11.3.3) and evaluate in particular the AUC, to quantify the attendance of the observers in the distortion regions during the quality assessment task. A higher AUC in this context reflects a stronger focus of the observers on the distortions and thus, a higher likelihood that the distortions have been consciously attended. The AUC analysis is conducted on a frame-by-frame basis for each video and simple statistics, such as the mean, are further determined to highlight interesting observations.

13.2.1 Creation of frame-based saliency maps

As a basis for analysis, the gaze patterns of all observers were first transferred into VFP, by means of clustering, and then converted into SM using a Gaussian filter kernel, similar to the procedures explained in Section 11.1. The difference to the SM created for the still images is the temporal change of the visual scene that is to be accounted for in case of the videos. The main difficulty in this respect is to distinguish between smooth pursuit eye movements, that are deployed to move the FoA along with moving objects, and the saccadic eye movements, that are used to shift the FoA between different objects in the scene. We chose a threshold of 25 dva/s to distinguish between smooth pursuit and saccadic eye movements. The reason for this being, that there were no object motions in the sequences that exceeded this speed and as such, higher eye movement speeds were attributed to saccades.

13.2.2 Frame-based ROC analysis and AUC computation

In Section 11.3.3, the ROC analysis was conducted to quantify the relative amount of saliency in the ROI and the BG, where the SM pixels within the ROI were considered to belong to the true positive class and the pixels outside the ROI considered to belong to the true negative class. In a similar fashion, we perform the ROC analysis here to quantify the relative amount of saliency in the distorted frame regions and the undistorted frame regions. We define the SM pixels within the distortion regions to belong to the true positive class and the SM pixels outside the distortion regions to belong to the true negative class.

The ROC analysis is performed on a frame-by-frame basis for all the sequences of the four distortion classes, SEQ$_{S,0.4}$, SEQ$_{N,0.4}$, SEQ$_{S,1.2}$, and SEQ$_{N,1.2}$, taking into account the different distortion locations in the salient and non-salient regions and the range of distorted frames for all 20 contents. It should be noted that the ROC was not just conducted on the distorted frames but on all 150 frames of each sequence, thus facilitating the analysis of shifts in attention due to the distorted video frames as compared to the undistorted video frames. In this respect, the AUC is consulted as a measure of the amount of saliency in the distorted regions as compared to the undistorted regions, with a higher AUC indicating a stronger focus of the observers on the distortion regions. With all distortions introduced through the packet loss being in the suprathreshold regime, one can assume the attention of the observers to be shifted towards the distortions during their course of appearance. As a result, the AUC of the distorted frames would rise in comparison to the AUC of the undistorted frames. The AUC computations are in the following denoted as AUC$_{S,0.4}$, AUC$_{N,0.4}$, AUC$_{S,1.2}$, and AUC$_{N,1.2}$, in relation to the respective distortion classes of the video sequences.

13.2.3 Attentional shifts due to distortions: an illustrative example

Before going into the details of the ROC analysis results, an illustrative example of viewing behaviour that we have observed for a wide range of video sequences is discussed. For this purpose, representative SM are shown in Fig. 78, superimposed as HM on the corresponding distorted sequences, SEQ$^{(13)}_{S,1.2}$ and SEQ$^{(13)}_{N,1.2}$, in the left and the right column, respectively. The distortions in these sequences propagated over 30 frames, from frame 111 to frame 140.

The first row shows the SM for frame 101 of the respective videos, which does not contain any distortions and is presented here to illustrate the VA before the appearance of the distortions. It can be noted that the attention is mainly in the lower part of the frame and in particular around the region that has been identified as highly salient, based on the task-free eye tracking data (see Section 12.1.1). However, the focus of the observers is not quite as condensed as in the task-free experiment (see Fig. 71), which might be due to the quality task under which the observers in experiment E5 were viewing the videos. Thus, the FoA might be moved to some degree to other regions of the frame to conduct the quality analysis.

The second row, showing frame 111, represents the first frame in which the

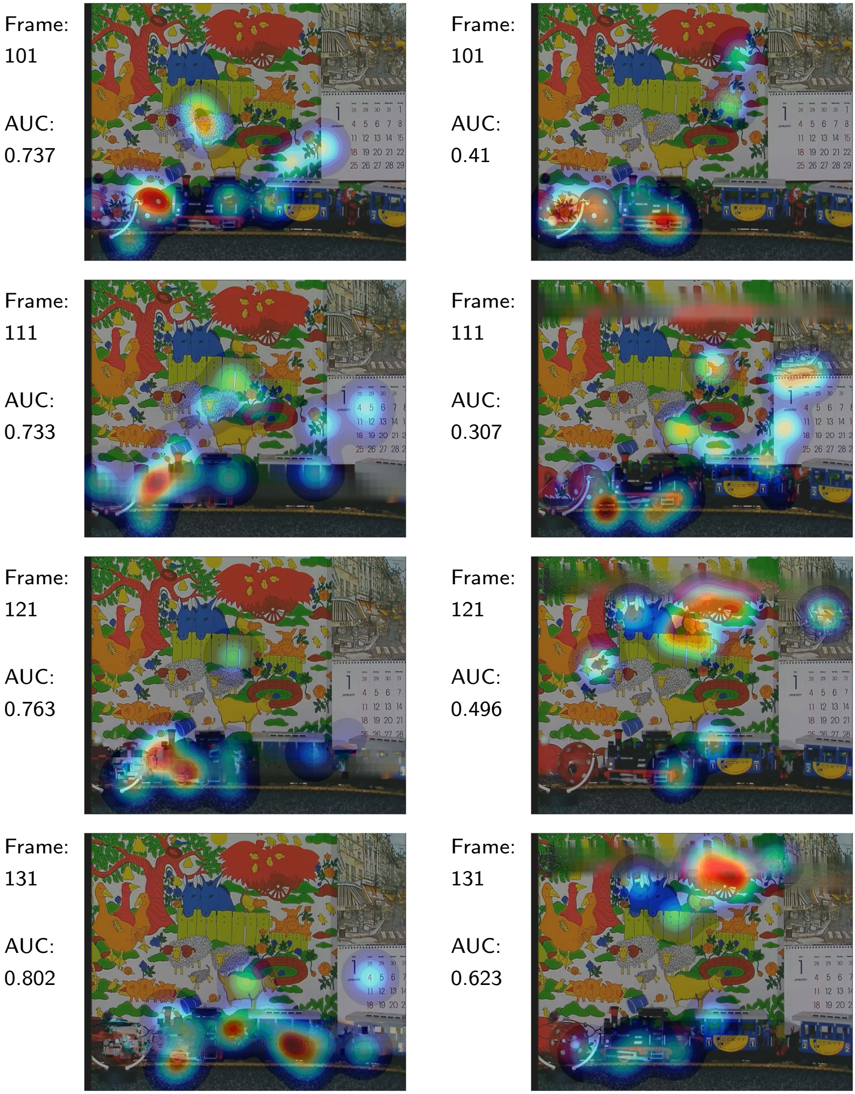


Figure 78: Example frames of SEQ$^{(13)}_{S,1.2}$ (left) and SEQ$^{(13)}_{N,1.2}$ (right) with distortions ranging from frame 111 to frame 140.

packet loss distortions appear. In the left column, the distortions clearly appear at the bottom of the frame, around the highly salient region. In the right column, the distortions appear towards the top of the frame, in the non-salient region.

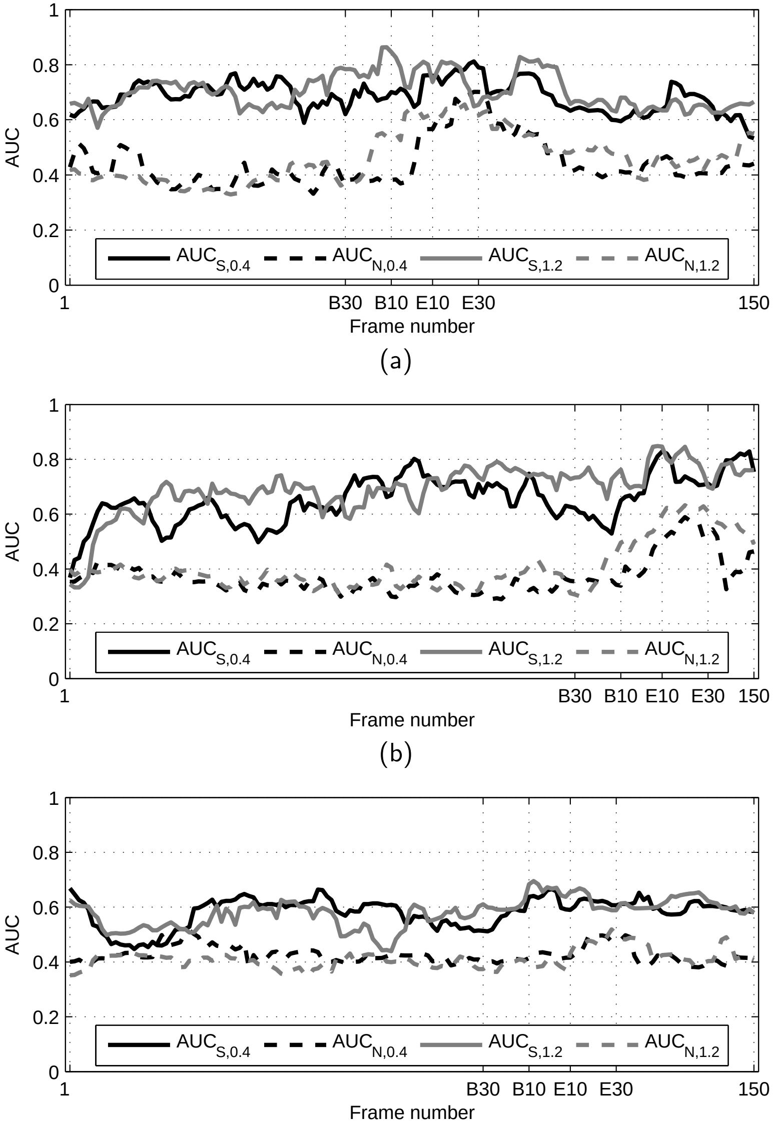
The third and fourth row, respectively, show frames 121 and frame 131 of the according sequences. It can clearly be seen in the left column that for SEQ$^{(13)}_{S,1.2}$ the FoA remains in the bottom of the frame, and in fact, even more so than in the previous frames. On the other hand, the right column shows that for SEQ$^{(13)}_{N,1.2}$ the FoA gradually shifts towards the distortions at the top of the frame, away from the salient region at the bottom.

The corresponding AUC values beside each frame reflect this behaviour. To be precise, the AUC values for SEQ$^{(13)}_{S,1.2}$ in the left column are fairly high for all four frames and increase only marginally with the appearance of the distortions. The AUC values for SEQ$^{(13)}_{N,1.2}$ are considerably smaller for the early frames and increase strongly for later frames, as the FoA is drawn away from the content saliency to the localised packet loss distortions. Similar observations have been made for a wide range of sequences, however, not all of them exhibit these phenomena. A more detailed analysis of all sequences based on the computation of the AUC is presented in the following sections.

13.2.4 Temporal progression of the AUC

To illustrate the progression of the AUC over the course of a sequence, the frame-by-frame AUC for three example videos are shown in Fig. 79, of which the one in the middle, SEQ$^{(13)}$, corresponds to the example frames presented in Fig. 78. The sequence numbers relate to the numbers provided in Fig. 70 and Fig. 71. For each sequence, four frame-based AUC curves are plotted, corresponding to the different distortion classes (SEQ$_{S,0.4}$, SEQ$_{N,0.4}$, SEQ$_{S,1.2}$, SEQ$_{N,1.2}$). The abscissae denote the frame numbers with labels provided for the beginning and the end of the temporal distortion ranges. Here, B30 and E30 denote the beginning and the end of the distortions in the GOP30 coded videos and B10 and E10 denote the beginning and the end of the distortions in the GOP10 coded videos.

As one can see from Fig. 79 (a) and (b), there is indeed a noticeable increase of the AUC in the distorted frames as compared to the undistorted frames. However, this increase is considerably larger for the sequences that are distorted in the non-salient region, SEQ$_{N,0.4}$ and SEQ$_{N,1.2}$, showing a stronger shift in the observers FoA. This can be explained by the saliency being generally higher in the distortion regions for the salient region distorted sequences, SEQ$_{S,0.4}$ and SEQ$_{S,1.2}$, and as such, the viewers' focus is already there and does not need to be shifted to the



(a)

(b)

(c)

Figure 79: Frame-based AUC computation for all four distortion classes of the video sequences: (a) SEQ$^{(11)}$, (b) SEQ$^{(13)}$, and (c) SEQ$^{(16)}$.

same extent as for the sequences $\mathsf{SEQ}_{N,0.4}$ and $\mathsf{SEQ}_{N,1.2}$.

It is further worth pointing out the delay with which the AUC rises in relation to the appearance of the distortions. From Fig. 79 (a) and (b) one can observe that the AUC starts to increase about 5 frames after the first distorted frame, which is apparent for both the GOP10 and the GOP30 coded sequences. This delay corresponds to about 200 ms, given the frame rate of 25 fps with which the videos were presented. Similar delays were observed for the other video sequences.

These observations made for video sequences $\mathsf{SEQ}^{(11)}$ and $\mathsf{SEQ}^{(13)}$ do not hold to the same degree for $\mathsf{SEQ}^{(16)}$, as can be observed in Fig. 79 (c), where the AUC does not rise notably during the distorted video frames. Unlike video sequences $\mathsf{SEQ}^{(11)}$ and $\mathsf{SEQ}^{(13)}$, video sequence $\mathsf{SEQ}^{(16)}$ contains very high and fast motion in conjunction with a highly textured background. As such, the distortions in the non-salient regions do not attract as much attention, or are not even detected at all, which is also evident in Fig. 77 with $\mathsf{SEQ}^{(16)}$ having received many 'Imperceptible' ratings.

13.2.5 Impact of the content saliency and distortion duration

Figure 79 shows that for all three sequences, the $\mathsf{AUC}_{S,0.4}$ and $\mathsf{AUC}_{S,1.2}$, corresponding to the salient region distorted sequences, are generally higher as compared to the $\mathsf{AUC}_{N,0.4}$ and $\mathsf{AUC}_{N,1.2}$, corresponding to the non-salient region distorted sequences. To verify if this is a general phenomenon for all 20 sequence contents, we compute the average AUC, $\mu^{(AUC)}$, over all 150 frames for each video. The results are presented in Fig. 80, showing, that indeed the average AUC is consistently higher for $\mathsf{SEQ}_{S,0.4}$ and $\mathsf{SEQ}_{S,1.2}$ as compared to $\mathsf{SEQ}_{N,0.4}$ and $\mathsf{SEQ}_{N,1.2}$.

Intuitively one would expect this result, as the distortions in $\mathsf{SEQ}_{S,0.4}$ and $\mathsf{SEQ}_{S,1.2}$ are in the regions of high saliency from the task-free eye tracking experiment (see Section 12.1.1). However, it should be emphasised here again, that the gaze patterns obtained in experiment E5 do not directly reflect the content saliency but instead the viewing behaviour under quality assessment task. Thus, the generally higher AUC in the sequences $\mathsf{SEQ}_{S,0.4}$ and $\mathsf{SEQ}_{S,1.2}$ reflects the fact, that the human observers tend to analyse the quality of the video sequences in the highly salient regions, rather than in the non-salient regions. This is in strong agreement with the results from Section 11.3 where we found that human observers rate image quality within ROI. To be more precise, the observers tend to start the assessment in the ROI and once inspected they, to some degree, move on to other parts of the image. As the visual scene constantly changes in the case of video sequences, the effect of moving on to a non-salient region

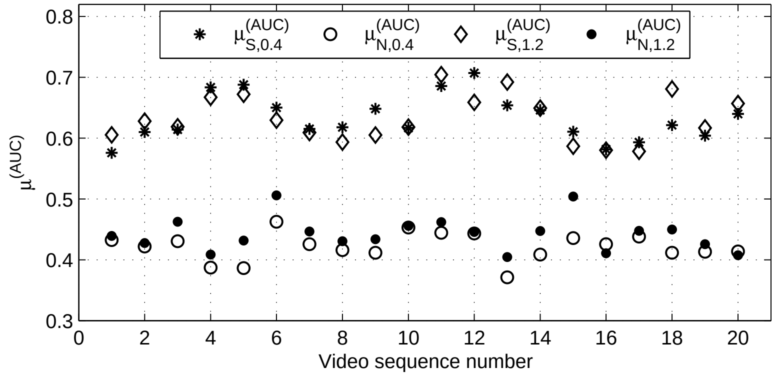


Figure 80: AUC averaged over all 150 frames for all sequences and distortion classes.

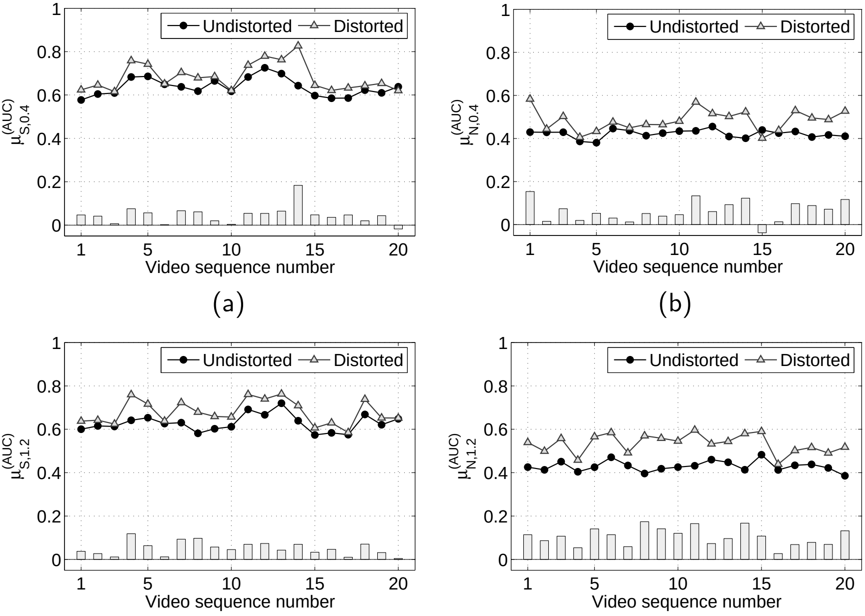
seems somewhat suppressed, as the AUC remains generally higher for SEQ$_{S,0.4}$ and SEQ$_{S,1.2}$ throughout the duration of the sequences.

Figure 80 also shows that SEQ$_{S,0.4}$ and SEQ$_{S,1.2}$ generally experience similar AUC values, and so do SEQ$_{N,0.4}$ and SEQ$_{N,1.2}$. However, SEQ$_{S,0.4}$ and SEQ$_{S,1.2}$ do not exhibit a consistent pattern regarding one of the two having a higher AUC than the other. The AUC$_{N,1.2}$, on the other hand, seems to be mostly higher than AUC$_{N,0.4}$, which indicates that the longer distortions were on average longer attended by the observers, as compared to the shorter distortions.

13.2.6 Attendance of distorted versus undistorted frames

To further quantify the amount to which the FoA shifts with respect to the distortions, we compute the average AUC $\mu^{(AUC)}$ independently for the distorted frames and the undistorted frames. The average AUC for the distorted frames, $\mu^{(AUC,D)}$, is thus computed over 10 and 30 frames, respectively, for the GOP10 and GOP30 coded sequences. The remaining frames are used to compute the average AUC for the undistorted frames, $\mu^{(AUC,U)}$. Given the delay in the shift of the FoA, the computation of the average AUC of the distorted frames is actually also shifted by 5 frames, or analogously, by 200 ms.

The average AUC for the undistorted and the distorted frames are shown in Fig. 81 for all 20 sequence contents and the 4 distortion classes. The bar plot at the bottom of each figure represents the difference between the respective average



(a) (b)

(c) (d)

Figure 81: AUC average over all distorted and undistorted frames for all sequences and distortion classes: (a) SEQ$_{S,0.4}$, (b) SEQ$_{N,0.4}$, (c) SEQ$_{S,1.2}$, and (d) SEQ$_{N,1.2}$. The bars at the bottom of each plot emphasise the difference between the average of the distorted and undistorted frames.

AUC for the undistorted and the distorted frames

$$\Delta\mu^{(AUC)} = \mu^{(AUC,D)} - \mu^{(AUC,U)}. \tag{102}$$

It can be seen that for all distortion classes, the average AUC is generally higher for the salient region distorted sequences. The AUC difference $\Delta\mu^{(AUC)}$ also highlight earlier observations, that the shift of the FoA is stronger for the non-salient region distorted videos, and in particular, in the case of the long distortions (SEQ$_{N,1.2}$).

Table 39: Pearson linear correlation coefficients between the AUC of the distorted frames, $\mu^{(AUC,D)}$, for the different distortion classes.

	$\mu_{S,1.2}^{(AUC,D)}$	$\mu_{N,0.4}^{(AUC,D)}$
$\mu_{S,0.4}^{(AUC,D)}$	0.764	0.01
$\mu_{N,1.2}^{(AUC,D)}$	0.032	0.271

13.2.7 Correlation analysis of average AUC in distorted frames

In Section 13.1, we showed that the distribution of impairment ratings changes considerably more with the saliency of the distortion region, as compared to the duration of the distortions (see Fig. 74). It was further shown that the average MOS exhibited larger differences with respect to the content saliency compared to the distortion duration (see Table 37). This predominance of the content saliency over the distortion duration is, to some degree, also apparent in the distortion attendance as analysed in this section. Particularly the averaged AUC presented in Fig. 80 clearly highlight the difference of the AUC with respect to the content saliency and the similarity of the AUC regarding the distortion duration.

To provide further insight regarding the impact of content saliency and distortion duration on the distortion attendance of the human observers, we compute the Pearson linear correlation coefficient ρ_P between the AUC of the distorted frames, $\mu^{(AUC,D)}$, over all 20 sequence contents. The results are presented in Table 39. With regards to the distortion duration; the correlation between $\mu_{S,0.4}^{(AUC,D)}$ and $\mu_{S,1.2}^{(AUC,D)}$ exhibits a fairly large value of 0.764, indicating that the attendance in the salient region distorted sequences SEQ$_{S,0.4}$ and SEQ$_{S,1.2}$ has been fairly similar in the case of both distortion durations. On the other hand, the correlation of 0.271 between $\mu_{N,0.4}^{(AUC,D)}$ and $\mu_{N,1.2}^{(AUC,D)}$ is very low, showing that in the non-salient region the attendance was not as stable between the two distortion durations. These observations may explain why the distinction between quality levels of the two distortion durations were more distinct in the case of sequences SEQ$_{S,0.4}$ and SEQ$_{S,1.2}$ (see Table 37), as these distortions were more consistently attended compared to the ones in the non-salient region.

From Table 39 it can further be seen that $\mu_{S,0.4}^{(AUC,D)}$ and $\mu_{N,0.4}^{(AUC,D)}$ are nearly uncorrelated. The same applies for $\mu_{S,1.2}^{(AUC,D)}$ and $\mu_{N,1.2}^{(AUC,D)}$. This indicates that there has been no common viewing pattern across the video sequences with respect to the two distortion regions (salient/non-salient) which suggests that the impact of content saliency is in fact somewhat larger as compared to the saliency caused by the transmission distortions. This can partly be attributed to the different degrees of visibility of the distortions and thus, the different degrees with which the FoA shifted towards these distortions.

14 Modelling Saliency Awareness for Video Quality Metrics

The outcomes of experiment E5 revealed that the content saliency of the region in which the localised packet loss distortions appear has a significant impact on the overall perceived annoyance of the distortions. In fact, the influence of the content saliency has been found to be more significant compared to the distortion duration. Thus, VQM that aim to accurately predict the perceived quality of a video sequence, in particular in the context of localised transmission errors, may not only focus on spatial and temporal distortion measures, but should also take into account the saliency of the video content.

The aim of this chapter is to determine the benefits of incorporating visual saliency information into existing VQM to improve their quality prediction performance. For this purpose, we consider a contemporary VQM, the temporal trajectory aware video quality measure (TetraVQM) [81]. This metric is based on numerous processing steps related to properties of the HVS, but does not take into account the content saliency. As such, TetraVQM is particularly suitable for an extension with a saliency awareness model.

The general saliency awareness framework considered in this work is shown in Fig. 82. Here, the white blocks denote the integral parts of most existing VQM. The grey blocks highlight the processing steps of the saliency awareness extension. It can be observed that the saliency awareness model is not integrated into the actual VQM, but instead constitutes a separate entity that is combined with the VQM in a final step. As such, the saliency awareness model can be added to existing VQM without having to conduct any changes to the VQM.

The saliency awareness model needs as input the regions of the visible distortions and also the saliency information of the visual content. In an applied scenario, the former could be determined from the distortion maps that are readily available from the VQM. The saliency information could be automatically predicted using VA models [129, 130]. In the context of this work, we omit these procedures and instead make use of the perfect knowledge that we have regarding the distortion region and the visual content saliency. To be precise, the distortion regions are known from the creation of the test sequences and the saliency information is obtained from the task-free eye tracking data. Thus, we avoid potential estimation errors of distortions and content saliency from objective methods, that would subsequently result in errors in the saliency awareness model.

Given the very recent conduction of the combined video quality and eye tracking experiment E5, the modelling and the results presented in this chapter are by

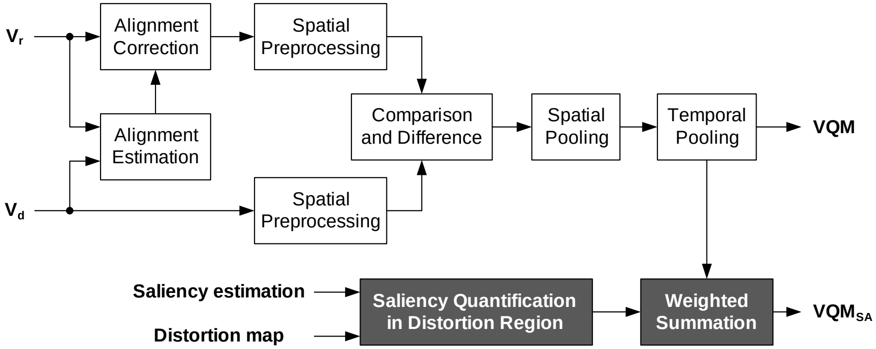


Figure 82: Extension of a conventional video quality metric (white blocks) with the saliency awareness model (grey blocks).

no means considered to be exhaustive. They are rather indicative and are considered to be a first step towards a more complete framework of saliency awareness for VQM that are not accounting for the saliency of the video content.

14.1 TetraVQM

TetraVQM [81] is an objective quality algorithm that is particularly well suited for enhancement with visual saliency because it already contains several steps that are motivated by the HVS. It has been designed for the prediction of video quality in multimedia scenarios, including the typical artifacts that occur in packet loss situations.

14.1.1 Essential processing steps

TetraVQM follows the FR approach and therefore uses the reference video sequence to predict the perceived visual quality of a corresponding distorted video sequence. The main focus is on temporal processing steps, e.g. the misalignment of the video sequences, frame freezes, frame skips, frame rate reduction, influence of scene cuts, and the tracking of the visibility of distorted objects. The processing starts with the spatial, temporal, and colour alignment of the two input videos, followed by the creation of a spatial distortion map for each video. The position and the severity of the degradations is then identified using the MSE.

A human observer perceives the video sequence as a continuous stream of

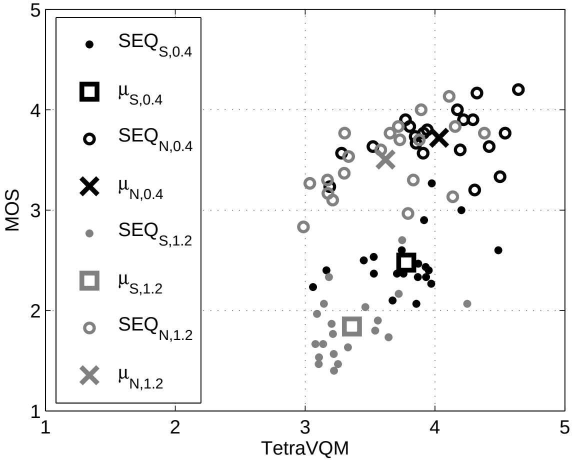


Figure 83: Scatter plot of MOS over TetraVQM, highlighting the four distortion classes and their mean values.

information, rather than image by image, and thus, the perceived severity of distortions depends on the duration that they have been seen by the observer. Furthermore, distortions which move together with an object are perceived as long lasting object degradations, rather than several isolated momentary points of distortions. Therefore, TetraVQM estimates the object motion and keeps track of the degradations over time. Each initial distortion map is then modified to account for the temporal visibility of the artifacts.

The spatial summation is performed by applying a filter that is based on the distribution of the cones in the fovea. Currently, the assumption is used that the viewer focuses on the point of the maximum perceived degradation. This was previously seen as the focal point of the observer. Thus, it is straightforward to improve the algorithm by applying a more sophisticated approach that uses the visual content saliency.

14.1.2 Omittance of content saliency

A scatter plot of the MOS from experiment E5 over TetraVQM is presented in Fig. 83. In this figure, the sequences corresponding to the four different subsets of distortions (SEQ$_{S,0.4}$, SEQ$_{N,0.4}$, SEQ$_{S,1.2}$, SEQ$_{N,1.2}$) are illustrated using different markers. In addition, the cluster means ($\mu_{S,0.4}$, $\mu_{N,0.4}$, $\mu_{S,1.2}$, $\mu_{N,1.2}$) are provided for all subsets.

The scatter plot highlights that TetraVQM accounts for the temporal duration of the distortions but not for the content saliency of the distortion region. The latter is evident given the big gap in MOS and the small gap in TetraVQM between the sequences with distortions in the salient ($\mathsf{SEQ}_{S,0.4}$,$\mathsf{SEQ}_{S,1.2}$) and non-salient ($\mathsf{SEQ}_{N,0.4}$,$\mathsf{SEQ}_{N,1.2}$) region. One could thus say that TetraVQM predicts the quality of the video sequences with distortions in the salient region too optimistically in relation to the video sequences with distortions in the non-salient region.

14.2 Saliency awareness model

The saliency awareness model is combined with a conventional VQM as shown in Fig. 82. The model consists of two integral parts, the first one being the saliency quantification in the distortion region, based on the provided input information about the distortions and the content saliency. The second part then consists of an appropriate pooling of the conventional VQM with the saliency awareness model.

In general, both the saliency quantification and the pooling stage can be implemented in various degrees of complexity. For instance, the saliency quantification in the distortion region could incorporate some interaction factor between the distortion visibility and the content saliency. In the scope of this work, we consider a simple saliency model in the form of a 'penalty term' that is added in relation to the amount of saliency in the distortion region as follows:

$$\mathsf{VQM}_{\mathsf{SA}} = \mathsf{VQM} - \alpha \cdot \mathsf{S} \tag{103}$$

where S denotes the saliency information. The idea behind (103) is to add a negative offset $\Delta = -\alpha \cdot \mathsf{S}$ to the VQM with respect to the amount of saliency information in the region where the distortions appear. This is based on the evidence from experiment E5, where it was shown that distortions in a more salient region are perceived as more annoying and as such, should receive a lower predicted quality score. The parameter S represents the saliency within the distortion regions of the videos and thus, determines the relative magnitude of the offset between different contents. The parameter α regulates the general degree with which the offset is performed and needs to be optimised for any particular VQM. The model outlined here is considered to be generally applicable to any VQM that does not take into account visual saliency.

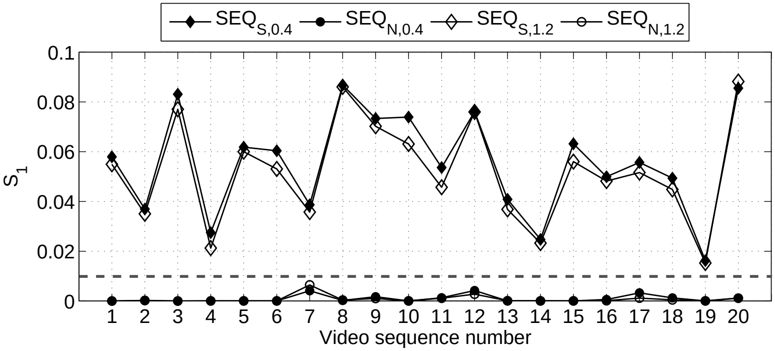


Figure 84: Saliency quantification S_1 for all 20 sequence contents in each subset.

In the following, we present two different saliency quantification methods that were found to considerably improve the quality prediction performance of TetraVQM.

14.2.1 Saliency quantification method S_1

The saliency awareness model using this first saliency quantification method, S_1, is in the following referred to as model M1. This method takes into account that the saliency within the distortion region varies between different videos. For this reason, the saliency in the distorted regions is quantified using the SM created from the gaze patterns. The mean saliency for each frame is then computed over the whole distortion region as

$$S_{1f} = \frac{1}{(lim_b - lim_t)(lim_r - lim_l)} \sum_{x=lim_b}^{lim_t} \sum_{y=lim_l}^{lim_r} S(x, y) \qquad (104)$$

where lim_b, lim_t, lim_l, and lim_r, respectively, denote the limits of the distortion region on the bottom, top, left, and right. In a temporal pooling step the mean over all degraded frames N_{df} is then computed as

$$S_1 = \frac{1}{N_{df}} \sum_{n=1}^{N_{df}} S_{1f}(n). \qquad (105)$$

The saliency magnitudes S_1 for all sequences are shown in Fig. 84. One can see that the sequences $\text{SEQ}_{S,0.4}$ and $\text{SEQ}_{S,1.2}$ contain a higher amount of saliency as compared to SEQ$_{N,0.4}$ and SEQ$_{N,1.2}$, however, the amount of saliency is not constant between the different sequences. As such, the offset changes with the saliency of the different sequences.

14.2.2 Saliency quantification method S$_2$

The saliency awareness model using the second saliency quantification method, S$_2$, is in the following referred to as model M2. This method does not distinguish between as many saliency levels as M1 does, but rather distinguishes only between two cases; salient region or non-salient region. This is realised with a threshold algorithm as follows:

$$\begin{aligned} \mathrm{S}_2 &= 1 \qquad \text{for} \qquad \mathrm{S}_1 \geq \tau \\ \mathrm{S}_2 &= 0 \qquad \text{for} \qquad \mathrm{S}_1 < \tau. \end{aligned} \tag{106}$$

Considering the results presented in Fig. 84, we define a threshold of $\tau = 0.01$ which separates the classes of saliency and non-saliency in the distorted image region. The threshold is indicated by the dashed grey line. As such, the VQM scores for the sequences SEQ$_{S,0.4}$ and SEQ$_{S,1.2}$ receive the same offset, whereas the VQM scores for the sequences SEQ$_{N,0.4}$ and SEQ$_{N,1.2}$ remain unaltered.

14.3 Performance evaluation

The quality prediction performance of TetraVQM is evaluated using three performance indicators; the RMSE, the Pearson linear correlation coefficient ρ_P, and the Spearman rank order correlation ρ_S. Prior to calculating the RMSE, a linear fit is applied in order to align the VQM output to the subjective rating scale. The optimal parameters α_{opt} for both models M1 and M2 are determined with respect to minimising RMSE between the saliency aware TetraVQM and the MOS using an exhaustive search. For further analysis of the saliency awareness models we also deployed them to the PSNR metric averaged over all frames of a video sequence. In the following, detailed results will be discussed only for TetraVQM.

The relation between the α and the RMSE is presented in Fig. 85 and the correspondence between α and the correlation coefficients is given in Fig. 86. The minimum RMSE and the maximum ρ_P and ρ_S are highlighted in the respective figures. The performance values are summarised in Table 40 for both TetraVQM and PSNR and their proposed enhancements as in TetraVQM$_{\mathrm{M1}}$, TetraVQM$_{\mathrm{M2}}$ and PSNR$_{\mathrm{M1}}$, PSNR$_{\mathrm{M2}}$, respectively. The performance results of TetraVQM and PSNR without the proposed enhancements indicate that these metrics are

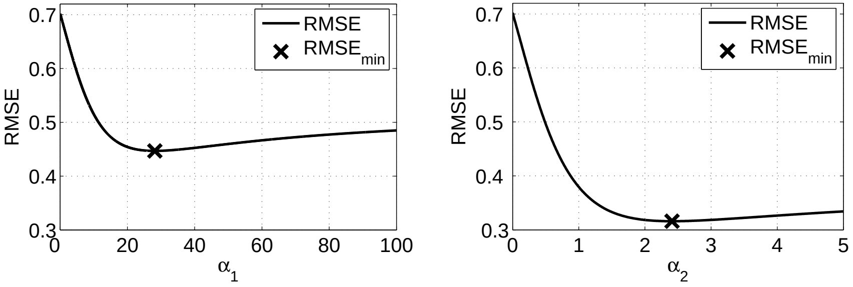


Figure 85: Root mean squared error (RMSE) versus α_1 and α_2.

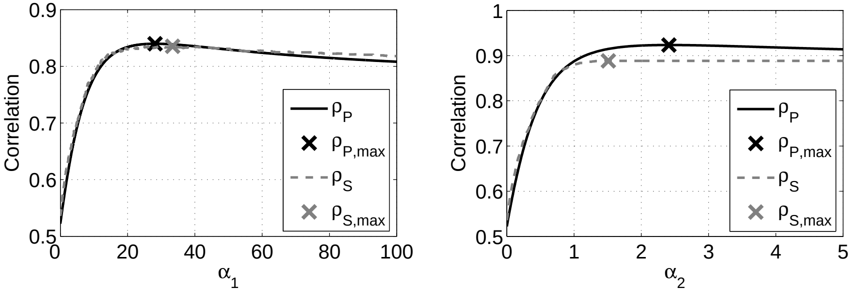


Figure 86: Pearson linear correlations (ρ_P) and Spearman rank order correlations (ρ_S) versus α_1 and α_2.

unable to predict the MOS given for the videos with the localised packet loss distortions. When comparing between TetraVQM and PSNR, it can be observed that TetraVQM consistently performs better than PSNR.

The results show that for both models M1 and M2, the RMSE can be largely decreased and the correlation coefficients ρ_P and ρ_S can be largely increased. Somewhat unexpected, model M2 achieves better results than model M1, even though M2 does not distinguish saliency levels between the distortion regions of the different sequences, but instead uses a constant offset for SEQ$_{S,0.4}$ and SEQ$_{S,1.2}$.

It should be noted that for model M2, the maximum ρ_S coincides with the α_2 for which the TetraVQM$_{\text{M2}}$ of all sequences SEQ$_{S,0.4}$ and SEQ$_{S,1.2}$ are shifted below the TetraVQM$_{\text{M2}}$ of the sequences SEQ$_{N,0.4}$ and SEQ$_{N,1.2}$. Thus, the

Table 40: Optimised parameters α_{opt} and quality prediction performance indicators (RMSE, ρ_P, ρ_S) for TetraVQM and PSNR.

Metric	α_{opt}	RMSE	ρ_P	ρ_S
TetraVQM	N/A	0.702	0.522	0.536
TetraVQM$_{M1}$	28.15	0.447	0.84	0.835
TetraVQM$_{M2}$	2.41	0.316	0.923	0.888
PSNR	N/A	0.75	0.414	0.451
PSNR$_{M1}$	418.61	0.465	0.825	0.83
PSNR$_{M2}$	35.08	0.332	0.915	0.88

rank order correlation of the objective quality scores with MOS is highest when all sequences with distortions in the salient regions are rated lower than the worst sequence with distortions in the non-salient region. This observation is in line with the conclusions drawn from the MOS of the subjective experiment (see Section 13.1 and Fig. 75).

Scatter plots of TetraVQM$_{M1}$ and TetraVQM$_{M2}$ are presented in Fig. 87 after deploying a linear mapping to the MOS. The scatter plot for TetraVQM$_{M2}$ shows two distinct point clouds for the two classes, salient and non-salient, which partially corresponds to the situation seen in Fig. 75. This is remarkable because the optimisation has been performed on the RMSE value and not on the correlation coefficients and thus, it is not an artifact of the training.

14.4 Limitations and outlook

Although the results presented in this chapter are very promising, there are some limitations that need to be considered. Firstly, the same subjective data has been used for training as well as for evaluation of the saliency awareness models. As such, the performance indicators presented in Table 40 provide an upper bound of the performance improvement that can be expected due to the saliency awareness framework. Further analysis based on separated training and validation sets will lead to more insight regarding the performance of the proposed framework.

The amount of saliency in the presented models has been quantified using the mean over the distortion region. Different quantification models might lead to even larger performance improvements than the ones that have been presented

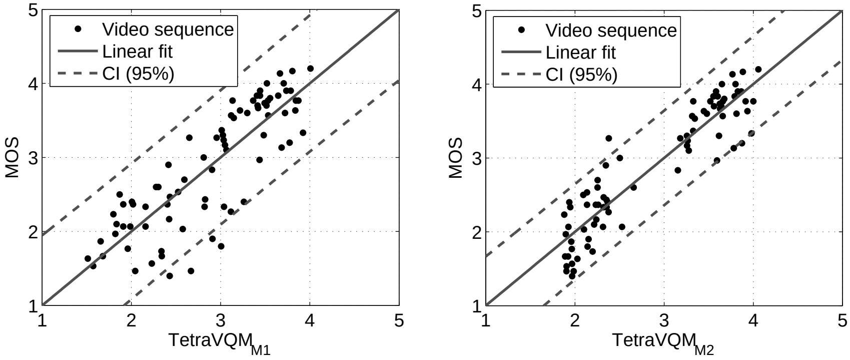


Figure 87: Scatter plots of MOS over $\text{TetraVQM}_{\text{M1}}$ and $\text{TetraVQM}_{\text{M2}}$, including 95% confidence intervals (CI).

here. Such models could, for instance, consider the variation of the saliency in the distortion region, rather than just taking the mean. The relative magnitudes of the saliency in relation to the distortion severeness could also be a factor worth looking at.

Finally, the presented framework combines the saliency model non-intrusively with the VQM. Although this has the advantage of not having to change the original metric, the interaction between the distortion measures and the HVS-based processing steps within the VQM cannot be fully integrated with the saliency model. Thus, further work can concentrate on implementing the saliency awareness directly into the algorithmic processing of the VQM, which may lead to further performance improvements.

15 Final Remarks

In the following sections, we briefly summarise the work that has been presented in this thesis and highlight the major contributions. We further discuss a few limitations that can be subject for future work. We end the thesis with some final conclusions.

15.1 Summary and contributions

In this thesis, we presented our work on modelling perceptual quality and visual saliency for image and video communication applications. The presented models were all developed and validated using as a ground truth data obtained from extensive subjective experiments. They were further designed within the constraints of image and video communication systems and in particular wireless networks. To be precise, the low computational processing power, the scarce channel bandwidth, and the complex distortion patterns were major considerations in the metric designs. The contributions of this thesis can be summarised as follows.

Part I presented reduced-reference quality metrics specifically designed for application in image and video communications. The metrics are of low computational cost and exploit minimal reference information, as compared to other reduced-reference metrics proposed in the literature. This is achieved while maintaining a superior quality prediction performance compared to other contemporary metrics in the context of image communications.

Part II proposes several techniques to further improve the quality prediction performance of the reduced-reference quality metrics. In particular, a multiobjective optimisation framework is proposed for determining optimal feature weights. The framework is applied to several quality metrics, showing consistent performance improvement while maintaining generalisation ability of the metrics. A region-of-interest framework is further proposed to be integrated into existing image quality metrics.

Part III aims at facilitating a better understanding regarding the deployed strategies of human observers when rating image quality. Firstly, the confidence of human observers when judging image quality is analysed in detail and simple computational models are derived to predict observer confidence, as a complement to typically computed confidence intervals. Secondly, the gaze patterns of human observers are analysed which were recorded during task-free and task-based (quality assessment task) eye tracking experiments, revealing valuable insight into the human viewing behaviour when presented natural image content in the absence and the presence of structural distortions.

Part IV, finally, evaluates the perceived annoyance of packet loss distortions in relation to the underlying content saliency of natural video sequences. The analysis is based on an extensive combined eye tracking and video quality assessment experiment. Based on the experiment outcomes, a saliency awareness model is further proposed to enhance existing video quality metrics by integrating content saliency information.

We further make some of the results from our subjective experiments publicly available to the research community. The outcomes of experiments E1 and E2 are made available in the Wireless Imaging Quality (WIQ) database, which is explained in Appendix A. The outcomes of experiment E3 are made available in the Region-of-Interest (ROI) database, which is explained in Appendix B. Finally, the gaze patterns recorded in experiment E4a are made available in the Visual Attention for Image Quality (VAIQ) database, which is explained in Appendix C.

15.2 Limitations and future work

The surveys presented in Chapter 1 of this thesis revealed that there is an abundant amount of different attributes and parameters that can, and maybe should, be accounted for in order to define successfully deployed perceptual quality metrics. Incorporating all factors, however, would lead to extremely complex computational metrics which would find no application in current image and video communication systems. Like any of the visual quality metrics reviewed in Section 1.4, we had to sacrifice some factors in the favour of others. To facilitate future research and possible extension of the presented models, we highlight in the following a few limitations that we are aware of:

- As most proposals in the literature thus far, the metrics presented in Part I and Part II of this thesis are based on the analysis of luminance values only. Thus, conducting similar tests on colour images and incorporating colour information into the models can be subject for future work [256, 257].

- The perceptual relevance weights of the metrics presented in Part I and Part II were derived for the particular case of JPEG compressed images. As such, the weights would need to be determined for other possible applications, for instance, for JPEG2000 coded images or H.264/AVC coded video sequences. Subjective experiments need to be conducted again to obtain a ground truth for the metric design. Incorporating several different coding systems into the weights optimisation would make the metrics more generally applicable, but likely reduce the quality prediction performance with respect to any of the codecs involved.

- The quality metrics and saliency models presented in this thesis focused on visual stimuli. In the case of video, auditory distortions also have a strong impact on the overall perceived quality. Furthermore, the attention of a human observer is strongly driven by auditory cues [170] in addition to the visual cues. Taking these considerations into account may lead to even more effective audio-visual quality metrics that correlate well with human perception [258, 259]. However, auditory cues to be incorporated into the quality and visual attention modelling were outside the scope of this thesis.

- There are still many open questions for future research regarding the gaze patterns obtained from both the image quality experiment presented in Chapter 9 and the video quality experiment reported in Chapter 12. In both cases it would be of great interest to establish closer relationships between the gaze patterns of the human observers and their quality judgements during the experiment. Such an analysis would serve to further understand the quality rating behaviour of human observers when presented natural image and video content in the presence of transmission distortions.

- The gaze patterns from experiments E4b and E5 were recorded under quality assessment task. Hence, the attention to the visual distortions does not reflect the pure saliency of the distortions in relation to the image or video content. It would therefore be very valuable to conduct a task-free eye tracking experiment in which the participants are shown images or videos containing transmission distortions. The outcomes would reflect more suitably the natural attention to these distortions, disregarding the search strategies deployed during quality assessment.

- In addition to the analysis provided in Section 11.2, there is much room for analysis regarding the task-free gaze patterns from the eye tracking experiment E4a presented in Chapter 9. As many visual attention models are designed and validated on saliency maps obtained from eye tracking, it is of crucial importance to obtain valid saliency maps that constitute a reliable ground truth. However, little is known about the convergence behaviour of saliency maps in dependence on the presentation time of the image. As such, a too short presentation time may result in an 'incomplete' saliency map whereas a too long presentation time may contain 'redundant' saliency information, thus unnecessarily prolonging expensive subjective experiments. Furthermore, the consistency of saliency maps obtained from eye tracking experiments of different laboratories should be evaluated to identify the reliability of different saliency maps as a ground truth for visual attention modelling. These are just some issues that need attention and in fact, at the time of submission of this thesis, the author has started a cooperative work with the University of Technology in Delft, The Netherlands, the University of Nantes, France, and the University of Western Sydney, Australia, to shed some more light onto these problems.

15.3 Conclusions

The field of visual quality assessment research has experienced tremendous advances over the past decades and especially in recent years. The increased deployment of perceptual quality assessment in image and video processing applications is promising for a wider acceptance of perceptual quality metrics as an alternative to PSNR. However, despite the efforts and the improvements there are no quality predictors yet that work reliable under a wide range of different scenarios. One of the most commonly neglected factors is visual attention, which has been shown in this thesis to have a considerable impact on the perceived visual quality and the prediction performance of image and video quality metrics. Colour [256] is another often disregarded attribute which, when taken into account, can be expected to further advance this field of research. In fact, Xia et al. [260] recently found that colour artifacts are amongst the most severely perceived distortions in natural video sequences and should thus not be neglected. We believe also that reduced-reference quality assessment should receive more attention as a method that allows for a well balanced compromise between full-reference and no-reference quality assessment, combining the advantages of both of them.

In any case, we are still far away from truly reliable and universally applicable visual quality metrics if, in fact, such metrics can be achieved, given the subjectivity of quality perception and the broad range of available multimedia applications nowadays.

Appendices

For the development of visual quality metrics and visual attention models, a ground truth is usually needed on which to design and validate the models on. In case of quality metrics, mean opinion scores obtained in subjective quality experiments are typically utilised. On the other hand, gaze patterns from eye tracking experiments often support the design and validation of visual attention models. For a better comparison of the objective methods between laboratories worldwide, it is desirable to have publicly available quality and eye tracking databases.

For this reason, we make some of our subjective databases publicly available to the research community. In particular, the outcomes of experiments E1 and E2 are made available in the **Wireless Imaging Quality (WIQ)** database, which is explained in more detail in Appendix A. The **Region-Of-Interest (ROI)** database contains the outcomes of experiment E3 and is discussed in Appendix B. Finally, the eye tracking data from experiment E4a is made available in the **Visual Attention for Image Quality (VAIQ)** database, which is introduced in Appendix C.

All three databases can be downloaded from the following web site:

http://www.bth.se/tek/rcg.nsf/pages/perceptual-databases

The passwords needed to unpack the files can be obtained by emailing the author of this thesis, Ulrich Engelke (ulrichengelke@gmail.com).

Appendix D presents the saliency maps created from the gaze patterns from eye tracking experiment E4b. The saliency maps are visualised as heat maps, overlayed on the image content.

A The Wireless Imaging Quality (WIQ) Database

A.1 Database description

The Wireless Imaging Quality (WIQ) database is based on the outcomes of experiments E1 and E2, which are described in detail in Chapter 2. The WIQ database contains the following data:

- Reference images from experiments E1 and E2
  (wiq_ref_images.zip, approx. 1.55 MB)
- Distorted images from experiment E1
  (wiq_dst_images_t01.zip, approx. 8.88 MB)
- Distorted images from experiment E2
  (wiq_dst_images_t02.zip, approx. 8.72 MB)
- Raw subjective image quality scores and mean opinion scores contained in a Matlab workspace (wiq_subjective_scores_matlab.zip, approx. 7 kB)
- Raw subjective image quality scores and mean opinion scores contained in an Excel spreadsheet (wiq_subjective_scores_excel.zip, approx. 45 kB)
- WIQ database readme file (wiq_readme.zip, approx. 3 kB)

A.2 Chief investigators

The chief investigators are:

- Ulrich Engelke, Blekinge Institute of Technology, Sweden
- Tubagus Maulana Kusuma, Gunadarma University, Indonesia
- Hans-Jürgen Zepernick, Blekinge Institute of Technology, Sweden

A.3 References for the WIQ database

In addition to the discussion in this thesis, the following publications contain detailed descriptions and analysis of the WIQ database:

U. Engelke, T. M. Kusuma, H.-J. Zepernick, and M. Caldera "Reduced-Reference Metric Design for Objective Perceptual Quality Assessment in Wireless Imaging," *Signal Processing: Image Communication*, vol. 24, no. 7, pp. 525-547, 2009.

U. Engelke, H.-J. Zepernick, and T. M. Kusuma “Subjective Quality Assessment for Wireless Image Communication: The Wireless Imaging Quality Database,” *in Proc. of International Workshop on Video Processing and Quality Metrics (VPQM)*, Scottsdale, USA, January 2010.

B The Region-of-Interest (ROI) Database

B.1 Database description

The Region-of-Interest (ROI) database is based on the outcomes of experiment E3, which is described in detail in Section 8.1. The ROI database contains the following data:

- ROI coordinates contained in a Matlab workspace (roi_coordinates_matlab.zip, approx. 3 kB)
- ROI coordinates contained in an Excel spreadsheet (roi_coordinates_excel.zip, approx. 17 kB)
- ROI database readme file (roi_readme.zip, approx. 2 kB)

B.2 Chief investigators

The chief investigators are:

- Ulrich Engelke, Blekinge Institute of Technology, Sweden
- Hans-Jürgen Zepernick, Blekinge Institute of Technology, Sweden

B.3 References for the ROI database

In addition to the discussion in this thesis, the following publications contain detailed descriptions and analysis of the ROI database:

U. Engelke and H.-J. Zepernick "A Framework for Optimal Region-of-Interest Based Quality Assessment in Wireless Imaging," *Journal of Electronic Imaging, Special Section on Image Quality*, vol. 19, no.1 , ID 011005, 2010.

U. Engelke and H.-J. Zepernick, "Optimal Region-of-Interest Based Visual Quality Assessment," *in Proc. of IS&T/SPIE Human Vision and Electronic Imaging XIV*, vol. 7240, San Jose, USA, January 2009.

B.4 Coordinates of all ROI selections

The coordinates of all ROI selections are presented in Tables 41 and 42, with the coordinate system origin being in the top left image corner.

Table 41: Coordinates of the ROI selections for the seven reference images as obtained from experiment E3 (participants 1-15).

#	Coord	Barbara		Elaine		Goldhill		Lena		Mandrill		Peppers		Tiffany	
1	x*	338	421	185	301	240	291	240	367	120	398	51	206	216	441
	y#	75	178	187	319	390	467	228	306	96	348	35	494	145	408
2	x	414	505	184	250	244	301	440	512	116	399	225	385	271	388
	y	91	227	265	324	387	463	251	512	26	96	160	331	291	382
3	x	319	501	187	304	251	481	241	373	112	373	186	431	220	419
	y	1	227	175	323	149	346	244	297	27	471	69	380	172	382
4	x	179	469	158	385	1	507	192	387	106	425	45	219	232	473
	y	2	206	151	366	27	250	220	405	17	137	45	505	135	422
5	x	327	424	179	314	275	355	230	358	110	409	188	424	207	437
	y	74	116	182	319	205	279	243	382	22	442	223	476	188	382
6	x	279	509	48	497	5	450	58	442	70	446	166	437	199	437
	y	1	205	3	510	60	476	32	475	19	120	178	490	126	405
7	x	229	496	240	507	337	491	191	373	108	411	239	382	189	438
	y	6	262	273	482	115	330	174	397	17	121	157	329	155	261
8	x	208	512	126	393	241	465	72	421	112	413	61	440	197	433
	y	1	512	90	373	153	489	42	512	16	491	10	494	138	255
9	x	296	512	143	390	61	478	223	367	122	390	73	425	216	444
	y	5	224	131	356	238	478	230	306	91	376	26	479	101	423
10	x	321	460	168	393	55	386	129	419	89	444	108	418	60	456
	y	32	247	153	379	175	456	133	461	11	484	13	475	53	464
11	x	317	471	8	175	256	360	222	364	123	411	86	195	197	449
	y	7	187	271	500	156	308	191	390	24	129	28	487	35	442
12	x	1	183	159	367	364	512	65	201	96	399	153	455	164	472
	y	88	299	173	366	338	466	170	329	1	146	164	376	129	303
13	x	259	498	199	315	106	272	204	423	101	446	21	210	208	453
	y	4	229	191	250	129	452	163	412	18	330	43	485	148	448
14	x	230	504	152	372	19	290	47	311	102	409	81	241	85	236
	y	269	508	166	352	120	475	171	498	14	106	31	287	345	488
15	x	310	496	194	322	89	464	238	368	109	418	208	437	249	418
	y	81	370	187	240	99	275	32	493	21	487	173	482	286	391

* The left/right x-coordinates for each image denote the respective left/right ROI limits.

# The left/right y-coordinates for each image denote the respective upper/lower ROI limits.

Table 42: Coordinates of the ROI selections for the seven reference images as obtained from experiment E3 (participants 16-30).

#	Coord	Barbara		Elaine		Goldhill		Lena		Mandrill		Peppers		Tiffany	
16	x*	291	472	143	421	250	291	201	361	100	392	86	190	205	436
	y#	4	217	86	379	384	477	166	399	29	104	90	466	163	248
17	x	421	466	178	331	173	263	236	367	111	417	118	393	233	414
	y	388	465	193	237	277	434	230	296	6	128	35	299	295	381
18	x	288	502	96	350	172	397	195	383	115	402	73	202	199	438
	y	4	280	104	347	253	444	184	401	30	468	51	479	151	403
19	x	186	512	87	512	199	512	60	446	115	402	91	428	165	487
	y	1	512	28	512	181	512	32	475	30	468	10	328	31	456
20	x	17	159	201	334	233	368	230	398	52	448	70	227	222	436
	y	112	253	262	318	372	471	219	304	15	496	16	224	144	380
21	x	242	512	75	456	1	460	96	431	99	400	66	426	158	486
	y	4	358	92	376	109	467	39	426	24	436	46	486	6	450
22	x	328	440	169	308	260	449	69	214	155	329	85	202	253	386
	y	9	181	186	237	201	321	213	512	94	443	43	472	306	376
23	x	7	315	67	452	272	458	55	401	62	441	61	252	194	442
	y	133	512	59	385	165	446	90	512	1	193	10	512	85	442
24	x	209	512	121	498	1	444	67	398	80	401	175	420	65	512
	y	1	512	1	512	56	449	50	512	1	500	177	495	30	512
25	x	325	477	157	371	146	300	219	379	103	409	65	412	209	439
	y	1	202	117	346	166	462	188	393	19	484	1	478	82	446
26	x	333	476	65	452	178	402	245	360	134	395	423	496	281	373
	y	21	200	18	249	282	428	245	284	117	338	142	272	310	369
27	x	15	174	158	402	237	307	192	398	128	390	71	188	272	386
	y	111	281	109	355	369	485	150	404	34	108	36	512	302	386
28	x	333	449	185	320	209	406	233	372	88	423	232	361	259	394
	y	2	186	186	239	49	167	201	394	17	119	160	302	296	378
29	x	330	427	188	305	266	420	237	372	135	357	66	404	210	441
	y	19	178	182	238	166	283	241	294	71	440	9	223	180	245
30	x	317	442	172	361	250	286	186	410	202	295	210	392	268	379
	y	68	118	198	308	403	456	236	293	124	335	179	485	298	384

* The left/right x-coordinates for each image denote the respective left/right ROI limits.

# The left/right y-coordinates for each image denote the respective upper/lower ROI limits.

C The Visual Attention for Image Quality (VAIQ) Database

C.1 Database description

The Visual Attention for Image Quality (VAIQ) database is based on the outcomes of experiment E4a, which is described in detail in Chapter 9. The VAIQ database contains the following data:

- Recorded gaze points contained in Excel spreadsheets (vaiq_gaze_points_excel.zip, approx. 16.9 MB)
- Recorded gaze points contained in a Matlab workspace (vaiq_gaze_points_matlab.zip, approx. 3.6 MB)
- Saliency maps Set 1 (vaiq_saliency_maps_1.zip, approx. 32.4 MB)
- Saliency maps Set 2 (vaiq_saliency_maps_2.zip, approx. 34.2 MB)
- Saliency maps Set 3 (vaiq_saliency_maps_3.zip, approx. 36.5 MB)
- Saliency images (vaiq_saliency_images.zip, approx. 13.6 MB)
- VAIQ readme file (vaiq_readme.zip, approx. 4 kB)

C.2 Chief investigators

The chief investigators are:

- Ulrich Engelke, Blekinge Institute of Technology, Sweden
- Anthony Maeder, University of Western Sydney, Australia
- Hans-Jürgen Zepernick, Blekinge Institute of Technology, Sweden

C.3 References for the VAIQ database

In addition to the discussion in this thesis, the following publications contain detailed descriptions and analysis of the VAIQ database:

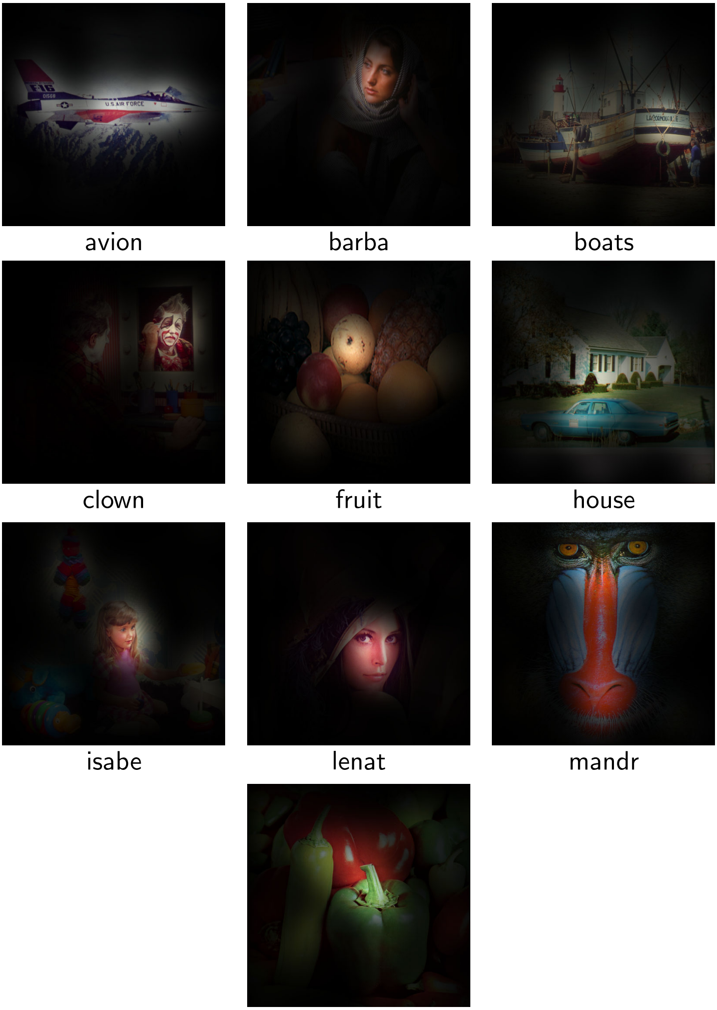
U. Engelke, A. J. Maeder, and H.-J. Zepernick, "Analysing Inter-Observer Saliency Variations in Task-Free Viewing of Natural Images," *in Proc. of IEEE International*

Conference on Image Processing (ICIP), Hong Kong, China, September 2010.

U. Engelke, A. J. Maeder, and H.-J. Zepernick, "Visual Attention Modelling for Subjective Image Quality Databases," *in Proc. of IEEE International Workshop on Multimedia Signal Processing (MMSP)*, Rio de Janeiro, Brazil, October 2009.

C.4 Saliency maps

The saliency maps for all images in the VAIQ database are visualised in Fig. 88 to Fig. 92. The saliency maps are divided into the different figures with respect to the databases that the corresponding images are from; IRCCyN/IVC [36], LIVE [37], and MICT [35]. The names under each image correspond to the original name from the respective database. More information on the image quality databases can be found in Table 24.



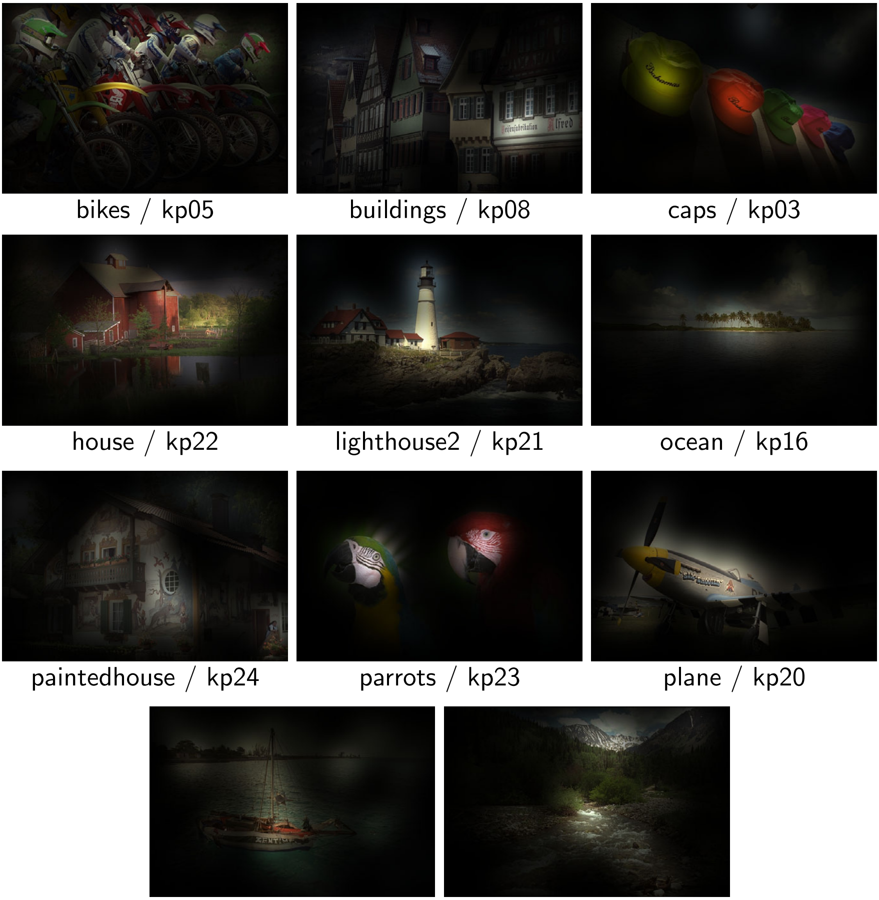
avion barba boats

clown fruit house

isabe lenat mandr

pimen

Figure 88: Saliency maps for the images from the IRCCyN/IVC [36] database.



bikes / kp05

buildings / kp08

caps / kp03

house / kp22

lighthouse2 / kp21

ocean / kp16

paintedhouse / kp24

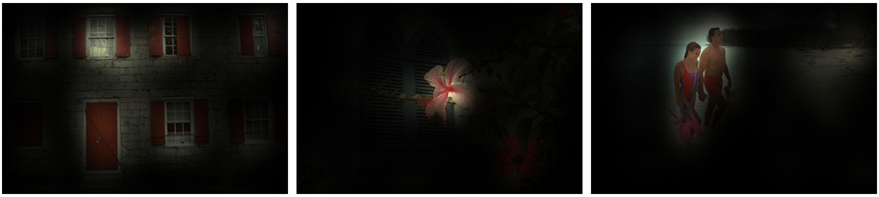
parrots / kp23

plane / kp20

sailing1 / kp06

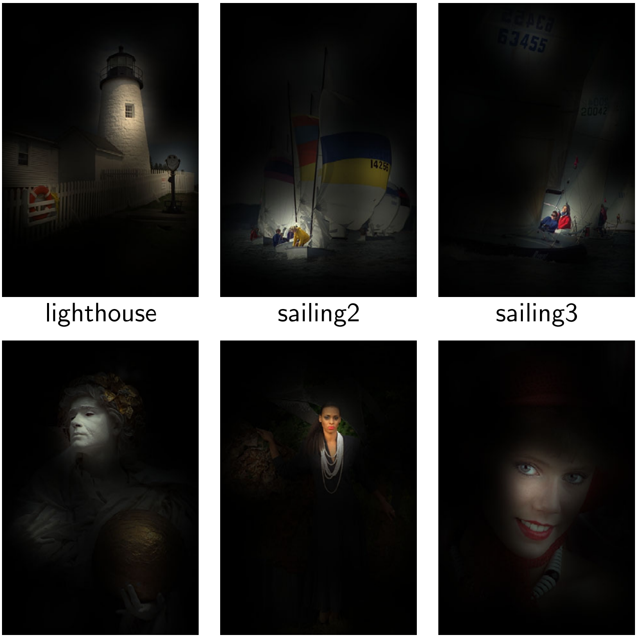
stream / kp13

Figure 89: Saliency maps for the images contained in both the LIVE [37] and the MICT [35] database (left label: LIVE; right label: MICT).



kp01 kp07 kp12

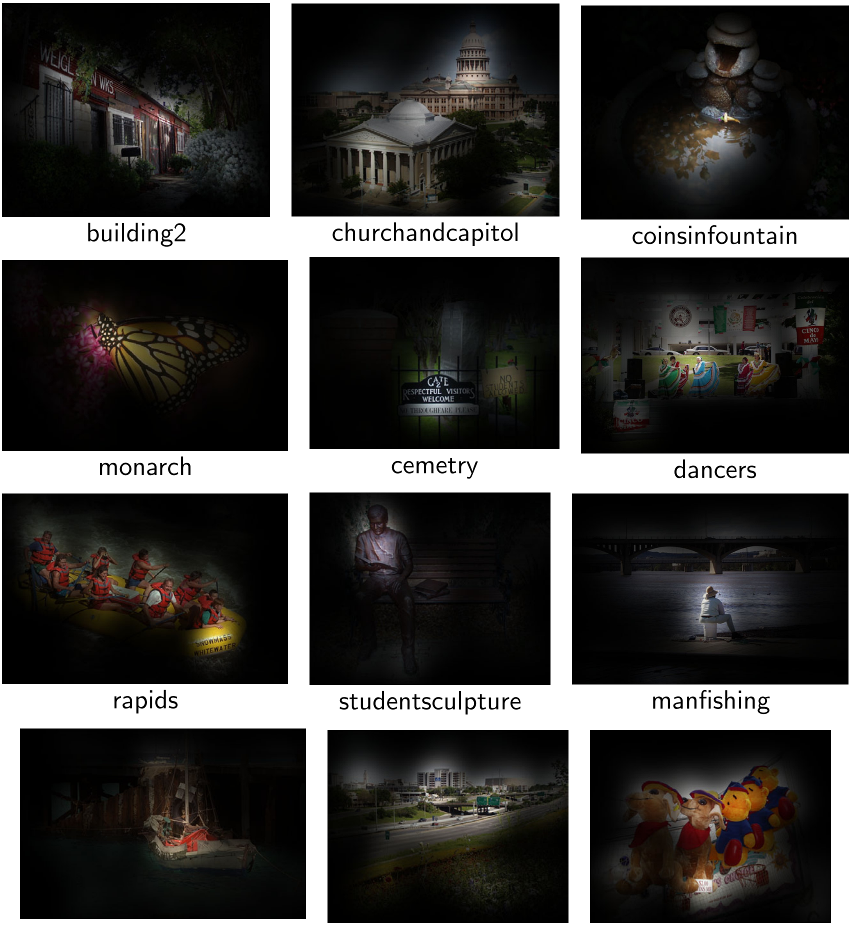
Figure 90: Saliency maps for the images exclusively from the MICT [35] database.



lighthouse sailing2 sailing3

statue woman womanhat

Figure 91: Saliency maps for the images exclusively from the LIVE [37] database.



building2 churchandcapitol coinsinfountain

monarch cemetry dancers

rapids studentsculpture manfishing

sailing4 flowersonih35 carnivaldolls

Figure 92: Saliency maps for the images exclusively from the LIVE [37] database (continued).

D Heat Maps for All Images from Experiment E4b

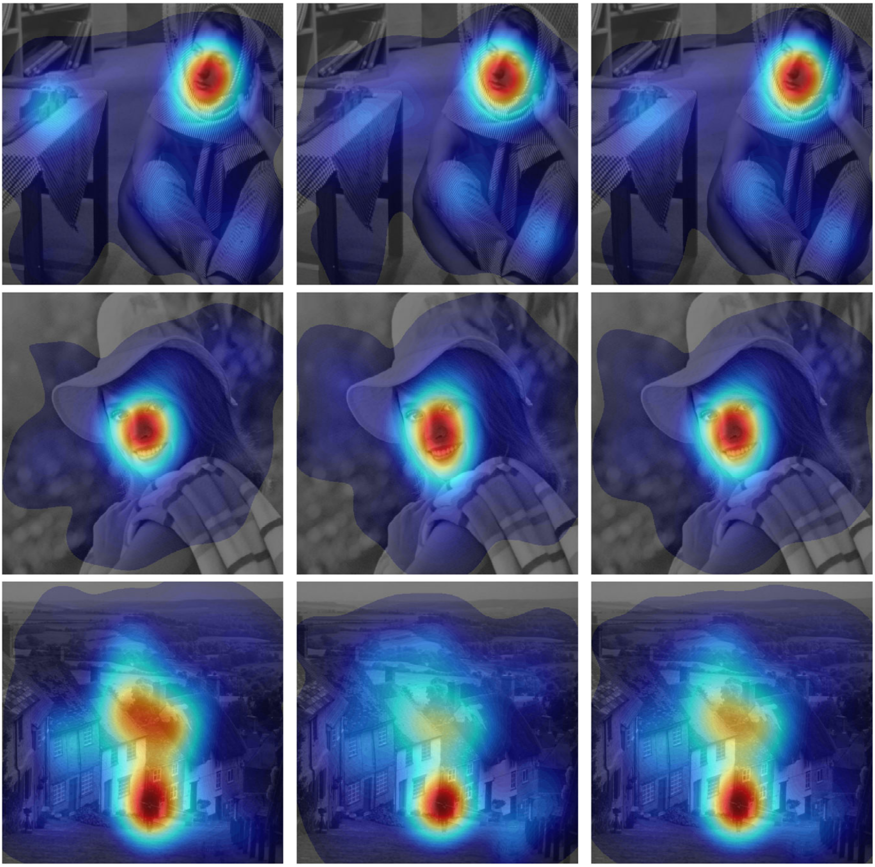


Figure 93: Heat maps for the reference images, based on gaze patterns from the first session (left), the second session (middle), and both sessions (right).

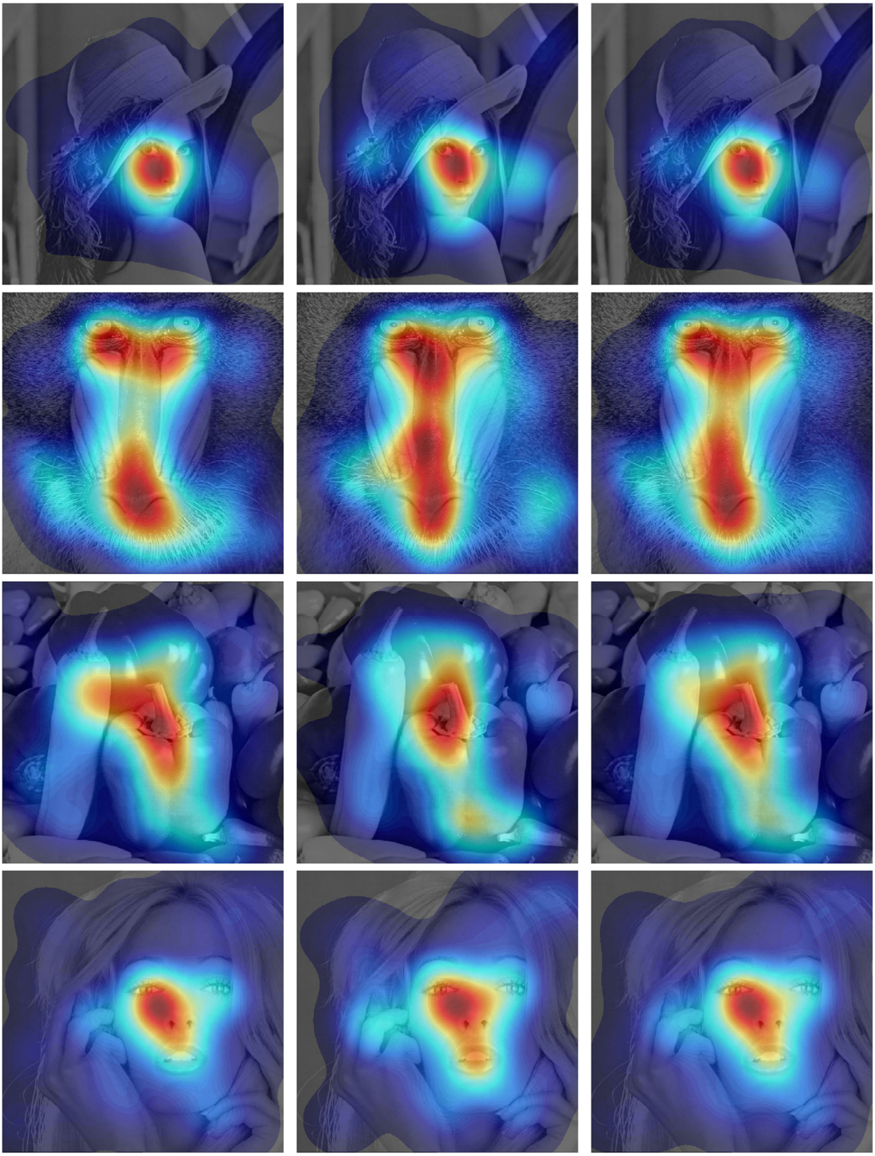


Figure 94: Heat maps for the reference images, based on gaze patterns from the first session (left), the second session (middle), and both sessions (right).

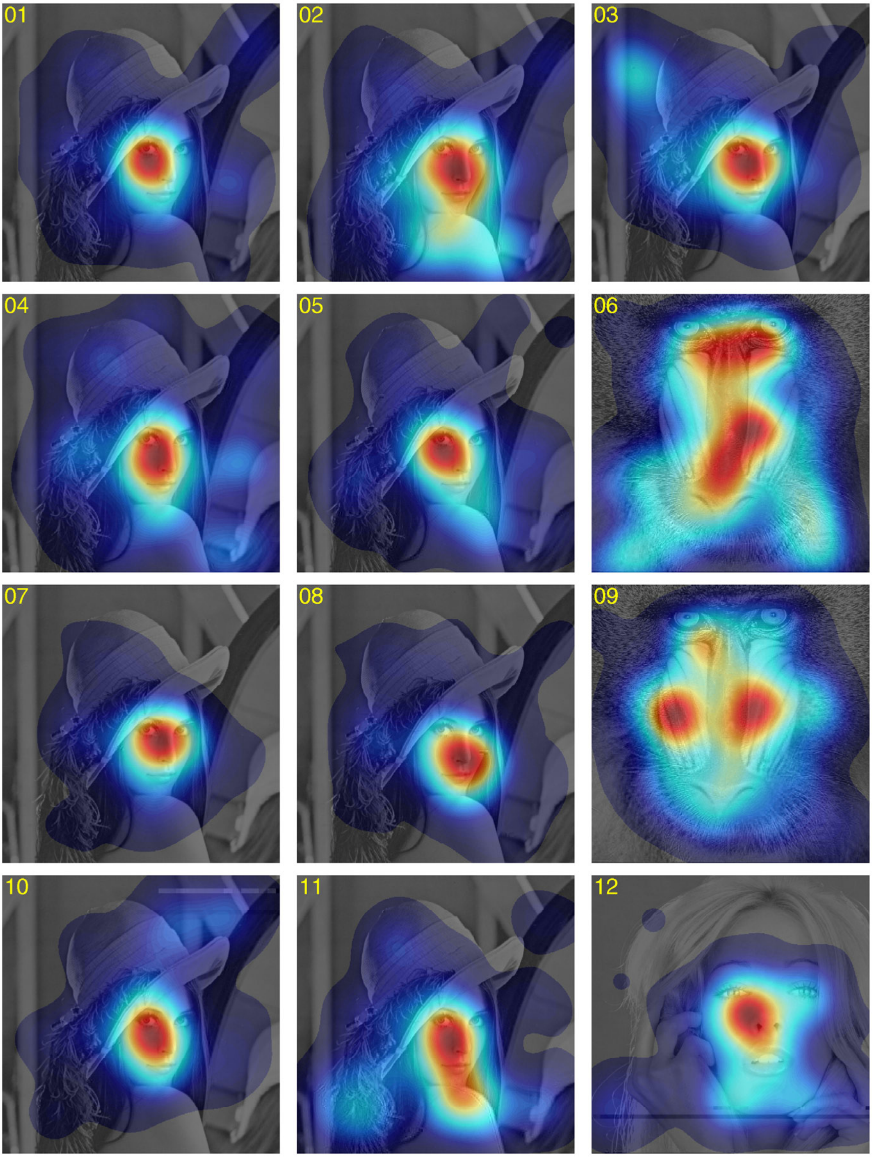


Figure 95: Heat maps for the distorted images 1-12.

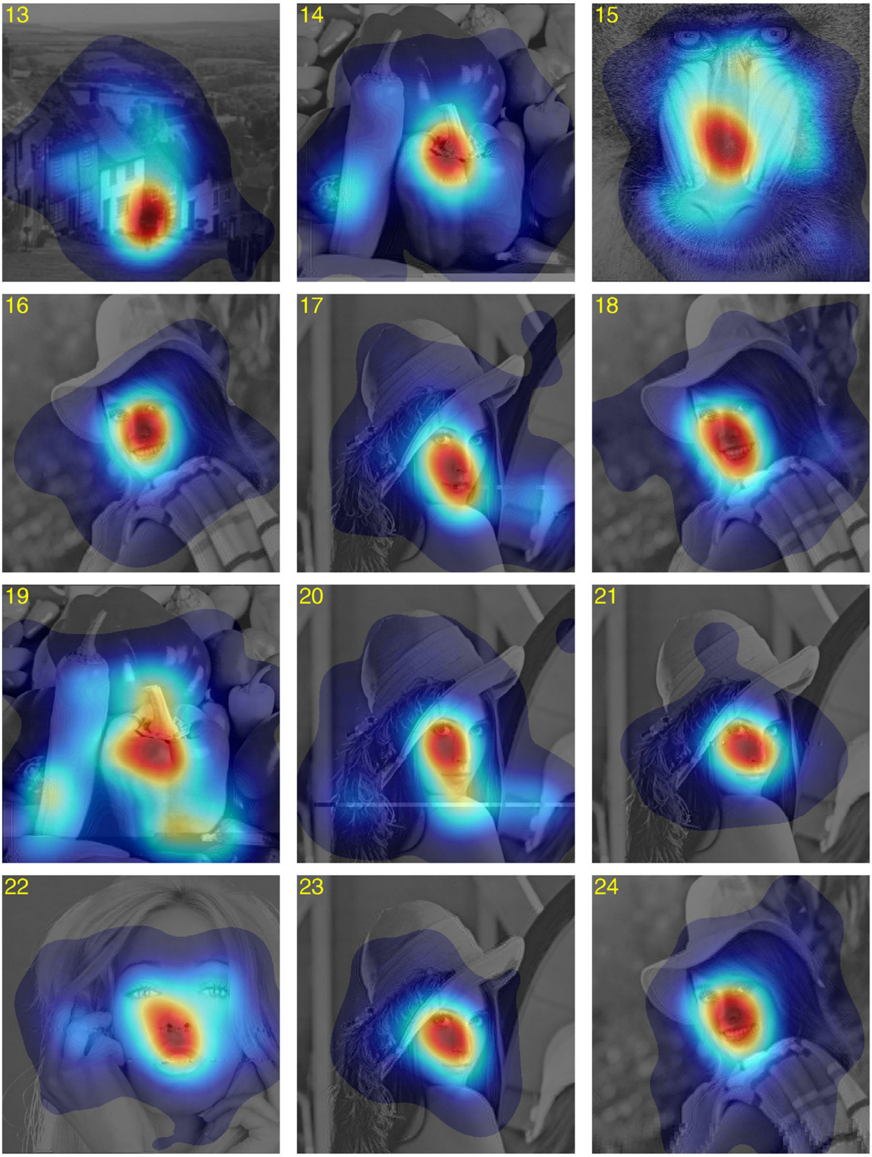


Figure 96: Heat maps for the distorted images 13-24.

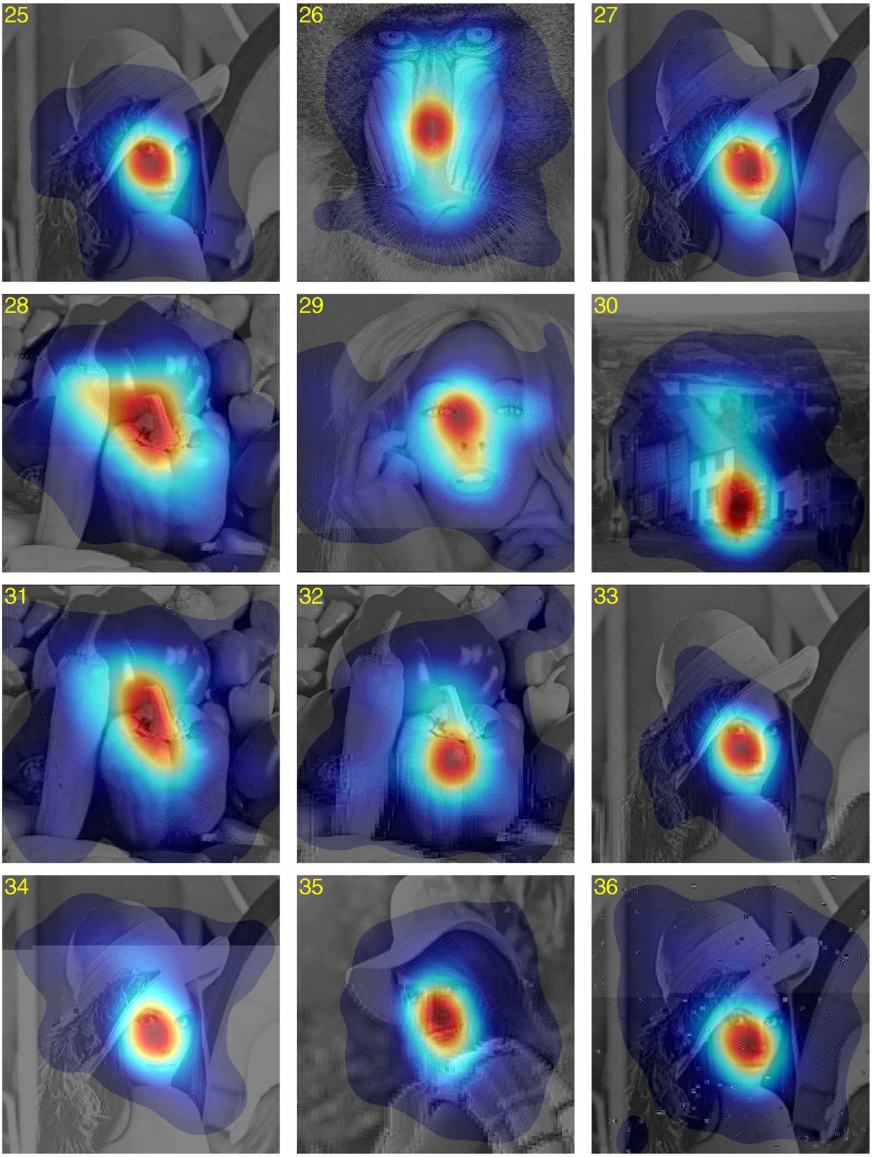


Figure 97: Heat maps for the distorted images 25-36.

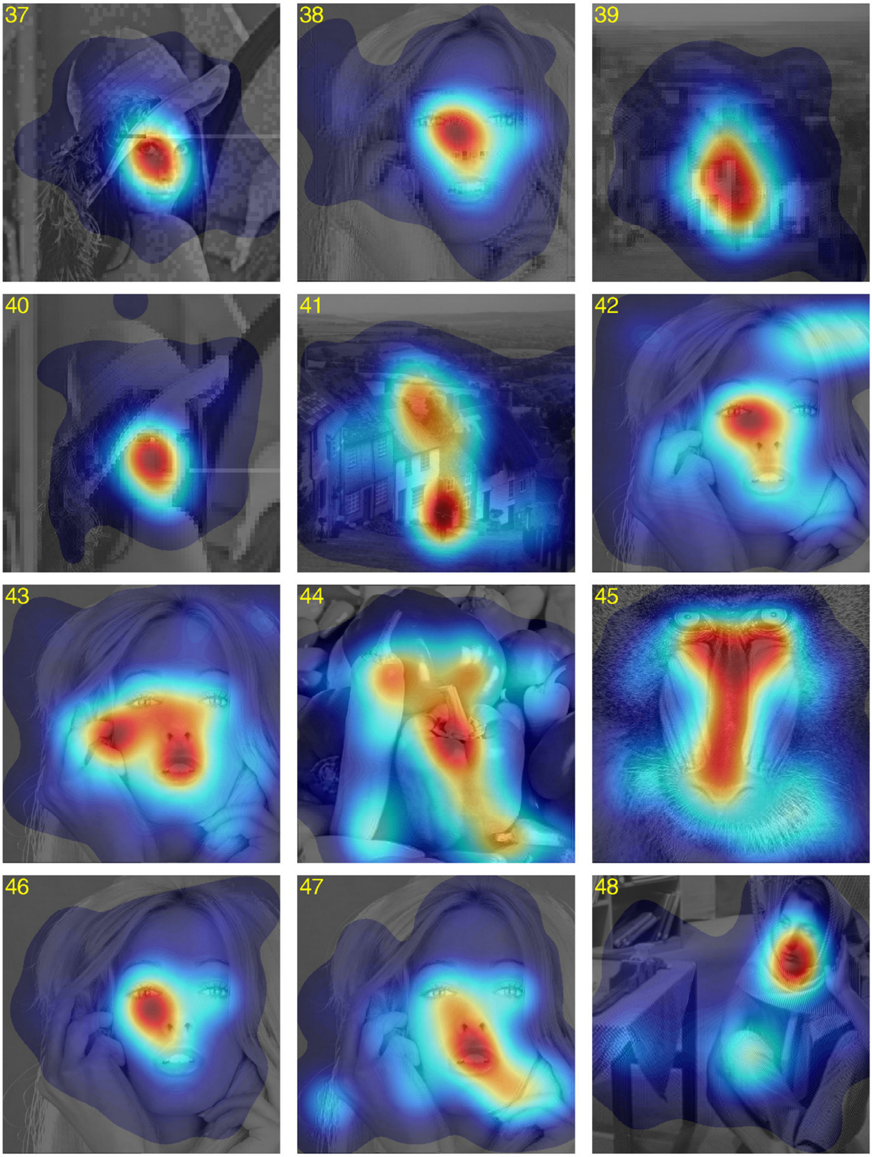


Figure 98: Heat maps for the distorted images 37-48.

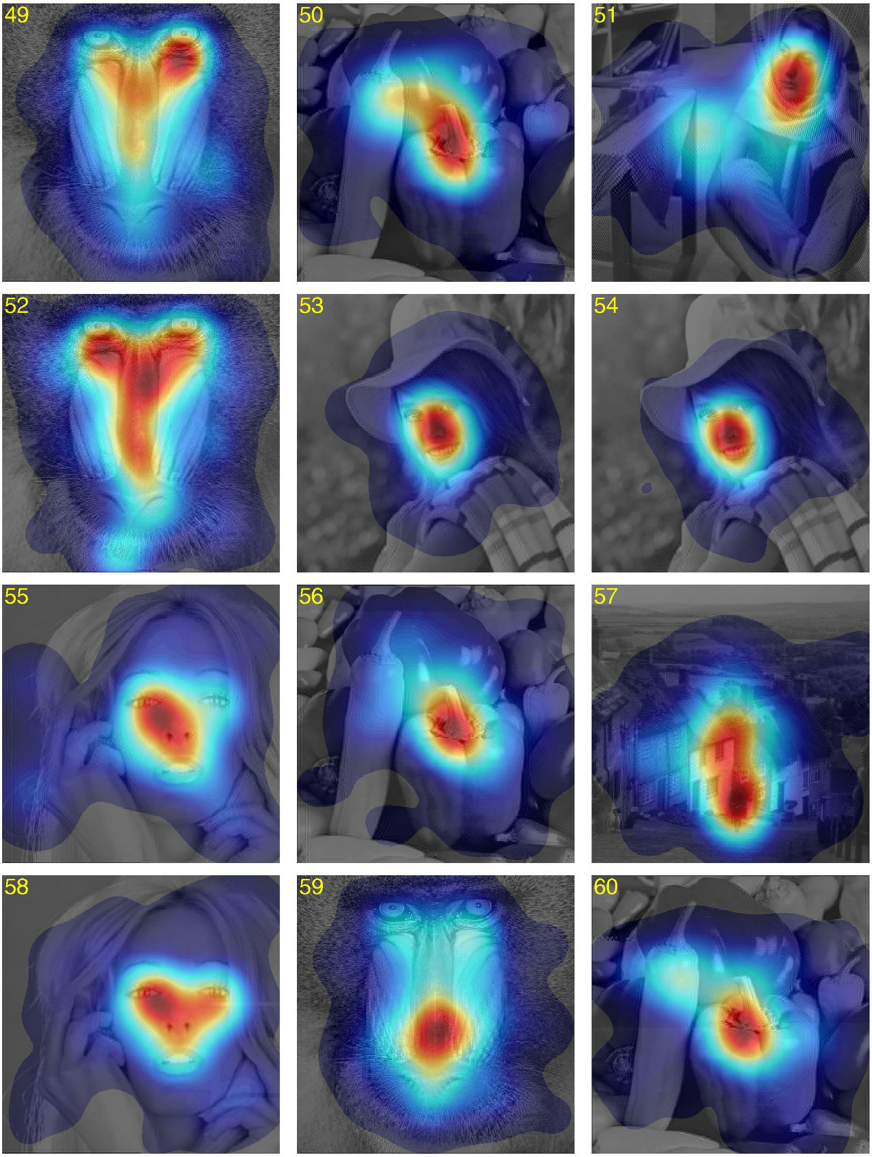


Figure 99: Heat maps for the distorted images 49-60.

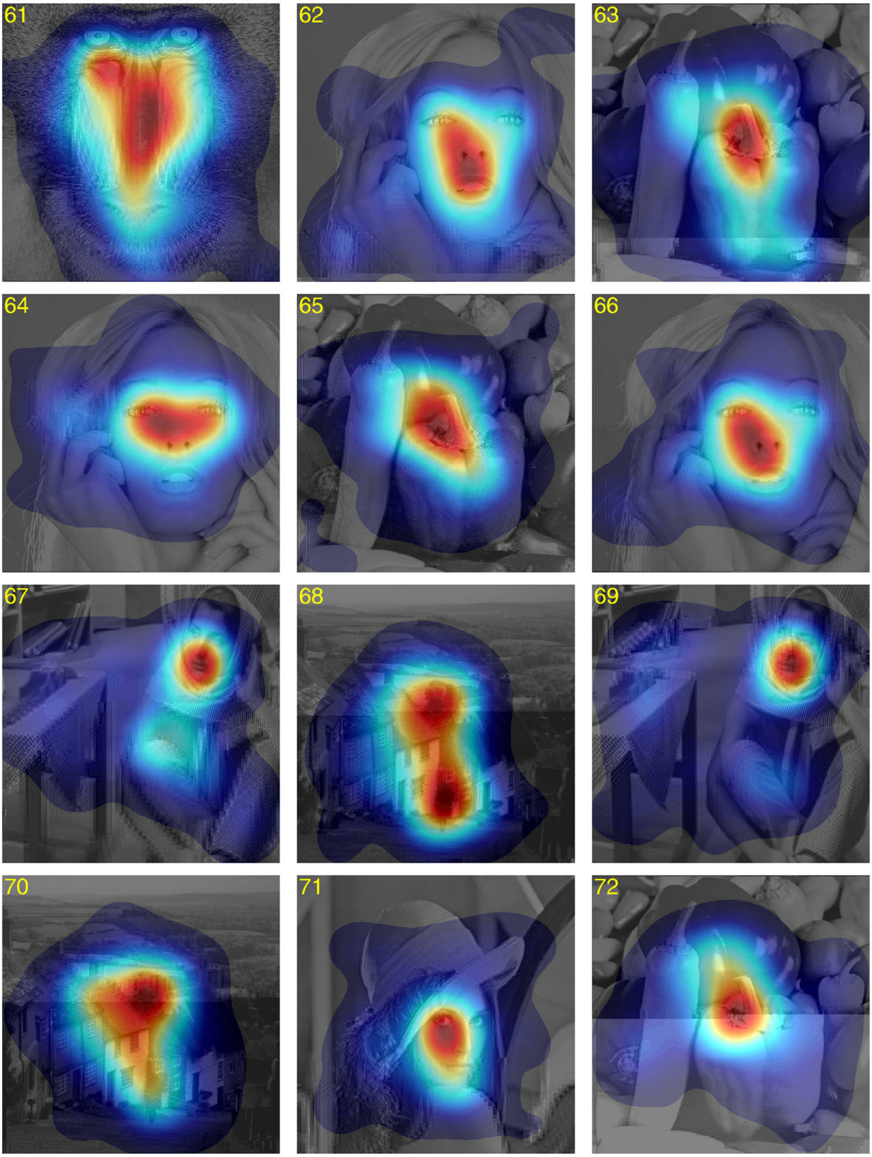


Figure 100: Heat maps for the distorted images 61-72.

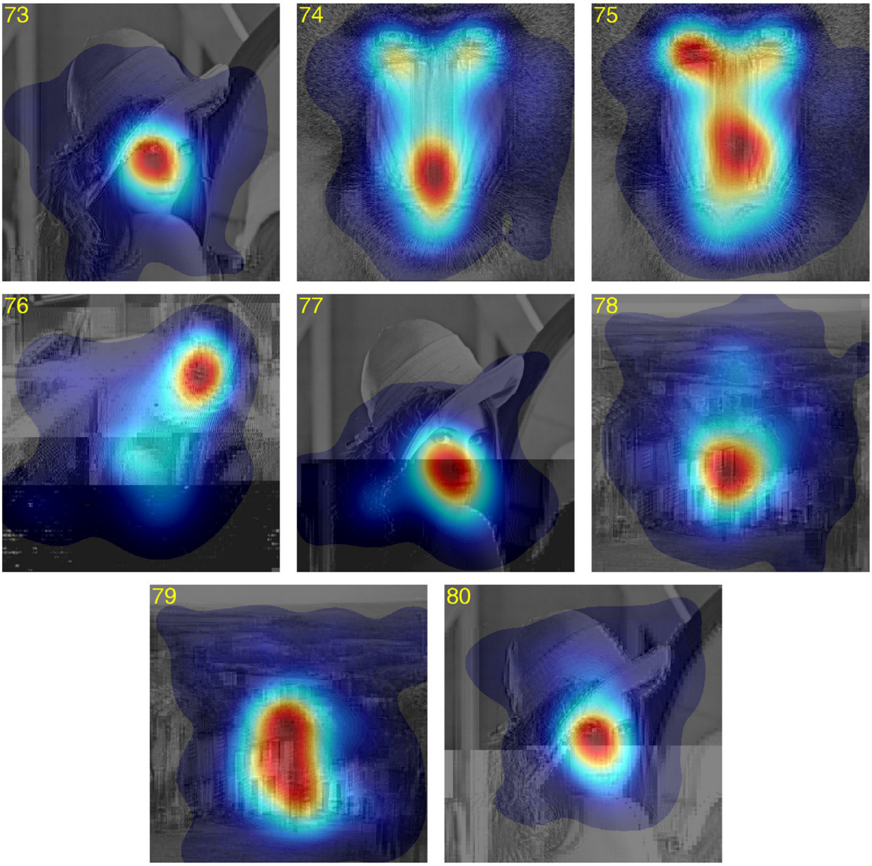


Figure 101: Heat maps for the distorted images 73-80.

References

[1] B. A. Wandell, *Foundations of Vision*. Sinauer Associates, Inc., 1995.

[2] A. C. Bovik (Ed.), *Handbook of Image and Video Processing*, 2nd ed. Academic Press, 2005.

[3] S. Winkler and F. Dufaux, "Video quality evaluation for mobile applications," in *Proc. of IS&T/SPIE Visual Communication and Image Processing*, vol. 5150, Jul. 2003, pp. 593–603.

[4] F. Pereira, "Sensations, perceptions and emotions towards quality of experience evaluation for consumer electronics video adaptations," in *Proc. of Int. Workshop on Video Processing and Quality Metrics for Consumer Electronics*, Jan. 2005.

[5] D. Soldani, M. Li and R. Cuny (Ed.), *QoS and QoE Management in UMTS Cellular Systems*. John Wiley & Sons, 2006.

[6] S. Buchinger, S. Kriglstein, and H. Hlavacs, "Comprehensive view on user studies: Survey and open issues for mobile tv," in *Proc. of ACM European Conf. on Changing Television Environments*, Jun. 2009.

[7] H. R. Wu and K. R. Rao (Ed.), *Digital Video Image Quality and Perceptual Coding*. CRC Press, 2006.

[8] A. W. Rix, A. Bourret, and M. P. Hollier, "Models of human perception," *Journal of BT Technology*, vol. 17, no. 1, pp. 24–34, Jan. 1999.

[9] S. Winkler, "Visual fidelity and perceived quality: Towards comprehensive metrics," in *Proc. of IS&T/SPIE Human Vision and Electronic Imaging*, vol. 4299, San Jose, CA, Jan. 2001, pp. 114–125.

[10] Z. Wang, A. C. Bovik, and L. Lu, "Why is image quality assessment so difficult?" in *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 4, May 2002, pp. 3313–3316.

[11] International Telecommunication Union, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow band telephone networks and speech codecs," ITU-T, Rec. P.862, Feb. 2001.

[12] ——, "Method for objective measurements of perceived audio quality," ITU-R, Rec. BS.1387-1, Dec. 2001.

[13] Q. Huynh-Thu and M. Ghanbari, "Scope of validity of PSNR in image/video quality assessment," *IET Electronics Letters*, vol. 44, no. 13, pp. 800–801, Jun. 2008.

[14] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. on Image Processing*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[15] D. M. Chandler, K. H. Lim, and S. S. Hemami, "Effects of spatial correlations and global precedence on the visual fidelity of distorted images," in *Proc. of IS&T/SPIE Human Vision and Electronic Imaging*, vol. 6057, Jan. 2006, pp. 131–145.

[16] U. Engelke, T. M. Kusuma, H.-J. Zepernick, and M. Caldera, "Reduced-reference metric design for objective perceptual quality assessment in wireless imaging," *Signal Processing: Image Communication*, vol. 24, no. 7, pp. 525–547, Jul. 2009.

[17] J. B. Martens and L. Meesters, "Image dissimilarity," *Signal Processing*, vol. 70, no. 3, pp. 155–176, Nov. 1998.

[18] S. Winkler, *Digital Video Quality - Vision Models and Metrics*. John Wiley & Sons, 2005.

[19] International Telecommunication Union, "Methodology for the subjective assessment of the quality of television pictures," ITU-R, Rec. BT.500-11, 2002.

[20] ——, "Subjective video quality assessment methods for multimedia applications," ITU-T, Rec. P.910, Sep. 1999.

[21] P. G. J. Barten, *Contrast Sensitivity of the Human Eye and Its Effects on Image Quality*. SPIE Press, 1999.

[22] Z. Yu, H. R. Wu, and T. Ferguson, "The influence of viewing distance on subjective impairment assessment," *IEEE Trans. on Broadcasting*, vol. 48, no. 4, pp. 331–336, Dec. 2002.

[23] M. Barkowsky, B. Eskofier, J. Bialkowski, and A. Kaup, "Influence of the presentation time on subjective votings of coded still images," in *Proc. of IEEE Int. Conf. on Image Processing*, Oct. 2006, pp. 429–432.

[24] S. H. Bae, T. N. Pappas, and B. H. Juang, "Subjective evaluation of spatial resolution and quantization noise tradeoffs," *IEEE Trans. on Image Processing*, vol. 18, no. 3, pp. 495–508, Mar. 2009.

[25] P. Brun, G. Hauske, and T. Stockhammer, "Subjective assessment of H.264/AVC video for low-bitrate multimedia messaging services," in *Proc. of IEEE Int. Conf. on Image Processing*, Oct. 2004, pp. 1145–1148.

[26] F. D. Simone, M. Naccari, M. Tagliasacchi, F. Dufaux, S. Tubaro, and T. Ebrahimi, "Subjective assessment of H.264/AVC video sequences transmitted over a noisy channel," in *Proc. of Int. Workshop on Quality of Multimedia Experience*, Jul. 2009.

[27] M. H. Pinson, S. Wolf, and G. Cermak, "HDTV subjective quality of H.264 vs. MPEG-2, with and without packet loss," *IEEE Trans. on Broadcasting*, vol. 56, no. 1, pp. 86–91, Mar. 2010.

[28] G. Zhai, J. Cai, W. Lin, X. Yang, W. Zhang, and M. Etoh, "Cross-dimensional perceptual quality assessment for low bitrate videos," *IEEE Trans. on Multimedia*, vol. 10, no. 7, pp. 1316–1324, Nov. 2008.

[29] M. C. Q. Farias, J. M. Foley, and S. K. Mitra, "Detectability and annoyance of synthetic blocky, blurry, noisy, and ringing artifacts," *IEEE Trans. on Signal Processing*, vol. 55, no. 6, pp. 2954–2964, Jun. 2007.

[30] S. Winkler and S. Süsstrunk, "Visibility of noise in natural images," in *Proc. of IS&T/SPIE Human Vision and Electronic Imaging IX*, vol. 5292, Jan. 2004, pp. 121–129.

[31] R. R. Pastrana-Vidal, J. C. Gicquel, J. L. Blin, and H. Cherifi, "Predicting subjective video quality from separated spatial and temporal assessment," in *Proc. of IS&T/SPIE Human Vision and Electronic Imaging XI*, vol. 6057, Feb. 2006.

[32] Q. Huynh-Thu and M. Ghanbari, "Temporal aspect of perceived quality in mobile video broadcasting," *IEEE Trans. on Broadcasting*, vol. 54, no. 3, pp. 641–651, Sep. 2008.

[33] T. Liu, Y. Wang, J. M. Boyce, H. Yang, and Z. Wu, "A novel video quality metric for low bit-rate video considering both coding and packet-loss artifacts," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 2, pp. 280–293, Apr. 2009.

[34] G. W. Cermak, "Consumer opinions about frequency of artifacts in digital video," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 2, pp. 336–343, Apr. 2009.

[35] Z. M. P. Sazzad, Y. Kawayoke, , and Y. Horita, "Image quality evaluation database," http://mict.eng.u-toyama.ac.jp/mict/index2.html, 2000.

[36] P. Le Callet and F. Autrusseau, "Subjective quality assessment IRC-CyN/IVC database," http://www.irccyn.ec-nantes.fr/ivcdb/, 2005.

[37] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik, "LIVE image quality assessment database release 2," http://live.ece.utexas.edu/research/quality, 2005.

[38] N. Ponomarenko, M. Carli, V. Lukin, K. Egiazarian, J. Astola, and F. Battisti, "Tampere image database 2008 TID2008, version 1.0," www.ponomarenko.info/tid2008.htm, 2008.

[39] D. M. Chandler, "Categorical subjective image quality database," http://vision.okstate.edu/csiq/, 2010.

[40] U. Engelke, H.-J. Zepernick, and T. M. Kusuma, "Subjective quality assessment for wireless image communication: The wireless imaging quality database," in *Proc. of Int. Workshop on Video Processing and Quality Metrics*, Jan. 2010.

[41] Video Quality Experts Group, "Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment," VQEG, Mar. 2000.

[42] T. Liu and Y. Wang, "Perceptual video quality in presence of packet loss," http://vision.poly.edu/index.html/index.php?n=HomePage.PerceptualVideoQualityInPresenceOfPacketLoss, 2009.

[43] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "LIVE video quality database," http://live.ece.utexas.edu/research/quality/live_video.html, 2009.

[44] A. K. Moorthy, K. Seshadrinathan, R. Soundararajan, and A. C. Bovik, "LIVE wireless video quality assessment database," http://live.ece.utexas.edu/research/quality/live_wireless_video.html, 2009.

[45] F. De Simone and T. Ebrahimi, "EPFL-PoliMI video quality assessment database," http://vqa.como.polimi.it/index.htm, 2009.

[46] L. Goldmann, F. De Simone, and T. Ebrahimi, "3D video quality assessment," http://mmspl.epfl.ch/page38842.html, 2010.

[47] Z. Wang and A. C. Bovik, *Modern Image Quality Assessment*. Morgan & Claypool, 2006.

[48] C. Zetzsche and G. Hauske, "Principal features of human vision in the context of image quality models," in *Proc. of Int. Conf. on Image Processing and Its Applications*, Jul. 1989, pp. 102–106.

[49] T. N. Pappas, R. J. Safranek, and J. Chen, "Perceptual criteria for image quality evaluation," in *Handbook of Image and Video Processing*, A. C. Bovik, Ed. Academic Press, 2005, ch. 8.2, pp. 939–960.

[50] T. Cornsweet, *Visual Perception*. Academic Press, 1970.

[51] H. R. Sheikh, A. C. Bovik, and G. de Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Trans. on Image Processing*, vol. 14, no. 12, pp. 2117–2128, Dec. 2005.

[52] M. P. Eckert and A. P. Bradley, "Perceptual quality metrics applied to still image compression," *Signal Processing*, vol. 70, pp. 177–200, Nov. 1998.

[53] B. W. Keelan, *Handbook of Image Quality: Characterization and Prediction*. CRC Press, 2002.

[54] N. Burningham, Z. Pizlo, and J. P. Allebach, "Image quality metrics," in *Encyclopedia of Imaging Science and Technology*. New York: Wiley, 2002, pp. 598–616.

[55] U. Engelke and H.-J. Zepernick, "Perceptual-based quality metrics for image and video services: A survey," in *Proc. of EuroNGI Conf. on Next Generation Internet Networks Design and Engineering Heterogeneity*, May 2007, pp. 190–197.

[56] S. Winkler and P. Mohandas, "The evolution of video quality measurement: From PSNR to hybrid metrics," *IEEE Trans. on Broadcasting*, vol. 54, no. 3, pp. 660–668, Sep. 2008.

[57] J. You, U. Reiter, M. M. Hannuksela, M. Gabbouj, and A. Perkis, "Perceptual-based quality assessment for audio–visual services: A survey," *Signal Processing: Image Communication*, Mar. 2010, doi: 10.1016/j.image.2010.02.002.

[58] J. Mannos and D. Sakrison, "The effects of a visual fidelity criterion on the encoding of images," *IEEE Trans. on Information Theory*, vol. 20, no. 4, pp. 525–536, Jul. 1974.

[59] F. Lukas and Z. Budrikis, "Picture quality prediction based on a visual model," *IEEE Trans. on Communications*, vol. 30, no. 7, pp. 1679–1692, Jul. 1982.

[60] S. Daly, "The visible differences predictor: An algorithm for the assessment of image fidelity," *in Digital Images and Human Vision, MIT Press, A.B. Watson (Ed.)*, pp. 179–206, 1993.

[61] A. Bradley, "A wavelet visible difference predictor," *IEEE Trans. on Image Processing*, vol. 8, no. 5, pp. 717–730, May 1999.

[62] P. C. Teo and D. J. Heeger, "Perceptual image distortion," in *Proc. of IEEE Int. Conf. on Image Processing*, vol. 2, Nov. 1994, pp. 982–986.

[63] E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger, "Shiftable multi-scale transforms," *IEEE Trans. on Information Theory*, vol. 38, no. 2, pp. 587–607, Mar. 1992.

[64] C. J. van den Branden Lambrecht and O. Verscheure, "Perceptual quality measure using a spatio-temporal model of the human visual system," in *Proc. of SPIE Digital Video Compression: Algorithms and Technologies*, vol. 2668, Jan. 1996, pp. 450–460.

[65] J. Lubin, "A human vision system model for objective picture quality measurements," in *Int. Broadcasting Convention*, Sep. 1997, pp. 498–503.

[66] S. Winkler, "A perceptual distortion metric for digital color images," in *Proc. of IEEE Int. Conf. on Image Processing*, Oct. 1998, pp. 399–403.

[67] ——, "A perceptual distortion metric for digital color video," in *Proc. of IS&T/SPIE Human Vision and Electronic Imaging IV*, Jan. 1999, pp. 175–184.

[68] A. B. Watson, J. Hu, and J. F. McGowan, "DVQ: A digital video quality metric based on human vision," *Journal of Electronic Imaging*, vol. 10, no. 1, pp. 20–29, Jan. 2001.

[69] J. B. Martens, "Multidimensional modeling of image quality," *Proc. of the IEEE*, vol. 90, no. 1, pp. 133–153, Jan. 2002.

[70] A. M. Eskicioglu and P. S. Fisher, "Image quality measures and their performance," *IEEE Trans. on Communications*, vol. 43, no. 12, pp. 2959–2965, Dec. 1995.

[71] T. Yamashita, M.Kameda, and M.Miyahara, "An objective picture quality scale for video images (PQS video) - definition of distortion factors," in *Proc. of IS&T/SPIE Visual Communications and Image Processing*, vol. 4067, May 2000, pp. 801–809.

[72] K. T. Tan and M. Ghanbari, "A multi-metric objective picture-quality measurement model for MPEG video," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 10, no. 7, pp. 1208–1213, Oct. 2000.

[73] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. on Image Processing*, vol. 15, no. 2, pp. 430–444, Feb. 2006.

[74] D. J. C. MacKay, *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

[75] W. Lin, L. Dong, and P. Xue, "Visual distortion gauge based on discrimination of noticeable contrast changes," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 15, no. 7, pp. 900–909, Jul. 2005.

[76] D. M. Chandler and S. S. Hemami, "VSNR: A wavelet-based visual signal-to-noise ratio for natural images," *IEEE Trans. on Image Processing*, vol. 16, no. 9, pp. 2284–2298, Sep. 2007.

[77] E. C. Larson and D. M. Chandler, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," *Journal of Electronic Imaging, Special Section on Image Quality*, vol. 19, no. 1, 011006, Jan. 2010.

[78] C. Li and A. C. Bovik, "Content weighted video quality assessment using a three component image model," *Journal of Electronic Imaging, Special Section on Image Quality*, vol. 19, no. 1, 011003, Jan. 2010.

[79] Y. F. Ou, Z. Ma, and Y. Wang, "A novel quality metric for compressed video considering both frame rate and quantization artifacts," in *Proc. of Int. Workshop on Video Processing and Quality Metrics for Consumer Electronics*, Jan. 2009.

[80] A. Ninassi, O. Le Meur, P. Le Callet, and D. Barba, "Considering temporal variations of spatial visual distortions in video quality assessment," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 2, pp. 253–265, Apr. 2009.

[81] M. Barkowsky, J. Bialkowski, B. Eskofier, R. Bitto, and A. Kaup, "Temporal trajectory aware video quality measure," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 2, pp. 266–279, Apr. 2009.

[82] K. Seshadrinathan and A. C. Bovik, "Motion-tuned spatio-temporal quality assessment of natural videos," *IEEE Trans. on Image Processing*, vol. 19, no. 2, pp. 335–350, Feb. 2010.

[83] Z. Wang, H. R. Sheikh, and A. C. Bovik, "No-reference perceptual quality assessment of JPEG compressed images," in *Proc. of IEEE Int. Conf. on Image Processing*, vol. 1, Sep. 2002, pp. 477–480.

[84] Y. Horita, M. Sato, Y. Kawayoke, P. Sazzad, and K. Shibata, "Image quality evaluation model based on local features and segmentation," in *Proc. of IEEE Int. Conf. on Image Processing*, Oct. 2006, pp. 405–408.

[85] H. Koumaras, A. Kourtis, and D. Martakos, "Evaluation of video quality based on objectively estimated metric," *Journal of Communications and Networks*, vol. 7, no. 3, pp. 235–242, Sep. 2005.

[86] F. Yang, S. Wan, Y. Chang, and H. R. Wu, "A novel objective no-reference metric for digital video quality assessment," *IEEE Signal Processing Letters*, vol. 12, no. 10, pp. 685–688, Oct. 2005.

[87] M. Ries, O. Nemethova, and M. Rupp, "Video quality estimation for mobile H.264/AVC video streaming," *Journal of Communications*, vol. 3, no. 1, pp. 41–50, Jan. 2008.

[88] H. Liu, J. Redi, H. Alers, R. Zunino, and I. Heynderickx, "No-reference image quality assessment based on localized gradient statistics: Application to JPEG and JPEG2000," in *Proc. of IS&T/SPIE Human Vision and Electronic Imaging XV*, Jan. 2010.

[89] H. R. Wu and M. Yuen, "A generalized block-edge impairment metric for video coding," *IEEE Signal Processing Letters*, vol. 4, no. 11, pp. 317–320, Nov. 1997.

[90] T. Vlachos, "Detection of blocking artifacts in compressed video," *IEE Electronics Letters*, vol. 36, no. 13, pp. 1106–1108, Jun. 2000.

[91] A. Leontaris, P. C. Cosman, and A. R. Reibman, "Quality evaluation of motion-compensated edge artifacts in compressed video," *IEEE Trans. on Image Processing*, vol. 16, no. 4, pp. 943–956, Apr. 2007.

[92] X. Marichal, W. Y. Ma, and H. J. Zhang, "Blur determination in the compressed domain using DCT information," in *Proc. of IEEE Int. Conf. on Image Processing*, vol. 2, Sep. 1999, pp. 386–390.

[93] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi, "Perceptual blur and ringing metrics: Application to JPEG2000," *Signal Processing: Image Communication*, vol. 19, pp. 163–172, Feb. 2004.

[94] H. Liu, N. Klomp, and I. Heynderickx, “A no-reference metric for perceived ringing artifacts in images,” *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 20, no. 4, pp. 529–539, Apr. 2010.

[95] J. Caviedes and F. Oberti, “A new sharpness metric based on local kurtosis, edge and energy information,” *Signal Processing: Image Communication*, vol. 19, pp. 147–161, Feb. 2004.

[96] R. Ferzli and L. J. Karam, “A no-reference objective image sharpness metric based on the notion of just noticeable blur (JNB),” *IEEE Trans. on Image Processing*, vol. 18, no. 4, pp. 717–728, Apr. 2009.

[97] N. D. Narvekar and L. J. Karam, “A no reference perceptual image sharpness metric based on a cumulative probability of blur detection,” in *Proc. of Int. Workshop on Quality of Multimedia Experience*, Jul. 2009.

[98] M. C. Q. Farias and S. K. Mitra, “No-reference video quality metric based on artifact measurements,” in *Proc. of IEEE Int. Conf. on Image Processing*, vol. 3, Sep. 2005, pp. 141–144.

[99] P. Gastaldo, S. Rovetta, and R. Zunino, “Objective quality assessment of MPEG2 video streams by using CBP neural networks,” *IEEE Trans. on Neural Networks*, vol. 13, no. 4, pp. 939–947, Jul. 2002.

[100] S. Mohamed and G. Rubino, “A study of real-time packet video quality using random neural networks,” *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 12, no. 12, pp. 1071–1083, Dec. 2002.

[101] P. Le Callet, C. Viard-Gaudin, and D. Barba, “A convolutional neural network approach for objective video quality assessment,” *IEEE Trans. on Neural Networks*, vol. 17, no. 5, pp. 1316–1327, Sep. 2006.

[102] M. C. Q. Farias, M. Carli, and S. K. Mitra, “Objective video quality metric based on data hiding,” *IEEE Trans. on Consumer Electronics*, vol. 51, no. 3, pp. 983–992, Aug. 2005.

[103] A. Ninassi, P. Le Callet, and F. Autrusseau, “Pseudo no reference image quality metric using perceptual data hiding,” in *Proc. of IS&T/SPIE Human Vision and Electronic Imaging XI*, vol. 6057, Feb. 2006.

[104] I. Cox, M. L. Miller, and J. A. Bloom, *Digital Watermarking*. Morgan Kaufmann, 2002.

[105] M. H. Pinson and S. Wolf, “A new standardized method for objectively measuring video quality,” *IEEE Trans. on Broadcasting*, vol. 50, no. 3, pp. 312–322, Sep. 2004.

[106] Z. Wang and E. P. Simoncelli, “Reduced-reference image quality assessment using a wavelet-domain natural image statistic model,” in *Proc. of IS&T/SPIE Human Vision and Electronic Imaging X*, vol. 5666, Mar. 2005, pp. 149–159.

[107] Q. Li and Z. Wang, "Reduced-reference image quality assessment using divisive normalization-based image representation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 2, pp. 202–211, Apr. 2009.

[108] M. Masry, S. S. Hemami, and Y. Sermadevi, "A scalable wavelet-based video distortion metric and applications," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 16, no. 2, pp. 260–273, Feb. 2006.

[109] T. Yamada, Y. Miyamoto, M. Serizawa, and H. Harasaki, "Reduced-reference based video quality metrics using representative luminance values," in *Proc. of Int. Workshop on Video Processing and Quality Metrics for Consumer Electronics*, Jan. 2007.

[110] I. P. Gunawan and M. Ghanbari, "Reduced-reference video quality assessment using discriminative local harmonic strength with motion consideration," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 18, no. 1, pp. 71–83, Jan. 2008.

[111] M. Carnec, P. Le Callet, and D. Barba, "Objective quality assessment of color images based on a generic perceptual reduced reference," *Signal Processing: Image Communication*, vol. 23, no. 4, pp. 239–256, Apr. 2008.

[112] K. Chono, Y.-C. Lin, D. Varodayan, Y. Miyamoto, and B. Girod, "Reduced-reference image quality assessment using distributed source coding," in *Proc. of IEEE Int. Conf. on Multimedia and Expo*, Jun. 2008, pp. 609–612.

[113] T. L. Lin, S. Kanumuri, Y. Zhi, D. Poole, P. Cosman, and A. R. Reibman, "A versatile model for packet loss visibility and its application in packet prioritization," *IEEE Trans. on Image Processing*, vol. 19, no. 3, pp. 722–735, Mar. 2010.

[114] Z. Wang, G. Wu, H. R. Sheikh, E. P. Simoncelli, E. H. Yang, and A. C. Bovik, "Quality aware images," *IEEE Trans. on Image Processing*, vol. 15, no. 6, pp. 1680–1689, Jun. 2006.

[115] J. Wolfe, "Visual attention," in *Seeing*, K. K. D. Valois, Ed. Academic Press, 2000, pp. 335–386.

[116] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature Reviews Neuroscience*, vol. 2, pp. 192–203, 2001.

[117] S. Treue, "Visual attention: The where, what, how and why of saliency," *Current Opinion in Neurobiology*, vol. 13, no. 4, pp. 428–432, Aug. 2003.

[118] J. M. Wolfe and T. S. Horowitz, "What attributes guide the deployment of visual attention and how do they do it?" *Nature Reviews Neuroscience*, vol. 5, pp. 1–7, Jun. 2004.

[119] M. Cerf, E. P. Frady, and C. Koch, "Faces and text attract gaze independent of the task: Experimental data and computer model,"

Journal of Vision, vol. 9, no. 12:10, pp. 1–15, 2009. [Online]. Available: http://journalofvision.org/9/12/10/

[120] J. M. Wolfe, K. R. Cave, and S. L. Franzel, "Guided search: An alternative to the feature integration model for visual search," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 15, no. 3, pp. 419–433, Aug. 1989.

[121] W. Einhäuser, M. Spain, and P. Perona, "Objects predict fixations better than early saliency," *Journal of Vision*, vol. 8, no. 14:18, pp. 1–26, 2008. [Online]. Available: http://journalofvision.org/8/14/18/

[122] M. S. Castelhano, M. L. Mack, and J. M. Henderson, "Viewing task influences eye movement control during active scene perception," *Journal of Vision*, vol. 9, no. 3:6, pp. 1–15, 2009. [Online]. Available: http://journalofvision.org/9/3/6/

[123] T. Betz, T. C. Kietzmann, N. Wilming, and P. König, "Investigating task-dependent top-down effects on overt visual attention," *Journal of Vision*, vol. 10, no. 3:15, pp. 1–14, 2010. [Online]. Available: http://journalofvision.org/10/3/15/

[124] A. L. Yarbus, *Eye Movements and Vision*. Plenum, 1967.

[125] C. M. Privitera and L. W. Stark, "Algorithms for defining visual regions-of-interest: Comparison with eye fixations," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2000.

[126] T. Foulsham and G. Underwood, "What can saliency models predict about eye movements? Spatial and sequential aspects of fixations during encoding and recognition," *Journal of Vision*, vol. 8, no. 2:6, pp. 1–17, 2008. [Online]. Available: http://journalofvision.org/8/2/6/

[127] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive Psychology*, vol. 12, no. 1, pp. 97–136, Jan. 1980.

[128] C. Koch and S. Ullman, "Shifts in selection in visual attention: Towards the underlying neural circuitry," *Human Neurobiology*, vol. 4, no. 4, pp. 219–227, 1985.

[129] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[130] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau, "A coherent computational approach to model bottom-up visual attention," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 28, no. 5, pp. 802–817, May 2006.

[131] S. Marat, T. H. Phuoc, L. Granjon, N. Guyader, D. Pellerin, and A. Guérin-Dugué, “Modelling spatio-temporal saliency to predict gaze direction for short videos,” *Int. Journal of Computer Vision*, vol. 82, no. 3, pp. 231–243, May 2009.

[132] F. W. M. Stentiford, “An estimator for visual attention through competitive novelty with application to image compression,” in *Proc. of Picture Coding Symposium*, Apr. 2001.

[133] Y. Hu, D. Rajan, and L. T. Chia, “Robust subspace analysis for detecting visual attention regions in images,” in *Proc. of ACM Int. Conference on Multimedia*, Nov. 2005, pp. 716–724.

[134] Y. Zhai and M. Shah, “Visual attention detection in video sequences using spatiotemporal cues,” in *Proc. of ACM Int. Conf. on Multimedia*, Oct. 2006, pp. 815–824.

[135] U. Rajashekar, I. van der Linde, A. C. Bovik, and L. K. Cormack, “GAFFE: A gaze-attentive fixation finding engine,” *IEEE Trans. on Image Processing*, vol. 17, no. 4, pp. 564–573, Apr. 2008.

[136] C. Guo, Q. Ma, and L. Zhang, “Spatio-temporal saliency detection using phase spectrum of quaternion Fourier transform,” in *Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition*, Jun. 2008, pp. 1–8.

[137] R. Achanta, S. Hemami, F. Estrada, and S. Süsstrunk, “Frequency-tuned salient region detection,” in *Proc. of IEEE Int. Conf on Computer Vision and Pattern Recognition*, Jun. 2009, pp. 1597–1604.

[138] D. Liu and T. Chen, “DISCOV: A framework for discovering objects in video,” *IEEE Trans. on Multimedia*, vol. 10, no. 2, pp. 200–208, Feb. 2008.

[139] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, “SUN: A Bayesian framework for saliency using natural statistics,” *Journal of Vision*, vol. 8, no. 7:32, pp. 1–20, 2008. [Online]. Available: http://journalofvision.org/8/7/32/

[140] L. W. Renninger, P. Verghese, and J. Coughlan, “Where to look next? Eye movements reduce local uncertainty,” *Journal of Vision*, vol. 7, no. 3:6, pp. 1–17, 2007. [Online]. Available: http://journalofvision.org/7/3/6/

[141] N. D. B. Bruce and J. K. Tsotsos, “Saliency, attention, and visual search: An information theoretic approach,” *Journal of Vision*, vol. 9, no. 3:5, pp. 1–24, 2009. [Online]. Available: http://journalofvision.org/9/3/5/

[142] H. J. Seo and P. Milanfar, “Static and space-time visual saliency detection by self-resemblance,” *Journal of Vision*, vol. 9, no. 12:15, pp. 1–27, 2009. [Online]. Available: http://journalofvision.org/9/12/15/

[143] W. Kienzle, F. A. Wichmann, B. Schölkopf, and M. O. Franz, "A non-parametric approach to bottom-up visual saliency," in *Proc. of Advances in Neural Information Processing Systems*, Sep. 2007, pp. 689–696.

[144] F. Navalpakkam and L. Itti, "Modeling the influence of task on attention," *Vision Research*, vol. 45, no. 2, pp. 205–231, Jan. 2005.

[145] A. D. Hwang, E. C. Higgins, and M. Pomplun, "A model of top-down attentional control during visual search in complex scenes," *Journal of Vision*, vol. 9, no. 5:25, pp. 1–18, 2009. [Online]. Available: http://journalofvision.org/9/5/25/

[146] Y. F. Ma, X. S. Hua, L. Lu, and H. J. Zhang, "A generic framework of user attention model and its application in video summarization," *IEEE Trans. on Multimedia*, vol. 7, no. 5, pp. 907–919, Oct. 2005.

[147] C. D. Vleeschouwer, X. Marichal, T. Delmot, and B. Macq, "A fuzzy logic system able to detect interesting areas of a video sequence," in *Proc. of IS&T/SPIE Human Vision and Electronic Imaging II*, vol. 3016, Jan. 1997, pp. 234–245.

[148] W. Osberger and A. M. Rohaly, "Automatic detection of regions of interest in complex video sequences," in *Proc. of IS&T/SPIE Human Vision and Electronic Imaging VI*, vol. 4299, Jan. 2001, pp. 361–372.

[149] S. Pinneli and D. M. Chandler, "A Bayesian approach to predicting the perceived interest of objects," in *Proc. of IEEE Int. Conf. on Image Processing*, Oct. 2008, pp. 2584–2587.

[150] W. Osberger and A. J. Maeder, "Automatic identification of perceptually important regions in an image," in *Proc. of IEEE Int. Conf. on Pattern Recognition*, vol. 1, Aug. 1998, pp. 701–704.

[151] R. Barland and A. Saadane, "Blind quality metric using a perceptual importance map for JPEG-2000 compressed images," in *Proc. of IEEE Int. Conf. on Image Processing*, Oct. 2006, pp. 2941–2944.

[152] P. Arbelaez, C. Fowlkes, and D. Martin, "The Berkeley segmentation dataset and benchmark," http://www.eecs.berkeley.edu/Research/Projects/CS/vision/grouping/segbench/, 2007.

[153] U. Engelke, H.-J. Zepernick, and A. J. Maeder, "Visual attention modeling: Region-of-interest versus fixation patterns," in *Proc. of IEEE Picture Coding Symposium*, May 2009, pp. 1–4.

[154] C. M. Masciocchi, S. Mihalas, D. Parkhurst, and E. Niebur, "Everyone knows what is interesting: Salient locations which should be fixated," *Journal of Vision*, vol. 9, no. 11:25, pp. 1–22, 2009. [Online]. Available: http://journalofvision.org/9/11/25/

[155] J. Wang, D. M. Chandler, and P. Le Callet, "Quantifying the relationship between visual salience and visual importance," in *Proc. of IS&T/SPIE Human Vision and Electronic Imaging XV*, vol. 7527, Jan. 2010.

[156] U. Engelke, H.-J. Zepernick, and A. Maeder, "Visual fixation patterns in subjective quality assessment: The relative impact of image content and structural distortions," in *Proc. of Int. Symp. on Intelligent Signal Processing and Communications Systems*, Dec. 2010.

[157] F. Boulos, W. Chen, B. Parrein, and P. Le Callet, "A new H.264/AVC error resilience model based on regions of interest," in *Proc. of Int. Packet Video Workshop*, May 2009.

[158] H. P. Frey, C. Honey, and P. König, "What's color got to do with it? The influence of color on visual attention in different categories," *Journal of Vision*, vol. 8, no. 14:6, pp. 1–17, 2008. [Online]. Available: http://journalofvision.org/8/14/6/

[159] M. C. Doyle and R. J. Snowden, "Identification of visual stimuli is improved by accompanying auditory stimuli: The role of eye movements and sound location," *Perception*, vol. 30, no. 7, pp. 795–810, Jul. 2001.

[160] K. K. Evans and A. Treisman, "Natural cross-modal mappings between visual and auditory features," *Journal of Vision*, vol. 10, no. 1:6, pp. 1–12, 2010. [Online]. Available: http://journalofvision.org/10/1/6/

[161] H. Liu and I. Heynderickx, "Studying the added value of visual attention in objective image quality metrics based on eye movement data," in *Proc. of IEEE Int. Conf. on Image Processing*, Nov. 2009, pp. 3097–3100.

[162] A. Cavallaro and S. Winkler, "Segmentation-driven perceptual quality metrics," in *Proc. of IEEE Int. Conf. on Image Processing*, vol. 5, Oct. 2004, pp. 3543–3546.

[163] A. J. Maeder, "The image importance approach to human vision based image quality characterization," *Pattern Recognition Letters*, vol. 26, no. 3, pp. 347–354, Feb. 2005.

[164] A. K. Moorthy and A. C. Bovik, "Visual importance pooling for image quality assessment," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 2, pp. 193–201, Apr. 2009.

[165] Z. Lu, W. Lin, X. Yang, E. Ong, and S. Yao, "Modeling visual attention's modulatory aftereffects on visual sensitivity and quality evaluation," *IEEE Trans. on Image Processing*, vol. 14, no. 11, pp. 1928–1942, Nov. 2005.

[166] J. You, A. Perkis, M. Hannuksela, and M. Gabbouj, "Perceptual quality assessment based on visual attention analysis," in *Proc. of ACM Int. Conference on Multimedia*, Oct. 2009, pp. 561–564.

[167] Q. Ma, L. Zhang, and B. Wang, “New strategy for image and video quality assessment,” *Journal of Electronic Imaging, Special Section on Image Quality*, vol. 19, no. 1, 011019, Jan. 2010.

[168] X. Feng, T. Liu, D. Yang, and Y. Wang, “Saliency based objective quality assessment of decoded video affected by packet losses,” in *Proc. of IEEE Int. Conf. on Image Processing*, Oct. 2008, pp. 2560–2563.

[169] T. Liu, X. Feng, A. Reibman, and Y. Wang, “Saliency inspired modeling of packet-loss visibility in decoded videos,” in *Proc. of Int. Workshop on Video Processing and Quality Metrics for Consumer Electronics*, Jan. 2009.

[170] J. S. Lee, F. de Simone, and T. Ebrahimi, “Influence of audio-visual attention on perceived quality of standard definition multimedia content,” in *Proc. of Int. Workshop on Quality of Multimedia Experience*, Jul. 2009, pp. 13–18.

[171] A. Ninassi, O. L. Meur, P. Le Callet, and D. Barba, “Does where you gaze on an image affect your perception of quality? Applying visual attention to image quality metric,” in *Proc. of IEEE Int. Conf. on Image Processing*, vol. 2, Oct. 2007, pp. 169–172.

[172] E. C. Larson, C. Vu, and D. M. Chandler, “Can visual fixation patterns improve image fidelity assessment?” in *Proc. of IEEE Int. Conf. on Image Processing*, Oct. 2008, pp. 2572–2575.

[173] U. Engelke, M. Barkowsky, P. Le Callet, and H.-J. Zepernick, “Modelling saliency awareness for objective video quality assessment,” in *Proc. of Int. Workshop on Quality of Multimedia Experience*, Jun. 2010.

[174] U. Engelke and H.-J. Zepernick, “A framework for optimal region-of-interest based quality assessment in wireless imaging,” *Journal of Electronic Imaging, Special Section on Image Quality*, vol. 19, no. 1, 011005, Jan. 2010.

[175] U. Engelke, A. Maeder, and H.-J. Zepernick, “VAIQ: The visual attention for image quality database,” http://www.bth.se/tek/rcg.nsf/pages/vaiq-db, 2009.

[176] H. Liu and I. Heynderickx, “TUD image quality database: Eye-tracking release 1,” http://mmi.tudelft.nl/ ingrid/eyetracking1.html, 2010.

[177] T. M. Kusuma, “A perceptual-based objective quality metric for wireless imaging,” Ph.D. dissertation, Curtin University of Technology, Perth, Australia, 2005.

[178] U. Engelke and H.-J. Zepernick, “Quality evaluation in wireless imaging using feature-based objective metrics,” in *Proc. of IEEE Int. Symp. on Wireless Pervasive Computing*, Feb. 2007, pp. 367–372.

[179] ——, “Quality assessment of an adaptive filter for artifact reduction in mobile video sequences,” in *Proc. of IEEE Int. Symp. on Wireless Pervasive Computing*, Feb. 2007, pp. 360–366.

[180] ——, “Pareto optimal weighting of structural impairments for wireless imaging quality assessment,” in *Proc. of IEEE Int. Conf. on Image Processing*, Oct. 2008, pp. 373–376.

[181] ——, “An artificial neural network for quality assessment in wireless imaging based on extraction of structural information,” in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Apr. 2007, pp. 1249–1252.

[182] ——, “Multiobjective optimization of multiple scale visual quality processing,” in *Proc. of IEEE Int. Workshop on Multimedia Signal Processing*, Oct. 2008, pp. 212–217.

[183] ——, “Multi-resolution structural degradation metrics for perceptual image quality assessment,” in *Proc. of Picture Coding Symposium*, Nov. 2007.

[184] ——, “Optimal region-of-interest based visual quality assessment,” in *Proc. of IS&T/SPIE Human Vision and Electronic Imaging XIV*, vol. 7240, Jan. 2009.

[185] U. Engelke, X. N. Vuong, and H.-J. Zepernick, “Regional attention to structural degradations for perceptual image quality metric design,” in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Mar. 2008, pp. 869–872.

[186] U. Engelke, A. J. Maeder, and H.-J. Zepernick, “Visual attention modelling for subjective image quality databases,” in *Proc. of IEEE Int. Workshop on Multimedia Signal Processing*, Oct. 2009, pp. 1–6.

[187] ——, “On confidence and response times of human observers in subjective image quality assessment,” in *Proc. of IEEE Int. Conf. on Multimedia and Expo*, Jun. 2009, pp. 910–913.

[188] U. Engelke, A. Maeder, and H.-J. Zepernick, “Analysing inter-observer saliency variations in task-free viewing of natural images,” in *Proc. of IEEE Int. Conf. on Image Processing*, Sep. 2010.

[189] U. Engelke, A. J. Maeder, and H.-J. Zepernick, “The effect of spatial distortion distributions on human viewing behaviour when judging image quality,” in *Proc. of European Conf. on Visual Perception*, Aug. 2009, p. 22.

[190] U. Engelke, R. Pepion, P. Le Callet, and H.-J. Zepernick, “Linking distortion perception and visual saliency in H.264/AVC coded video containing packet loss,” in *Proc. of SPIE/IEEE Int. Conf. on Visual Communications and Image Processing*, Jul. 2010.

[191] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 560–576, Jul. 2003.

[192] S. Lin and D. J. Costello, *Error Control Coding: Fundamentals and Applications*. Prentice Hall, 1983.

[193] A. F. Molisch, *Wireless Communications*. Wiley-IEEE Press, 2005.

[194] S. S. Hemami, D. M. Chandler, B. Chern, and J. A. Moses, "Suprathreshold visual psychophysics and structure-based visual masking," in *Proc. of IS&T/SPIE Visual Communications and Image Processing*, vol. 6077, Jan. 2006.

[195] International Telecommunication Union, "Specifications and alignment procedures for setting of brightness and contrast of displays," ITU-R, Rec. BT.814, 1994.

[196] ——, "Specification of a signal for measurement of the contrast ratio of displays," ITU-R, Rec. BT.815, 1994.

[197] Cambridge Research Systems, "ColorCAL colorimeter," http://www.crsltd.com/catalog/colorcal/index.html, 2010.

[198] DisplayMate Technologies Corp, "DisplayMate Multimedia Edition," http://www.displaymate.com/infodmm.html, 2010.

[199] International Telecommunication Union, "Subjective assessment of standard definition digital television (SDTV) systems," ITU-R, Rec. BS.1129-2, 1998.

[200] Video Quality Experts Group, "Final report from the Video Quality Experts Group on the validation of objective models of multimedia quality assessment, phase I," VQEG, Sep. 2008.

[201] P. Corriveau, A. Webster, A. M. Rohaly, and J. Libert, "Video quality experts group: The quest for valid objective methods," in *Proc. of IS&T/SPIE Human Vision and Electronic Imaging V*, vol. 3959, Jan. 2000, pp. 129–139.

[202] Video Quality Experts Group, "Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment, phase II," VQEG, Aug. 2003.

[203] S. Winkler, "On the properties of subjective ratings in video quality experiments," in *Proc. of Int. Workshop on Quality of Multimedia Experience*, Jul. 2009, pp. 139–144.

[204] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi, "A no-reference perceptual blur metric," in *Proc. of IEEE Int. Conf. on Image Processing*, vol. 3, Sep. 2002, pp. 57–60.

[205] S. Saha and R. Vemuri, "An analysis on the effect of image features on lossy coding performance," *IEEE Signal Processing Letters*, vol. 7, no. 5, pp. 104–107, May 2000.

[206] J.-R. Ohm, *Multimedia Communication Technology: Representation, Transmission and Identification of Multimedia Signals*. Springer, 2004.

[207] M. C. Q. Farias, M. S. Moore, J. M. Foley, and S. K. Mitra, "Perceptual contributions of blocky, blurry, and fuzzy impairments to overall annoyance," in *Proc. of IS&T/SPIE Human Vision and Electronic Imaging IX*, vol. 5292, Jan. 2004, pp. 109–120.

[208] M. C. Q. Farias, J. M. Foley, and S. K. Mitra, "Perceptual analysis of video impairments that combine blocky, blurry, noisy, and ringing synthetic artifacts," in *Proc. of IS&T/SPIE Human Vision and Electronic Imaging X*, vol. 5666, Jan. 2005, pp. 107–118.

[209] H. de Ridder and M. C. Willemsen, "Percentage scaling: A new method for evaluating multiple impaired images," in *Proc. of IS&T/SPIE Human Vision and Electronic Imaging V*, vol. 3959, Jan. 2000, pp. 68–77.

[210] H. de Ridder, "Minkowski-metrics as a combination rule for digital-image-coding impairments," in *Proc. of IS&T/SPIE Human Vision, Visual Processing, and Digital Display III*, vol. 1666, Jan. 1992, pp. 16–26.

[211] A. Saadane, "Toward a unified fidelity metric of still-coded images," *Journal of Electronic Imaging*, vol. 16, no. 1, Jan. 2007, doi: 10.1117/1.2437728.

[212] N. . ITS, "A3: Objective video quality measurement using a peak-signal-to-noise-ratio (PSNR) full reference technique," *ATIS T1.TR.PP.74-2001*, 2001.

[213] F. Pereira and T. Ebrahimi (Ed.), *The MPEG-4 Book*. Prentice Hall PTR, 2002.

[214] International Telecommunication Union, "Advanced video coding for generic audiovisual services," ITU-T, Rec. H.264, Nov. 2007.

[215] ——, "Examples for H.263 encoder/decoder implementations, appendix III," ITU-T, Rec. H.263, Jun. 2000.

[216] A. Rossholm and K. Andersson, "Adaptive de-blocking de-ringing filter," in *Proc. of IEEE Int. Conf. on Image Processing*, Sep. 2005, pp. 1042–1045.

[217] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 2nd ed. Prentice Hall, 2002.

[218] C. M. Fonseca and P. J. Fleming, "Multiobjective optimization," in *Evolutionary Computation 2: Advanced Algorithms and Operations*, T. Back, D. B. Fogel, and Z. Michalewicz, Eds. Taylor & Francis, 2000, ch. 5, pp. 25–37.

[219] L. A. Zadeh, "Optimality and non-scalar-valued performance criteria," *IEEE Trans. on Automatic Control*, vol. 8, no. 1, pp. 59–60, Jan. 1963.

[220] F. W. Gembicki and Y. Y. Haimes, "Approach to performance and sensitivity multiobjective optimization: The goal attainment method," *IEEE Trans. on Automatic Control*, vol. 20, no. 6, pp. 769–771, Dec. 1975.

[221] R. Fletcher, *Practical Methods of Optimization*, 2nd ed. John Wiley & Sons, 1987.

[222] S. Haykin, *Neural Networks: A Comprehensive Foundation*, 1st ed. Macmillan, 1994.

[223] P. Gastaldo and R. Zunino, "Neural networks for the no-reference assessment of perceived quality," *Journal of Electronic Imaging*, vol. 14, no. 3, Jul. 2005.

[224] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[225] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks*, vol. 2, no. 5, pp. 359–366, 1989.

[226] B. Widrow and M. A. Lehr, "30 years of adaptive neural networks: Perceptron, madaline, and backpropagation," *Proc. of the IEEE*, vol. 78, no. 9, pp. 1415–1442, Sep. 1990.

[227] K. Levenberg, "A method for the solution of certain non-linear problems in least squares," *The Quarterly of Applied Mathematics*, vol. 2, pp. 164–168, 1944.

[228] D. W. Marquardt, "An algorithm for least-squares estimation of nonlinear parameters," *Journal of Applied Mathematics*, vol. 11, no. 2, pp. 431–441, Jun. 1963.

[229] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

[230] E. H. Adelson, C. H. Anderson, J. R. Bergen, P. J. Burt, and J. M. Ogden, "Pyramid methods in image processing," *RCA Engineer*, vol. 29, no. 6, pp. 33–41, 1984.

[231] P. J. Burt and E. H. Adelson, "The Laplacian pyramid as a compact image code," *IEEE Trans. on Communications*, vol. 31, no. 4, Apr. 1983.

[232] K. Seshadrinathan and A. C. Bovik, "Unifying analysis of full reference image quality assessment," in *Proc. of IEEE Int. Conf. on Image Processing*, Oct. 2008, pp. 1200–1203.

[233] EyeTech Digital Systems, "TM3 eye tracker," http://www.eyetechds.com/, 2009.

[234] G. W. Cermak, “An experimental test of two models of attribute integration,” *Personality and Social Psychology*, Jul. 1983.

[235] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. New York: Chapman & Hall/CRC, 1993.

[236] N. Metropolis and S. Ulam, “The Monte Carlo method,” *Journal of the American Statistical Association*, vol. 44, no. 247, pp. 335–341, Sep. 1949.

[237] B. Efron and R. Tibshirani, “Bootstrap methods for standard errors, confidence intervals and other measures of statistical accuracy,” *Statistical Science*, vol. 1, no. 1, p. 54–75, Feb. 1986.

[238] A. W. Bowman and A. Azzalini, *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations*. Oxford University Press, 1997.

[239] R. Kohavi, “A study of cross-validation and Bootstrap for accuracy estimation and model selection,” in *Proc. of Int. Joint Conf. on Artificial Intelligence*, Aug. 1995, pp. 1137–1143.

[240] L. Elazary and L. Itti, “Interesting objects are visually salient,” *Journal of Vision*, vol. 8, no. 3:3, pp. 1–15, 2008. [Online]. Available: http://journalofvision.org/8/3/3/

[241] C. Fookes, A. Maeder, S. Sridharan, and G. Mamic, “Gaze based personal identification,” in *Behavioral Biometrics for Human Identification: Intelligent Applications*. IGI Global, 2009.

[242] J. Cohen, P. Cohen, S. G. West, and L. S. Aiken, *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, 3rd ed. Lawrence Erlbaum Associates, 2003.

[243] C. T. Vu, E. C. Larson, and D. M. Chandler, “Visual fixation patterns when judging image quality: Effects of distortion type, amount, and subject experience,” in *Proc. of IEEE Southwest Symposium on Image Analysis and Interpretation*, Mar. 2008, pp. 73–76.

[244] T. Fawcett, “An introduction to ROC analysis,” *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[245] L. Itti and P. Baldi, “A principled approach to detecting surprising events in video,” in *Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition*, vol. 1, Jun. 2005, pp. 631–637.

[246] Cambridge Research Systems, “Tools for vision science,” http://www.crsltd.com/catalog/eye-trackers/index.html, 2009.

[247] Heinrich Hertz Institute Berlin, “H.264/AVC reference software JM 16.1,” http://iphome.hhi.de/suehring/tml/, 2009.

[248] Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG, "SVC/AVC Loss Simulator," http://wftp3.itu.int/av-arch/jvt-site/2005_10_nice/, 2005.

[249] Video Clarity, "ClearView," http://www.videoclarity.com/products.html, 2010.

[250] SensoMotoric Instruments, "iView X™Hi-Speed," http://www.smivision.com/en/eye-gaze-tracking-systems/products/iview-x-hi-speed.html, 2010.

[251] S. Kanumuri, P. C. Cosman, A. R. Reibman, and V. A. Vaishampayan, "Modeling packet-loss visibility in MPEG-2 video," *IEEE Trans. on Multimedia*, vol. 8, no. 2, pp. 341–355, Apr. 2006.

[252] K. Brunnström and B. N. Schenkman, "Comparison of the predictions of a spatiotemporal model with the detection of distortion in small moving images," *Optical Engineering*, vol. 41, no. 3, pp. 711–722, Mar. 2002.

[253] Y. Wang, Z. Wu, and J. M. Boyce, "Modeling of transmission-loss-induced distortion in decoded video," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 16, no. 6, pp. 716–732, Jun. 2006.

[254] S. E. Maxwell and H. D. Delaney, *Designing Experiments and Analyzing Data: A Model Comparison Perspective*, 2nd ed. Lawrence Erlbaum Associates, 2004.

[255] M. S. Moore, J. M. Foley, and S. K. Mitra, "Defect visibility and content importance: Effects on perceived impairment," *Signal Processing: Image Communication*, vol. 19, pp. 185–203, Feb. 2004.

[256] M. Fairchild, *Color Appearance Models*, 2nd ed. Chichester, UK: Wiley IS&T, 2005.

[257] D. Hasler and S. Süsstrunk, "Measuring colourfulness in natural images," in *Proc. of IS&T/SPIE Human Vision and Electronic Imaging XIII*, vol. 5007, Jan. 2003, pp. 87–95.

[258] S. Winkler and C. Faller, "Perceived audiovisual quality of low-bitrate multimedia content," *IEEE Trans. on Multimedia*, vol. 8, no. 5, pp. 973–980, Oct. 2006.

[259] M. N. Garcia and A. Raake, "Impairment-factor-based audio-visual quality model for IPTV," in *Proc. of Int. Workshop on Quality of Multimedia Experience*, Jul. 2009, pp. 1–6.

[260] J. Xia, Y. Shi, K. Teunissen, and I. Heynderickx, "Perceivable artifacts in compressed video and their relation to video quality," *Signal Processing: Image Communication*, vol. 24, no. 7, pp. 548–556, Aug. 2009.

ABSTRACT

The evolution of advanced radio transmission technologies for third and future generation mobile radio systems has paved the way for the delivery of mobile multimedia services. This is further enabled through contemporary video coding standards, such as H.264/AVC, allowing wireless image and video applications to become a reality on modern mobile devices. The extensive amount of data needed to represent the visual content and the scarce channel bandwidth constitute great challenges for network operators to deliver an intended quality of service. Appropriate metrics are thus instrumental for service providers to monitor the quality as experienced by the end user. This thesis focuses on subjective and objective assessment methods of perceived visual quality in image and video communication. The content of the thesis can be broadly divided into four parts.

Firstly, the focus is on the development of image quality metrics that predict perceived quality degradations due to transmission errors. The metrics follow the reduced-reference approach, thus, allowing to measure quality loss during image communication with only little overhead as side information. The metrics are designed and validated using subjective quality ratings from two experiments. The distortion assessment performance is further demonstrated through an application for filter design.

The second part of the thesis then investigates various methodologies to further improve the quality prediction performance of the metrics. In this respect, several properties of the human visual system are investigated and incorporated into the metric design. It is shown that the quality prediction performance can be considerably improved using these methodologies.

The third part is devoted to analysing the impact of the complex distortion patterns on the overall perceived quality, following two goals. Firstly, the confidence of human observers is analysed to identify the difficulties during assessment of the distorted images, showing, that indeed the level of confidence is highly dependent on the level of visual quality. Secondly, the impact of content saliency on the perceived quality is identified using region-of-interest selections and eye tracking data from two independent subjective experiments. It is revealed, that the saliency of the distortion region indeed has an impact on the overall quality perception and also on the viewing behaviour of human observers when rating image quality.

Finally, the quality perception of H.264/AVC coded video containing packet loss is analysed based on the results of a combined subjective video quality and eye tracking experiment. It is shown that the distortion location in relation to the content saliency has a tremendous impact on the overall perceived quality. Based on these findings, a framework for saliency aware video quality assessment is proposed that strongly improves the quality prediction performance of existing video quality metrics.

BLEKINGE TEKNISKA HÖGSKOLA · BTH ·

ISSN 1653-2090
ISBN 978-91-7295-185-3

2010:06