Data-Driven Estimation and Neurophysiological Assessment of Perceived Visual Quality

vorgelegt von
Dipl.-Ing. Sebastian Bosse
geb. in Mettingen

Von der Fakultät IV - Elektrotechnik und Informatik
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktor der Ingenieurwissenschaften
- Dr.-Ing. -

genehmigte Dissertation

Promotionsausschuss:

Vorsitzender:	Prof. Dr. Benjamin Blankertz
1. Gutachter:	Prof. Dr.-Ing. Thomas Wiegand
2. Gutachter:	Prof. Dr. Klaus-Robert Müller
3. Gutachter:	Prof. Anthony M. Norcia, Ph.D.

Tag der wissenschaftlichen Aussprache: 3. Juli 2018

Berlin, 2018

Das Rumgehopse hier ist ja nur
das Vermeiden einer Berufswahl.
Es wird Zeit, sich was Neues einfallen zu lassen.

— René Pollesch, *Service/No Service*

Acknowledgements

This thesis is the result of an intellectual adventure that started with the development of coding tools for video compression and led to the measurement of EEG signals — the little I expected this trajectory, the more I enjoyed it! I am deeply grateful to Prof. Dr.-Ing. Thomas Wiegand for giving me the opportunity to work in his group/department/institute, for his encouragement to dare this adventure, for stimulating discussions and for his constant trust in my work.
During this journey I met many people that paved my way and to whom I owe my greatest gratitude: Prof. Dr. Klaus-Robert Müller was of continuous support throughout the work on this thesis with his scientific advices, friendly encouragement and contagious joy in science – and by regularly assuring me that 'there is a life after the Ph.D.'.
I am very grateful to Prof. Anthony M. Norcia, Ph.D. for allowing me to visit his lab and for sharing his curiosity in natural image statistics and the human visual system. I found my stay at the Stanford Vision and Neuro-Development Lab among of the best experiences in my academic 'career'.
My work greatly benefited from the discussions with Prof. Dr. Gabriel Curio — and his sharp witted comments. Dr. Detlev Marpe gave me the liberty and space in his department necessary for this work, and frequently reminded me to submit this thesis. Dr. Wojciech Samek was of tremendous help as a group leader and colleague. I am very thankful for his trust in my work and his constant encouragement to share my results with the research community.
During my work at HHI I had the privilege to supervise excellent student researchers. Milena Bagdasarian, Sören Becker, Michael Dietzel, Dominique Maniry, and Zacharias Fisches supported me finishing the thesis by running subjective tests, helping implementing algorithms and challenging my ideas.
Many colleagues and friends proof-read drafts of this thesis and helped me transforming it in something resembling a dissertation. Thank you all, especially Dr. Wojciech Samek and Peter J. Kohler, Ph.D., for valuable feedback.
Without the help of Gabriele Thiele and Dominique Jojade I would probably be stuck in some administrative rabbit hole. Unfortunately it is impossible to list all other colleagues at HHI who supported me during the work on my thesis, be it with professional discussions in front of the coffee machine or slightly less professional discussions in front of a bar.

Most of all I have to thank Britta and Thea: Britta proves that perception is something worth to have and does not even need science for that; Thea takes away all the doubts and does not only show that learning-based systems actually work, but also that they can have the most beautiful smile.

Abstract

Perceptual quality is one of the key aspects of modern multimedia communication systems. Nevertheless, the questions of how to reliably assess quality as perceived by humans, how to computationally estimate perceived quality and how to incorporate computational models of quality in multimedia systems can still not be answered satisfactorily, despite decades of research. However, these problems are connected, because multimedia systems are typically evaluated in quality assessment studies, the outcome of quality assessment studies informs the design of computational quality models and computational quality models are in turn used for the optimization of multimedia systems.

This dissertation contributes to the current state of research in several ways. First, a novel neural network-based end-to-end optimized model for image quality estimation is proposed. The proposed method achieves prediction performance that is superior to the state-of-the-art for no-reference as well as for full-reference quality estimation.

The second contribution is a formal definition of distortion sensitivity that leads to the derivation of a computationally graceful, perception-based adaptation that can be applied to any given quality model. The proposed framework relates the functional psychometric outcome of quality assessment to a local weighting that can be used to improve the accuracy of quality estimation. A neural network-based method for estimating local weights is proposed and evaluated for the estimation of image quality.

In a third contribution, the concept of distortion sensitivity is transferred to rate-distortion theory for lossy compression. A perceptual bit allocation scheme for block-based video compression is derived and experimentally evaluated for compression of still images. Significant bit rate savings are achieved compared to the state of the art, at identical perceptual quality . However, the results suggest that the performance of data-driven quality models crucially depends on the availability of labeled training data. It is shown that, for generating such data, conventional psychophysical assessment of perceived quality inherently suffers from several flaws. Thus, in a fourth contribution, a neurophysiological quality assessment method based on steady-state visual evoked potentials is proposed. The proposed assessment method achieves significant correlations to conventionally obtained quality scores. Extracted neural markers of quality are statistically equivalent to MOS values in a linear prediction model. This paves the way towards novel neurophysiological methods for reliable assessment of visual quality.

Zusammenfassung

Wahrgenommene Qualität ist ein Schlüsselaspekt moderner Multimediakommunikationssysteme. Jedoch sind die Fragen, wie man zu verlässlichen Qualitätsurteilen kommt, wie man Qualität der menschlichen Wahrnehmung entsprechend computergestützt schätzen kann, und wie man computergestützte Modelle für Qualität in Multimediasystemen verwenden kann trotz langjähriger Forschung nicht zufriedenstellen geklärt. Die genannten Fragestellungen sind eng miteinander verknüpft, denn typischerweise werden Multimediasysteme auf der Basis von Qualitätsbewertungen durch Menschen evaluiert, die Ergebnisse solcher Qualitätsuntersuchungen werden für den Entwurf rechnergestützter Qualitätsmodelle genutzt und diese rechnergestützten Qualitätsmodelle dienen schließlich wiederum der Optimierung von Multimediasystemen.

Diese Dissertation trägt in mehrfacher Hinsicht zur aktuellen Forschung bei. Der erste Beitrag dieser Arbeit ist der Entwurf eines neuen, auf einem neuronalen Netz basierenden und Ende-zu-Ende optimierten rechnergestützten Models zur Bildqualitätsschätzung. Die Vorhersagegenauigkeit der vorgeschlagenen Methode übertrifft die anderer, dem Stand der Technik entsprechender Methoden, dies sowohl in Kenntnis als auch in Unkenntnis des Referenzbildes. Der zweite Beitrag der Arbeit ist die formale Definition von Störungsempfindlichkeit, die zur Herleitung eines rechnerisch vorteilhaften Qualitätsmodells führt. Der vorgeschlagene konzeptionelle Rahmen verknüpft die funktional-psychometrische Beschreibung von Qualitätswahrnehmung mit einer lokalen Gewichtung zur perzeptuellen Anpassung von gegebenen Qualitätsmodellen. Zur Schätzung lokaler Gewichte für die Qualitätsvorhersage wird eine auf einem neuronalen Netz basierende Methode vorgestellt und untersucht.

In einem dritten Beitrag wird das Konzept der Störungssensitivität auf die Rate-Verzerrungs-Theorie der verlustbehafteten Kompression übertragen. Ein Schema zur perzeptuellen Bitallokation in blockbasierter Videokompression wird hergeleitet und am Beispiel der Standbildkompression experimentell untersucht. Bei gleicher perzeptueller Qualität werden im Vergleich zum Stand der Technik deutliche Bitratengewinne gezeigt.

Dennoch legen die Ergebnisse nahe, dass die Leistungsfähigkeit von datengetriebenen Qualitätsmodellen maßgeblich von der Verfügbarkeit annotierter Trainingsdaten abhängt. Es wird jedoch gezeigt, dass konventionelle psychophysikalische Qualitätsbewertungsmethoden inhärente Nachteile aufweisen. Deshalb wird in einem vierten Beitrag eine elektroenzephalographische Qualitätsbewertungsmethode entwickelt, die auf zustandsstabilen visuell evozierten Potentialen beruht. Die vorgestellte Methode zeigt signifikante Korrelationen zu konventionell ermittelten Qualitätsurteilen. Extrahierte neurale Marker wahrgenommener

Qualität sind in einem einfachen linearen Prädiktionsmodell statistisch nicht von MOS-Werten zu unterscheiden. Dies zeigt einen Weg zu neuartigen neurophysiologischen Methoden für die Beurteilung von visueller Qualität auf.

Contents

List of Tables

List of Figures

1 Introduction

1.1 Overview

Multimedia services are an integral part of modern society and digital communication shapes the way we learn and teach, work and entertain, interact and cooperate. Services and applications such as video streaming, video conferencing, social media, and connected devices such as smart phones and tablets are ubiquitous, and emerging technologies, such as virtual reality (VR) and augmented reality (AR) are becoming more and more important. Video is a core modality of all of these multimedia applications and the amount of visual signals captured, stored, transmitted and consumed is tremendous: In 2013, about 660 billion pictures were taken, and this number almost doubled to 1.2 trillion pictures in 2017 and is predicted to grow even further [Statistica, 2017]; in 2016, images and videos accounted for 74% of all consumer internet traffic, and, with a growth of 31%, is predicted to rise to 82% of all consumer internet traffic [Cisco, 2017].

Because multimedia signals are consumed mostly by humans, visual quality is crucial to overall user satisfaction. However, for transmission and storage at bit rates suitable for today's channels and memory devices, these signals are digitized and potentially compressed. Compression algorithms achieve bitrate reduction by redundancy reduction and irrelevance reduction. While the former exploits statistical structures in the visual signal and is losslessly reversible by the receiver, the latter removes actual information from the signal, that cannot be recovered by the receiver. The information loss introduces distortions to the signal. If the removed information is in fact *not* irrelevant, these distortions become visible to humans. To efficiently control the trade-off between bit rate and distortion, it is crucial to measure the perceived impairments [Wiegand and Schwarz, 2016].
The measurement of perceived impairments, or, reversely termed: perceived quality, is not only essential for the operation of compression algorithms, but also for benchmarking and monitoring the performance of multimedia systems or parts thereof [Wang, 2011].

However, the seemingly easy problem of measuring perceived quality can present astonishingly difficult challenges. While individual humans are very fast and typically very confident

about their subjective quality judgment, the reliable and robust assessment of quality holds many pitfalls and computational estimation of quality as perceived by humans is surprisingly challenging. Despite decades of research on the topic, assessment methods are still debated, prediction approaches are far from being satisfactory and a definite quantitative model for quality is not forthcoming.

In this thesis, perceptual quality is investigated from three complementary perspectives and 1. novel data-driven methods for computational quality estimation are developed, 2. a new psychophysiological approach to quality assessment is proposed, and 3. a data-driven computational quality model is applied to perceptually optimized image compression. Under the premise that quality in multimedia is nothing more than perceptual quality, if not stated differently, the term *quality* will always denote *perceptual quality*.

1.2 Outline of the Thesis

The thesis is structured as follows:

Chapter 2 introduces the concept of quality in multimedia, describes psychophysical quality assessment and discusses its shortcomings. After a short introduction to psychophysiology, several measurement methods are introduced and motivated as a solution for overcoming these shortcomings.

Chapter 3 discusses computational approaches to the estimation of quality, introduces the concept of end-to-end data-driven quality assessment, and reviews the state-of-the-art quality estimation method. Image quality databases and performance metrics are presented.

Chapter 4 proposes novel end-to-end trained methods for full-reference (FR) and no-reference (NR) image quality prediction. The performance of the methods are improved by incorporating a jointly optimized spatially weighted pooling strategy. The superior performance of the proposed methods is benchmarked and compared to other state-of-the-art methods.

Chapter 5 introduces the concept of distortion sensitivity in quality estimation, derives a functional psychometrical definition and proposes a data-driven approach for its estimation. The resulting method for quality estimation is evaluated. Moreover, it is shown how the concept bridges from the psychometrical description of quality to perceptually efficient bit allocation in image and video compression. The superior performance of the proposed bit allocation scheme is shown in an image compression context.

Chapter 6 studies the use of steady-state visual evoked potential (SSVEP) for neurophysiological image quality assessment. For unsupervised extraction of meaningful spatial components from the electroencephalography (EEG) data, spatio-spectral decomposition (SSD) is reformulated in the frequency domain to allow for direct application to SSVEPs. A statistical screening method for rejecting unreliable subjects is proposed. The proposed methods show comparable results to psychophysical quality assessment.

Chapter 7 concludes the thesis with a summary and discussion of the presented results and gives an outlook on future work.

1.3 Publications

The following list contains all contributions broadly related to this thesis made by the author to the scientific literature in the fields of image and video compression, quality estimation and quality assessment.

Journal Articles

Bosse, S., Becker, S., Müller, K.-R., Samek, W., and Wiegand, T. (2018d). Estimation of distortion sensitivity for visual quality prediction using a convolutional neural network. *Digital Signal Processing, submitted*

Bosse, S., Arndt, S., Engelke, U., Martini, M., Ramzan, N., and Brunnström, K. (2018a). The VQEG testplan for psychophysiological video quality assessment. *Quality and User Experience, submitted*

Bosse, S., Maniry, D., Müller, K.-R., Wiegand, T., and Samek, W. (2018e). Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on Image Processing*, 27(1):206–219

Reisenhofer, R., Bosse, S., Kutyniok, G., and Wiegand, T. (2018). A Haar wavelet-based perceptual similarity index for image quality assessment. *Signal Processing: Image Communication*, 61:33–43

Bosse, S., Acqualagna, L., Samek, W., Porbadnigk, A. K., Curio, G., Blankertz, B., Müller, K.-R., and Wiegand, T. (2017a). Assessing perceived image quality using steady-state visual evoked potentials and spatio-spectral decomposition. *IEEE Transactions on Circuits and Systems for Video Technology*, 8215(c):1–1

Engelke, U., Darcy, D., Mulliken, G., Bosse, S., Martini, M., Arndt, S., Antons, J.-N., Chan, K., Ramzan, N., and Brunnström, K. (2017). Psychophysiology-Based QoE Assessment: A Survey. *IEEE Journal of Selected Topics in Signal Processing*, 11(1):6–21

Avarvand, F. S., Bosse, S., Müller, K.-R., Schäfer, R., Nolte, G., Wiegand, T., Curio, G., and Samek, W. (2017a). Objective quality assessment of stereoscopic images with vertical disparity using EEG. *Journal of Neural Engineering*, 14(4):046009

Acqualagna, L., Bosse, S., Porbadnigk, A. K., Curio, G., Müller, K.-R., Wiegand, T., and Blankertz, B. (2015). EEG-based classification of video quality perception using steady state visual evoked

potentials (SSVEPs). *Journal of Neural Engineering*, 12(2):026012 (**Shared First Authorship**)

Scholler, S., Bosse, S., Treder, M. S., Blankertz, B., Curio, G., Müller, K.-R., and Wiegand, T. (2012). Toward a direct measure of video quality perception using EEG. *IEEE Transactions on Image Processing*, 21(5):2619–29

Peer-Reviewed Contributions to Conferences

Bosse, S., Becker, S., Fisches, Z., Samek, W., and Wiegand, T. (2018c). Neural network-based estimation of distortion sensitivity for image quality prediction. In *Proceeding of the IEEE International Conference on Image Processing (ICIP), accepted for publication*

Bosse, S., Bagdasarian, M., Samek, W., Curio, G., and Wiegand, T. (2018b). On the stimulation frequency in SSVEP-based image quality assessment. In *10th International Conference on Quality of Multimedia Experience (QoMEX), accepted for publication*

Bosse, S., Siekmann, M., Samek, W., and Wiegand, T. (2017c). A perceptually relevant shearlet-based adaptation of the PSNR. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pages 315–319

Shahbazi, F., Bosse, S., Nolte, G., Wiegand, T., and Samek, W. (2017). Quality assessment of 3D visualizations with vertical disparity: An ERP approach. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, pages 4391–94

Avarvand, F. S., Bosse, S., Nolte, G., Wiegand, T., and Samek, W. (2017b). Measuring the quality of 3D visualizations using EEG: A time-frequency approach. In *Proceedings of the 7th Graz Brain-Computer Interface Conference*, pages 441–446. Verlag der TU Graz

Bosse, S., Siekmann, M., Rasch, J., Wiegand, T., and Samek, W. (2016e). Quality assessment of image patches distorted by image Compression using crowdsourcing. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6

Bosse, S., Maniry, D., Müller, K.-R. R., Wiegand, T., and Samek, W. (2016b). Neural network-based full-reference image quality assessment. In *Proceedings of the Picture Coding Symposium (PCS)*, pages 1–5

Bosse, S., Chen, Q., Siekmann, M., Samek, W., and Wiegand, T. (2016a). Shearlet-based reduced reference image quality assessment. In *Proceeding of the IEEE International Conference on Image Processing (ICIP)*, pages 2052–2056

Bosse, S., Maniry, D., Wiegand, T., and Samek, W. (2016c). A deep neural network for image quality assessment. In *Proceedings of the IEEE International Conference on Image Processing*

(ICIP), pages 3773–3777

Bosse, S., Müller, K.-R., Wiegand, T., and Samek, W. (2016d). Brain-computer interfacing for multimedia quality assessment. In *Proc. of the IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2834–2839

Dias, A. S., Siekmann, M., Bosse, S., Schwarz, H., Marpe, D., and Mrak, M. (2015b). Rate-distortion optimised quantisation for HEVC using spatial just noticeable distortion. In *Proceedings of the European Signal Processing Conference (EUSIPCO)*, pages 110–114

Dias, A., Schwarz, S., Siekmann, M., Bosse, S., Schwarz, H., Marpe, D., Zubrzycki, J., and Mrak, M. (2015a). Perceptually Optimised Video Compression. In *IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 1–4

Bosse, S., Acqualagna, L., Porbadnigk, A., Blankertz, B., Curio, G., Müller, K.-R., and Wiegand, T. (2014). Neurally informed assessment of perceived natural texture image quality. In *Proceedings of IEEE International Conference on Image Processing (ICIP)*, pages 1987–1991

Other Contributions to Conferences

Bosse, S., Acqualagna, L., Porbadnigk, A. K., Curio, G., Müller, K.-R., Blankertz, B., and Wiegand, T. (2015). Neurophysiological assessment of perceived image quality using steady-state visual evoked potentials. In *Applications of Digital Image Processing XXXVIII*, volume 9599, pages 959914–959914

Contributions to Standardization

Bosse, S., Helmrich, C., Schwarz, H., Marpe, D., and Wiegand, T. (2017b). Perceptually optimized QP adaptation and associated distortion measure. In *Joint Video Exploration Team (JVET) of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, JVET-H0047*, Macao, China

1.4 Additional Publications

For the sake of completeness, the following list contains additional contributions that are less related to this thesis.

Müller, K., Schwarz, H., Marpe, D., Bartnik, C., Bosse, S., Brust, H., Hinz, T., Lakshman, H., Merkle, P., Rhee, F. H., Tech, G., Winken, M., and Wiegand, T. (2013). 3D high-efficiency video coding for multi-view video and depth data. *IEEE Transactions on Image Processing*, 22(9):3366–3378

Marpe, D., Schwarz, H., Bosse, S., Bross, B., Helle, P., Hinz, T., Kirchhoffer, H., Lakshman, H., Nguyen, T., Oudin, S., Siekmann, M., Suhring, K., Winken, M., and Wiegand, T. (2010b). Video

compression using nested quadtree structures, leaf merging, and improved techniques for motion representation and entropy coding. *IEEE Transactions on Circuits and Systems for Video Technology*, 20(12):1676–1687

Peer-Reviewed Contributions to Conferences

Schwarz, H., Bartnik, C., Bosse, S., Brust, H., Hinz, T., Lakshman, H., Merkle, P., Müller, K., Rhee, H., Tech, G., Winken, M., Marpe, D., and Wiegand, T. (2012b). Extension of High Efficiency Video Coding (HEVC) for multiview video and depth data. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pages 205–208

Schwarz, H., Bartnik, C., Bosse, S., Brust, H., Hinz, T., Lakshman, H., Marpe, D., Merkle, P., Müller, K., Rhee, H., Tech, G., Winken, M., and Wiegand, T. (2012a). 3D video coding using advanced prediction, depth modeling, and encoder control methods. In *Picture Coding Symposium (PCS)*, pages 1–4

Bosse, S., Schwarz, H., Hinz, T., and Wiegand, T. (2012). Encoder control for renderable regions in high efficiency multiview video plus depth coding. In *Picture Coding Symposium (PCS)*, pages 129–132

Marpe, D., Schwarz, H., Wiegand, T., Bosse, S., Bross, B., Helle, P., Hinz, T., Kirchhoffer, H., Lakshman, H., Nguyen, T., Oudin, S., Siekmann, M., Sühring, K., and Winken, M. (2011). Improved video compression technology and the emerging high efficiency video coding standard. In *Proc. of the IEEE International Conference on Consumer Electronics (ICCE)*, pages 52–56

Winken, M., Marpe, D., Schwarz, H., Wiegand, T., Boße, S., Bross, B., Helle, P., Hinz, T., Kirchhoffer, H., Lakshman, H., Nguyen, T., Oudin, S., Siekmann, M., and Sühring, K. (2011). Highly efficient video coding based on quadtree structures, improved motion compensation, and probability interval partitioning entropy coding. In *Proc. of the ITG Conference on Electronic Media Technology, CEMT*

Siekmann, M., Bosse, S., Schwarz, H., and Wiegand, T. (2010). Separable Wiener filter based adaptive in-loop filter for video coding. In *Picture Coding Symposium (PCS)*, pages 70–73

Marpe, D., Schwarz, H., Bosse, S., Bross, B., Helle, P., Hinz, T., Kirchhoffer, H., Lakshman, H., Nguyen, T., Oudin, S., Siekmann, M., Sühring, K., Winken, M., and Wiegand, T. (2010a). Highly efficient video compression using quadtree structures and improved techniques for motion representation and entropy coding. In *Picture Coding Symposium (PCS)*, pages 206–209

Contributions to Standardization

Albrecht, M., Bartnik, C., Bosse, S., Brandenburg, J., Bross, B., Erfurt, J., George, V., Haase, P., Helle, P., Helmrich, C., Henkel, A., Hinz, T., de Luxan Hernandes, S., Kaltenstadler, S., Keydel,

P., Kirchhoffer, H., Lehmann, C., Lim, W.-Q., Ma, J., Maniry, D., Marpe, D., Merkle, P., Nguyen, T., Pfaff, J., Rasch, J., Rischke, R., Rudat, C., Schaefer, M., Schierl, T., Schwarz, H., Siekmann, M., Skupin, R., Stallenberger, B., Stegemann, J., Suehring, K., Tech, G., Venugopal, G., Walter, S., Wieckowiski, A., Wiegand, T., and Winken, M. (2018). Description of SDR, HDR and 360° video coding technology proposal by Fraunhofer HHI. In *Joint Video Exploration Team (JVET) of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, JVET-J0014*, San Diego, CA, USA

Schwarz, H., Bartnik, C., Bosse, S., Brust, H., Hinz, T., Lakshman, H., Marpe, D., Merke, P., Müller, K., Rhee, H., Tech, G., Winken, M., and Wiegand, T. (2011a). Description of 3D video coding technology proposal by Fraunhofer HHI (HEVC compatible configuration B). In *MPEG Meeting ISO/IEC JTC1/SC29/WG11, Doc. MPEG11/M22571*, Geneva, Switzerland

Schwarz, H., Bartnik, C., Bosse, S., Brust, H., Hinz, T., Lakshman, H., Marpe, D., Merke, P., Müller, K., Rhee, H., Tech, G., Winken, M., and Wiegand, T. (2011b). Description of 3D video coding technology proposal by Fraunhofer HHI (HEVC compatible configuration A),. In *MPEG Meeting ISO/IEC JTC1/SC29/WG11, Doc. MPEG11/M22570*, Geneva, Switzerland

Schwarz, H., Bartnik, C., Bosse, S., Brust, H., Hinz, T., Lakshman, H., Marpe, D., Merke, P., Müller, K., Rhee, H., Tech, G., Winken, M., and Wiegand, T. (2011c). Description of 3D video coding technology proposal by Fraunhofer HHI (MVC compatible). In *MPEG Meeting ISO/IEC JTC1/SC29/WG11, Doc. MPEG11/M22569*, Geneva, Switzerland

Stefanoski, N., Espinosa, P., Wang, O., Lang, M., Smolic, A., Bosse, S., Farre, M., Müller, K., Schwarz, H., Winken, M., and Wiegand, T. (2011). Description of 3D video coding technology proposal by Disney research Zurich and Fraunhofer HHI,. In *MPEG Meeting - ISO/IEC JTC1/SC29/WG11, Doc. MPEG11/M22668*, Geneva, Switzerland

Merke, P., Jäger, F., Bosse, S., and Müller, K. (2011). HEVC Anchors and Target Bit Rates for 3DV CfP. In *MPEG Meeting - ISO/IEC JTC1/SC29/WG11, Doc. MPEG11/M21217*, Turin, Italy

Winken, M., Bosse, S., Benjamin, B., Helle, P., Hinz, T., Kirchoffer, H., Lakshman, H., Marpe, D., Oudin, S., Preiss, M., Schwarz, H., Siekmann, M., Sühring, K., and Wiegand, T. (2010). Description of video coding technology proposal by Fraunhofer HHI. In *Joint Collaborative Team on Video Coding, JCTVC-A116*, Dresden, Germany

2 Perceptual Quality and Its Assessment

In most technical systems quality is typically and ultimately evaluated by humans. Hence, and as opposed to computational quality estimation, human quality assessment constitutes the ground truth of quality. This chapter discusses different aspects of quality assessment with emphasis on visual quality and visual quality assessment. Following a definition of quality, psychometric approaches of quality assessment are revisited; a critical discussion reveals several conceptual flaws and practical problems of these conventional approaches. A brief introduction into neurophysiology and EEG provides the foundation to survey the state of the art in psychophysiological quality assessment.

Several thoughts, examples and arguments in this chapter have been published earlier in

Bosse, S., Müller, K.-R., Wiegand, T., and Samek, W. (2016d). Brain-computer interfacing for multimedia quality assessment. In *Proc. of the IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2834–2839 ©2016 IEEE

and

Engelke, U., Darcy, D., Mulliken, G., Bosse, S., Martini, M., Arndt, S., Antons, J.-N., Chan, K., Ramzan, N., and Brunnström, K. (2017). Psychophysiology-Based QoE Assessment: A Survey. *IEEE Journal of Selected Topics in Signal Processing*, 11(1):6–21 ©2017 IEEE.

2.1 Perceptual Quality

The concept of quality is easy to understand but difficult to define[1]. In multimedia technology it is typically used to capture one aspect of the performance of a system or service. A widely used definition [Callet et al., 2012] specifies quality as *"[...] the outcome of an individual's comparison and judgment process [...] in terms of the evaluated excellence or goodness, of the degree of need fulfillment, and in terms of a 'quality event' "*, where a *quality event* is *"[a]n observable occurance"*, e.g. a multimedia signal. Factors driving this judgment process are

[1] "I know it when I see it"–Potter Stewart.

manifold and not all of them, such as situational or socio-cultural contexts [Garcia et al., 2014], can be controlled for by means of technology. However, other factors have a clear relation with the *quality event*. For visual quality, e.g., resolution, frame rate or visibility and compression artifacts play a crucial role [Garcia et al., 2014]. Quality of multimodal, e.g. audiovisual, signals is strongly related to the independent qualities of the individual monomodal signals but less so to factorial multimodal interactions [Hands, 2004, You et al., 2010]. The outcome of this judgment process is not binary, as humans can conceive of quality as a gradual quantity (the *goodness* of the previously quoted definition) and differentiate different levels of quality.

Obviously, quality is a perceptual quantity [Garcia et al., 2014], and the outcome of sensorial, perceptual and cognitive processes. These comprise low-level and high-level processing, such as contrast sensitivity or attention for visual signals [Palmer, 1999]. However, quality perception is a conscious experience made on the basis of these preceding sensorial and cognitive processes and that results in a judgment on a signal. With the exception of extreme cases where impairments might destroy semantic information, this judgment does not carry any information about the content of the signal itself and in natural viewing situations quality is typically not explicitly attended to. The precise interactions between different perceptual processes are still unclear [Palmer, 1999] and it is even more unknown how exactly these processes lead to the perception of quality in the case of multimodal stimuli. Moreover, the internal reference used for comparison when arriving at a quality judgement is still unknown. However, despite these conceptual difficulties, quality needs to be quantified for the evaluation of technical systems or services. The next section outlines psychophysical approaches to quality assessment that are widely used.

2.2 Psychophysical Assessment of Perceptual Quality

Due to the lack of satisfactory models for perception and quality formation the reliable assessment of perceptual quality builds on psychophysical judgement tests. In these tests a human observer is presented with a signal and is asked to give an overt judgement response based on the quality of the presented signal. The stimulus conditions presented in quality tests are defined by the types of impairment, the magnitude of impairment and the source reference signal. The magnitude of impairment is typically controlled for by an objective parameter, such as e.g. the quantization parameter (QP), the bitrate or the peak signal-to-noise ratio (PSNR) for the evaluation of video compression schemes.
In order to allow for meaningful and reproducible test results, procedures for psychophysical quality assessment are defined by various recommendations of the International Telecommunication Union (ITU) for different signal modalities and systems. Two prominent and widely referenced examples of these recommendations are are concerned with the assessment of visual quality [ITU-R Rec. BT.500-13, 2012, ITU-T Rec. P.910, 2008]. These recommendations prescribe viewing conditions (comprising objective parameters such as background chromaticity, ratios of between the luminance of the inactive screen and signal's peak luminance, and viewing distance), test methods and data processing for psychophysical quality tests.

2.2.1 Test Procedures

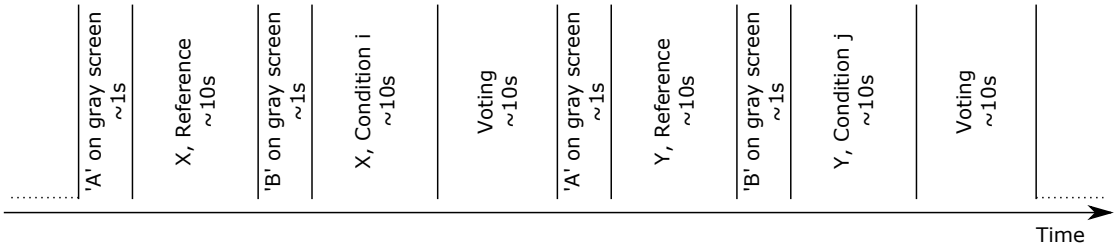


Figure 2.1 – Temporal structure of the DCR test procedure for stimulus X under condition i and stimulus Y under condition j.

Regardless of the test procedure used, psychophysical quality assessment tests should not take more than 30 minutes per participant in order to avoid increasing unreliability of the assessors due to increasing weariness [ITU-R Rec. BT.500-13, 2012]. The first conditions presented to the observer should be chosen to be representative of the test material in order to make the participant familiar with rating scale, task and range of distortions. Ratings collected during this training/calibration phase are not considered for further analysis.
Generally, two categories of assessment procedures, namely double stimulus assessment and single stimulus assessment, can be distinguished. In double stimulus assessment, such as Degradation Category Rating (DCR) [ITU-T Rec. P.910, 2008] (also known as Double Stimulus Impairment Scale (DSIS) [ITU-R Rec. BT.500-13, 2012]) or Double Stimulus Continuous Quality Scale (DSCQS) [ITU-R Rec. BT.500-13, 2012], the participant is presented with a test condition (e.g. a distorted version of reference video or image) in relation to its reference condition. Reference and test conditions can be presented consecutively or side by side, the latter referred to as Simultaneous Presentation (SP).
In the DCR case, see Fig. 2.1 for a sketch of the temporal structure, the test participant is informed about which of conditions is the test and which is the reference condition and is asked to report their quality judgement after both conditions were presented. Observers report their judgment in terms of impairment of the test condition with respect to the reference condition on a categorical scale that is semantically annotated as 1-*Very annoying*, 2-*Annoying*, 3-*Slightly annoying*, 4-*Perceptible but not annoying* and 5-*Imperceptible*. In the DSCQS case the presentation order (or left/right hand sided position on screen for SP) of reference and test condition is randomized and an assessor is asked for a quality judgment for both stimuli on a continuous rating scale. In contrast to ratings given in DCR, not the absolute value of the ratings, but difference between them is of interest and used for further processing.
In single stimulus assessment methods, e.g. Single Stimulus (SS) [ITU-R Rec. BT.500-13, 2012] or Absolute Cateogry Rating (ACR) [ITU-T Rec. P.910, 2008], the quality of the test condition is assessed by the participant without comparison to a reference condition. In ACR, the opinion scores are given with respect to the absolute quality and the rating scale carries semantic annotations as 1-*Bad*, 2-*Poor*, 3-*Fair*, 4-*Good* and 5-*Excellent*. In order to allow the test participants to report their judgments with finer granularity, categorical grades are sometimes scaled to 9 points [ITU-T Rec. P.910, 2008]; grades 2, 4, 6, 8 are not assigned semantic annotations then.

However, studies show that differences between finer and coarser gradings are insignificant [Huynh-Thu et al., 2011]. Double and single stimulus assessment can be conducted with or without presenting a Hidden Reference (HR) as a test condition in order to detect inconsistencies or to allow for the calculation of differential quality scores (see Section 2.2.2).
More exhaustive methods ask the participant for comparative ratings of all possible pair-wise combinations of conditions. This is referred to as Paired Comparison (PC). For some applications, such as surveillance or medical systems, also utility or performance-based grading schemes are studied and defined [Knoche et al., 1999, ITU-T Rec. P.912, 2016].
Despite the efforts of (pre-)standardization bodies such as ITU-T Study Group (SG) 9 and SG 12, ITU-R SG 6, and the Video Quality Expert Group (VQEG) to provide rational guidelines for psychophysiological quality assessment we will see in in Chapter 3 that these recommendations are not as widely acted on in detail as one would hope for.

2.2.2 Data Analysis

In order to arrive at a statistically robust and reliable quantification of perceptual quality, ratings need to be collected for more than one subject. Recommendations on the number of participants in psychophysical quality studies vary within quite a large range of 6 to 40 [ITU-T Rec. P.911, 1998]. At least 15 observers are recommended by [ITU-R Rec. BT.500-13, 2012], but in a cross-lab study [Pinson et al., 2012] a number of 24 test participants is determined for achieving statistical consistency across different labs. Collected quality ratings are screened in order to identify unreliable subjects that are rejected based on a 'hard' outlier detection [ITU-R Rec. BT.500-13, 2012]. More sophisticated methods that are better able to model subjects biases and inconsistencies are currently under investigation [Li and Bampis, 2017].
The final quality value assigned to each condition identified by its source reference k and distortion type/level j is then reported as the (in)famous mean opinion score (MOS) [ITU-R Rec. BT.500-13, 2012]

$$\mathrm{MOS}_{kj} = \frac{1}{I}\sum_{i=0}^{I-1} \mathrm{OS}_{ikj}, \tag{2.1}$$

the average over the subject- and condition-wise individual opinion scores OS_{ikj} of all I subjects after screening. Typically, the reliability of the MOS is quantified based on the 95% confidence interval assuming Gaussianity of the subject-wise ratings per condition,

$$\mathrm{CI}_{kj} = 1.96\frac{\sigma_{kj}}{\sqrt{I}} \tag{2.2}$$

with σ_{kj} being the standard deviation of the condition denoted by k and j across all subjects. If ratings for the reference conditions are available as well, e.g. in an HR setup, instead of the

MOS the differential mean opinion scores (DMOS) may be reported

$$\mathrm{DMOS}_{kj} = \mathrm{MOS}_{k,ref} - \mathrm{MOS}_{kj} \tag{2.3}$$

2.2.3 Critical Acclaim

Psychophysical test procedures are relatively easy to implement and to conduct and are therefore widely used. The MOS is the de-facto metric for the quantification of perceptual quality of multimedia signals and the outcome of psychophysical tests is considered as ground truth when it comes to evaluation, validation or benchmarking of multimedia systems or services [Streijl et al., 2016].
Despite its popularity and common usage, psychophysical tests in general and the MOS in particular suffer from several inherent flaws. Most fundamentally, psychophysical assessment methods, rely on assessors' subjective introspection, judgement process and subsequent overt conscious responses; hence, they provide only limited, if any, insight into the internal perceptual and cognitive processing underlying the decision making in quality perception. Another conceptual issue is that the explicit task of giving a judgement response has a disrupting effect because it, as a matter of principle, interferes with natural viewing behaviour, and as such, with natural viewing experience (informally sometimes referred to as "Schrödinger's cat of quality assessment"). This challenges the conclusions that can be drawn from psychophysical quality tests for real-world applications, in which, as discussed previously, quality is typically attended to only latently. In immersive media, e.g. VR this mismatch between quality assessment and application may even be more severe as the question 'are you immersed?' potentially breaks immersiveness [Slater, 2004].
Another limitation of psychophysical approaches to multimedia quality assessment is the restriction to supra-threshold stimuli (consciously detectable stimuli) and the insensitivity to sub-threshold stimuli (consciously undetectable stimuli). This is critical when it comes to short-term assessment of non-instantaneously perceivable phenomena such as visual fatigue or nausea [Urvoy et al., 2013].

In addition to these considerations concerning psychophysical assessment in general, there are several specific problems with the use of MOS as a measure for perceptual quality. Quality ratings are collected per subject and condition employing Likert-type rating scales [Likert, 1932]. However, these rating scales do not conform to the laws of fundamental measurements as known from natural sciences [Riskey, 1986]. Thus, Likert-type categorical rating scales do not quantify absolute metrics but are influenced by situational and contextual factors, such as the range of stimuli, the stimulus frequency, or the number of categories [Riskey, 1986]. Moreover, responses from categorical rating scales should be considered as ordinal data, not interval data [Stevens, 1946]. Therefore it is doubtful that the used scale intervals coincide with the corresponding semantic intervals [Knoche et al., 1999]. This is even less clear across language or cultural boarders.
Critically for the MOS, typically reported summary statistics such as the mean and the standard

deviation are inappropriate for ordinal data [Stevens, 1946]. This might compromise the validity of conclusions drawn from psychophysical quality tests reporting the MOS and/or Standard Deviation of Opinion Scores (SOS) [Hoßfeld et al., 2011].

Even pragmatically accepting this conceptual dubiousness of the MOS for the quantification of perceptual quality, the experimenter is left with several practical complications. The variance of quality judgements across subjects requires the recruitment of sufficient participants to achieve adequately small confidence intervals of the MOS. The variance stems from the label noise introduced by inconsistencies of the subjects ratings. These inconsistency might occur between subjects, e.g. due to different expectations, decision strategies, experiences or differences in attentiveness, [Pinson et al., 2012, Janowski and Pinson, 2015, Li and Bampis, 2017] but also for individual subjects, e.g. due to a general lack of attentiveness or weariness increasing over the course of the assessment session [Janowski and Pinson, 2015]. Further, the variance does not only depend on the participants, but also on the content of the source reference material considered in the test [Winkler, 2014]. In order to cope with this variance, at least 15 participants are prescribed for visual quality assessment [ITU-R Rec. BT.500-13, 2012] (other authors recommend 24 participants [Pinson et al., 2012]), but the precise number depends on the target confidence interval and is usually not known a-priori [Winkler, 2014]. Another problem is the time limitation of psychophysical assessment: "Look at a stimulus — form a quality rating — report it — repeat" is a surprisingly exhausting exercise. In order to prevent subjects from reporting (overly) unreliable quality judgements due to weariness (and by this to further increase the variance of the ratings), psychophysical tests should not take longer than 30 minutes [ITU-R Rec. BT.500-13, 2012]. This time constraint severely limits the number of conditions that can be presented as the process of reporting an overt response consumes time as well. A third problem for the experimenter employing the MOS is related to previously discussed mismatch between quality assessment and real-world media reception: Individual quality ratings are typically reported *after* the presentation of a stimulus. With regard to time, a given rating thus does not represent a differential, but rather an integral quantity. This renders the MOS unfeasible for assessing momentary values of quality. Note that a simplistic solution such as a continuously operated quality slider [Garcia Freitas et al., 2015] would aggravate the antagonism between the task of quality assessment and a natural viewing experience.

Quality focussed design of technical systems brings another perspective on practical limitations of psychophysical quality assessment: For benchmarking, evaluating or comparing multimedia systems or services, psychophysical methods are cumbersome, tedious and expensive. For designing systems psychophysical quality assessment does not provide models of quality that are actionable. This becomes even more critical when an optimization criterion is moved to the heart of an algorithm, as it is the case for bit allocation in image or video coding [Wiegand and Schwarz, 2016]. Psychophysical (and also in the next section discussed psychophysiological) quality assessment is of no help here, even if a human were somehow integrated into any process of in-loop optimization the judgement response would just be too slow. Computational and potentially real-time capable approaches to quality estimation will

be discussed in Chapter 3.

2.3 Psychophysiological Quality Assessment

In order to overcome the previously discussed limitations of psychophysical quality assessment and to objectify quality assessment, researchers started to study the application of methods from psychophysiology for quality assessment. This section gives a brief introduction into the fields of psychophysiology and EEG and surveys the state of the art of psychophysiological quality assessment.

2.3.1 Psychophysiology Background

Psychophysiology is concerned with the identification and measurement of physiological bases and correlates of psychological processes and subjective experiences [Cacioppo et al., 2016]. Some psychophysiological measurements can be directly or indirectly interpreted with regard to processes underlying the perception of quality and relevant signals may be related to the central nervous system (CNS), autonomous nervous system (ANS) or eye movements [Cacioppo et al., 2016]. The sympathetic division of the ANS and its 'fight or flight' response [Jansen et al., 1995] can be quantified by using e.g. electrocardiography (ECG) (relating to excitement or fatigue [Cacioppo et al., 2016]), galvanic skin response (GSR) (relating to arousal [Cacioppo et al., 2016]) and pupil diameter. Pupil diameter also belongs to the class of eye measurements that were shown to have a relationship to cognition [Cacioppo et al., 2016]. Eye movements, for instance, provide valuable insight into overt visual attention [Yarbus, 1967], eye blink rates relate to visual fatigue [Bang et al., 2014], and pupil dilation to cognitive load [Hess and Polt, 1964] and long-term memory processes [Kafkas, 2012]. Neurophysiological methods are targeted specifically to the CNS and provide insight into the neural underpinnings of sensorial, perceptual and cognitive processing in the brain. *Inner psychophysics* as a neural foundation of *outer psychophysics* were postulated already 1907 by Gustav Theodor Fechner [Fechner, 1907]. While the neural code and its underlying computational mechanisms are not yet well understood, many features encoded at different levels of the sensory pathway have been mapped out with increasing detail. For instance, neuronal ensembles at ascending levels of the visual pathway are tuned to progressively larger and more complex visual features [Palmer, 1999, Wandell, 1995]. Neural correlates of perceptual decision-making, attention and features thereof have been researched and identified in various experiments [Dmochowski and Norcia, 2015, Kohler et al., 2018].

Many of these well-established neurophysiological signatures are particularly promising among psychophysiological signals for assessing quality, as most multimedia signals interface directly with the sensory pathway. Ultimately, neurophysiological approaches could enable researchers and engineers to directly read out the quality related neural response, be it sensory, perceptual or cognitive, of an individual to a multimedia signal as a quantifiable metric. As sketched in Fig. 2.2, such a measurement would avoid many of the problems of psychophysical

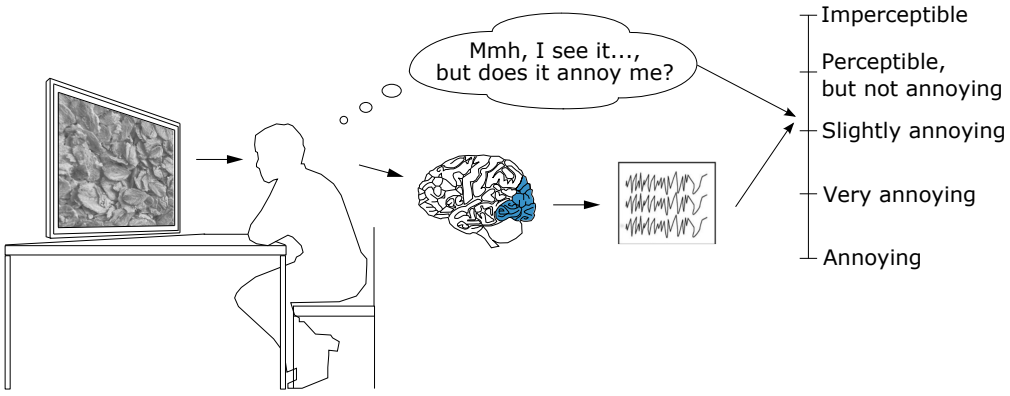


Figure 2.2 – Rating scales do not necessarily reflect internal experience. Neurophysiological measurements such as EEG could bypass the need for overt categorical responses. ©2016 IEEE

assessment methods outlined in the previous section by bypassing the need of an overt response.

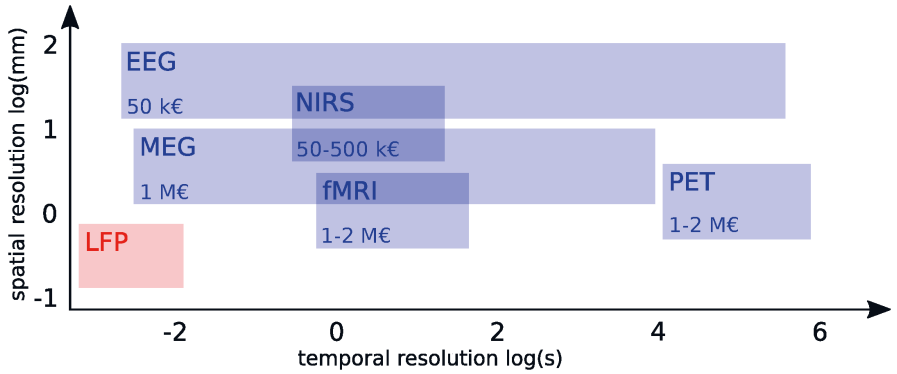


Figure 2.3 – Comparison of temporal and spatial resolution and acquisition costs of selected neuroimaging methods. Red shadings denote invasive methods, blue indicates non-invasive methods.

Different means of measuring neurophysiological signals are available exploiting different physiological mechanisms. Ethical and proportionality considerations rule out the use of invasive methods such as electrocorticography (ECoG) or local field potential (LFP). The same holds for positron emission tomography (PET), that, although being non-invasive, involves exposure to ionizing radiation. Near-infrared spectroscopy (NIRS), functional magnetic resonance imaging (fMRI) and PET measure the hemodynamic response, changes in blood flow in active brain areas. Although providing a comparably high spatial resolution, hemodynamic measures are limited in terms of temporal resolution. Note that especially fMRI and PET are also very costly in acquisition and operation. EEG and magnetoencephalography (MEG) pick up changes in voltage potentials or changes in the magnetic fields, respectively, on the skull, resulting from displacement of electrical charges in neural populations due to neural activity. While EEG and MEG provide excellent temporal resolution, its spatial resolution is limited.

Although MEG has some advantages over EEG [Lopes da Silva, 2013], until recently [Boto et al., 2018], its operation required a magnetically shielded room and devices are bulky, stationary and costly. A comparative summary of different neurophysiological measurements in terms of invasiveness, spatial and temporal resolution and cost is shown in Fig. 2.3. Due to its low level of invasiveness, its relative modest cost in acquisition and operation, its adequately high temporal resolution and recent advances in non-invasive EEG-based brain computer interfacing (BCI) [Lee et al., 2015, Müller et al., 2008] EEG is practically the most appropriate and promising method for psychophysiological quality assessment.

2.3.2 Electroencephalography

2.3.2.1 Physiological and Psychological Processes Underlying the EEG

The brain contains around 86 billion neurons [Herculano-Houzel, 2009]. Neurons are electrically excitable cells, consisting of the cell body called soma, cellular input extensions called dendrites, a cable-like output extension called axon, and endings of the axon connecting to the dendrites of other cells called axon terminals [Blum and Rutkove, 2007]. Neurons communicate by neurotransmission. A typical neuron has a resting potential of approximately $-70\,\mathrm{mV}$ across the cell membrane [Blum and Rutkove, 2007]. Voltage-gated ion channels control the exchange of ions with the extracellular milieu. If the excitatory drive integrated over the dendrites sufficiently depolarizes the membrane potential over a threshold of around $-55\,\mathrm{mV}$, sodium ion channels open and sodium ions flow into the cell. This nonlinear process produces a rapid rise and fall in the membrane potential, called an action potential [Blum and Rutkove, 2007]. The action potential travels along the neuron's axon and contributes to the depolarization of subsequent neuron's membrane potential, potentially triggering the rise of another action potential [Blum and Rutkove, 2007]. At the synaptic junction in the axon terminals, positively or negatively charged ions travel into the post- or presynaptic terminal, resulting in excitatory or inhibitory postynaptic potential (EPSP or IPSP), respectively [Nunez and Srinivasan, 2006]. EPSP and IPSP induce an extracellular current into the opposing direction. If strong enough, the temporal and spatial accumulation of EPSP and IPSP currents give rise to potential differences between two locations on the scalp that can be measured over time as the *electroencephalographic* signal [Nunez and Srinivasan, 2006].

Electrophysiological measurements have been studied for many decades with Hans Berger credited as the first to record EEG in humans, starting in 1924 [Berger, 1933]. While the precise links between EEG and psychological phenomena are still only incompletely understood, several relations are widely accepted. For example, power modulations in different frequency bands of the EEG signal covary with cognitive states, e.g. the θ-band (4 Hz to 8 Hz) is associated to attention and drowsiness, and the α-band (8 Hz to 12 Hz) is related to alertness, relaxation and fatigue [Ward, 2003]. α-waves were also among the first oscillations reported [Berger, 1933].
Whereas neural oscillations are temporally rather non-localized, event-related potentials

(ERPs) are large scale electrical events that consist of stereotypic changes in electrical activity that are time-locked with sensory stimuli and related cognitive events [Luck, 2014]. The relations of ERPs to high- and low-level sensory and cognitive processing have been thoroughly characterised [Handy, 2005, Luck, 2014]. They are characterized by their time-dependent amplitude according to a common nomenclature, with the first letter referring to the polarity of a particular component and subsequent number(s) indicating latency (in ms) or relative position in the order of components [Luck, 2014]. I.e. the well-known P300 component exhibits a positive peak around 300 ms after stimulus onset [Polich, 2012], but the latency can be significantly higher and reach more than 300 ms. The amplitude and delay of its subcomponent P3b is known to increase with decreased expectation of a stimulus, thus indicating the novelty of a task related stimulus. The amplitude P3a is related to attentional shifts and task unrelated novelty [Polich, 2009]. Other ERPs have been shown to be involved with object representation and memory operations in a variety of task behaviours [Handy, 2005, Luck and Kappenman, 2011]. ERPs are analysed in the time-domain and characterized by peak amplitude and latency. Given its temporal isolation and the need for a baseline reference, in ERP studies stimuli are separated from each other by relatively long interstimulus intervals [Luck, 2014].

Another type of neurophysiological response to temporally isolated visual stimuli is the SSVEP [Regan, 1977, Norcia et al., 2015]. While transient ERPs are typically observed in response to surprising or novel stimuli, SSVEPs are observed in response to sustained, periodic stimuli. Periodic stimulation results in increased narrowband EEG spectral power at the stimulation frequency and its harmonics (integer multiples of the stimulation frequency) [Regan and Regan, 1988], and can be, as suggested by the name, very stable in amplitude and phase [Regan, 1966]. SSVEPs are typically defined by their amplitude, phase and spatial channel distribution for the tagged frequency and its associated harmonics [Norcia et al., 2015]. This makes it natural to analyse SSVEP in the frequency domain rather than in the time domain [Regan and Regan, 1988]. While the amplitude of the SSVEP is related to the magnitude of the perceptual response, the phase is related to processing delays [Norcia et al., 2015]. As the signal is only contained in the response components harmonically related to the stimulation frequency, the noise in a recording can be easily estimated by the amplitude of the response component in the frequency bins neighbored to harmonics of the stimulation frequency [Meigen and Bach, 1999]. Responses in real EEG recordings are contaminated by noise. However, the fact that the response itself is narrowband, while noise sources are broadband explains the reported high signal-to-noise-ratio (SNR) of SSVEP recordings, relative to broadband ERP-responses [Norcia et al., 2015, Meigen and Bach, 1999]. While traditionally the SSVEP was considered to be related to sensory processes and low-level vision [Regan, 1989], it can also be used to study visual processes of higher level, i.e. motion [Ales and Norcia, 2009], face [Alonso-Prieto et al., 2013], object [Farzin et al., 2012] perception or multi-sensory integration [Regan et al., 1995] and can be used for BCI [Won et al., 2015, Kwak et al., 2015, Müller-Putz et al., 2006].

The signal-to-noise ratio (SNR) of EEG signals is typically very low. Amplitudes of evoked potentials are very small laying in the range of several microvolts, buried in the EEG background activity with an amplitude range of tens of microvolts making single trial quality assessment

difficult. As evoked potentials are time locked with the stimulus onset they can be resolved against the background activity and other types of non-phase locked noise by averaging the recorded signals across several trials [Handy, 2005, Blankertz et al., 2011].

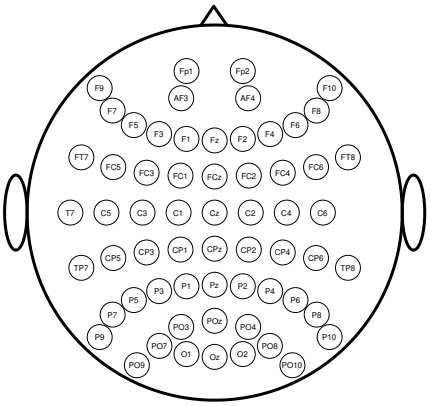


Figure 2.4 – Electrode locations in the scalp as specified by the international 10-20 system [Sharbrough et al., 1991]. Odd numbers denote electrodes on the left hemisphere, even numbers denote electrodes on the right hemisphere. Underlying cortical regions are represented by the prefix of the electrode names: F is frontal, P is parietal, O is occipital, T is temporal, C is central.

EEG signals are typically recorded simultaneously at different scalp locations. Fig. 2.4 shows electrode position for a 64 channel 10-20 system [Sharbrough et al., 1991]. The interpretation of the EEG signal with regard to the involvement of neural populations of different cortical regions is usually based on the spatial distribution of the activity of the recorded signal. For computational reasons, the dimensionality of recorded EEG data is commonly reduced for further analysis or processing of the EEG signal. A still widely used approach for dimensionality reduction is the selection of specific subset of channels. However, this approach has the disadvantage that it assumes relevant information only in the vicinity of the selected channel. Channel selection moreover neglects subject-wise differences in anatomy and recording-wise differences in channel positions (i.e. due to slight cap misalignments). Further, it relies on explicit neuroanatomical assumptions.
More recent methods from the field of BCI apply a combined analysis of signals measured at different scalp locations using spatial filters. These filters project the recorded data from the sensor-space to a subspace of reduced dimensionality determined by an optimization criterion that enhances signal recovery [Blankertz et al., 2011, Haufe et al., 2014b]. Linear filters also improve the interpretability of spatial activity distributions [Haufe et al., 2014a] and extracted components serve as rationally 'selected' channels.

2.3.3 State of the Art

As the previous section showed, the type of desired EEG responses and the investigated signal modality will invariably guide experimental design and data analysis.

Audio Quality Audio signals were among the first modalities studied in the context of EEG-based quality assessment with ERPs being used as a quantitative measure for perceived quality of audio signals [Porbadnigk et al., 2010, Antons et al., 2010]. As stimulus, a phoneme /a/ at varying quality levels was presented to the subjects for 160 ms. In an oddball paradigm it is shown that with decreasing quality of the stimulus the latency of the P300 is decreasing while its amplitude is increasing. By applying a classifier based on shrinkage linear discriminant analysis (LDA) [Blankertz et al., 2011, Blankertz et al., 2008], distortions below the threshold of conscious perception are detected for 2 out of 11 subjects [Porbadnigk et al., 2010]. This work is extended to longer and more realistic auditory stimuli of lengths of 200, 1200 and 8000 ms (phoneme, word, sentence) subject to transmission distortions and similar effects as for phoneme length stimuli are found [Antons et al., 2012b]. Due to the low SNR of EEG signals, commonly a lot of trials have to be collected. Single trial methods were explored for the analysis of neural correlates of speech quality and the detection of perceived distortions [Porbadnigk et al., 2013]. It is shown that for auditory stimuli components reflecting the perception of distortions are assessable in early components already. Further, by combining behavioral and neural data, the sensitivity of the experiment is significantly increased. The influence of low-quality audio signals on fatigue was studied [Antons et al., 2012a], where α- and θ-activity is shown to be increased for audio signals subject to distortions introduced by bandwidth limitations in transmission systems.

Similar approaches to those in speech quality have been studied for video quality assessment. The brain response to JPEG compression artifacts in images was studied based on an oddball paradigm [Lindemann and Magnor, 2011]. As shown for speech signals, the elicited ERP component is reduced in latency and increased in amplitude for decreasing visual quality. In follow-up studies, similar results are reported for different kinds of artifacts in video clips [Mustafa et al., 2012b]. Using principal component analysis (PCA) for dimensionality reduction and support vector machine (SVM), a classification accuracy of 76.5% for the most obvious and of 73.5% for the less obvious distortions is reported for trials correctly detected behaviorally [Lindemann et al., 2011]. For different types of distortion, mean single-trials classification accuracy of up to 85% for distorted vs. undistorted images is achieved in [Mustafa et al., 2012a] using a wavelet-based approach. Neural correlates of coding distortions are studied in an oddball paradigm [Scholler et al., 2012], where recorded EEG data is filtered by an LDA filter [Blankertz et al., 2011, Blankertz et al., 2008]. Filter weights are obtained based on signed biserial correlation coefficients between trials with highest distortion and no distortion. For distortion magnitudes above the behavioral perception threshold, an area under the curve (AUC) close to 1 is obtained. Although the classifications accuracies in most studies [Scholler et al., 2012, Mustafa et al., 2012a, Lindemann et al., 2011] refer only to trials correctly classified

behaviorally, also classification accuracies above chance are reported for trials with a quality degradation below the behavioral perception threshold [Scholler et al., 2012]. As for speech signals [Porbadnigk et al., 2010], this suggests the potential of EEG to assess also subconscious processing of distortions in multimedia signals.
Similar results are reported for degradations introduced as changes in color saturation and changes in maximum luminance values of images [He et al., 2016].
Moon et al. [Moon and Lee, 2015] investigated the perception of high dynamic range (HDR) and standard dynamic range (SDR) videos and showed a significant power difference in the γ band. Viewing HDR content results in an increased correlation of power in the α band with the median luminance than viewing SDR content [Darcy et al., 2016]. The authors conclude a higher level of engagement towards HDR content.
Stereoscopy adds further quality related components to visual signals, as the perceived quality may be affected by crosstalk, misalignment of stereo image pairs, or the accommodation-vergence conflict [Urvoy et al., 2013]. The neural workload imposed to the viewer may result in visual discomfort or fatigue that might not become conscious within an instant [Urvoy et al., 2013]. The relation between the power in different oscillatory neural bands and MOS values for 3D videos subject to compression artifacts at two quality level was studied in a single channel analysis with reported correlations of $|r| \approx 0.25$ [Kroupi et al., 2014]. EEG recordings showed high frontal asymmetry in the α band, which reflected emotional affect towards the two different quality levels. Exploratory studies on the visual discomfort show a relation to changes in band power [Chen et al., 2013, Chen et al., 2014] and ERP [Li et al., 2008, Cho et al., 2012]. Using ERPs and spatial filters for feature extraction and shrinkage LDA for classification, a mean classification accuracy of 63.3% is reported for the neural classification of comfort zone in 3D viewing positions [Frey et al., 2015]. Other authors [Avarvand et al., 2017a] show a relation between the amplitude of the P1 component and vertical misalignment of stereo images. The influence of the shutter frequency of shutter glasses on the neural workload of the viewer is evaluated in [Wenzel et al., 2016]. Neural correlates of the flicker introduced by the opening and closing of the shutter glasses could be identified up to a frequency of 67.2 Hz, well above the behaviorally estimated flicker fusion threshold at 47.4 Hz. It is concluded that the risk of reduction in quality of experience and usability can be reduced by using higher frequencies for shutter glasses. Note that this study is conceptually different from the others, as no neural correlate of quality, but neural markers of flicker fusion are studied.

Multimodal Quality Donley et al. [Donley et al., 2015] studied the impact of various levels of synchrony of wind, vibration and light on audio-visual sequences. P300 amplitude and audiovisual quality were shown to be significantly correlated [Arndt et al., 2014a]. In an investigation of the effect of changes in audio and video quality using EEG and eye tracking parameters and conclude that α activity was correlated to video quality [Arndt et al., 2014b].

In summary, previous studies show promising results for the assessment of quality using EEG. Given the novelty of neurophysiological approaches for quality assessment, most of the work is exploratory, mainly reporting experimental designs and resulting correlations

between neural signals and behavioral responses. It is not clear yet how to interpret neural signals quantitatively in terms of acceptability of an impairment; prediction models have not been proposed nor evaluated yet. Rational methods for selecting and/or extracting the quality-related signal from the EEG channels [Blankertz et al., 2008, Müller et al., 2008] were studied and used only by few authors [Scholler et al., 2012, Porbadnigk et al., 2013, Porbadnigk et al., 2011]. Approaches studied so far do not use identical stimuli (even if the same modality is studied) making it hard to compare different methods, thus studies stand mostly for themselves. This is especially unfortunate as the space of experimental parameters, including experimental paradigm, temporal and spatial filtering, dimensionality reduction, outlier rejection, choice of recording, is large and difficult to sample without comparable stimuli. However, the Psychophysiological Quality Assessment (PsyPhyQA) project of VQEG, chaired by the author of this thesis, is developing a testplan comprising a publicly available set of defined stimuli in order to overcome this problem.

2.4 Lessons Learned

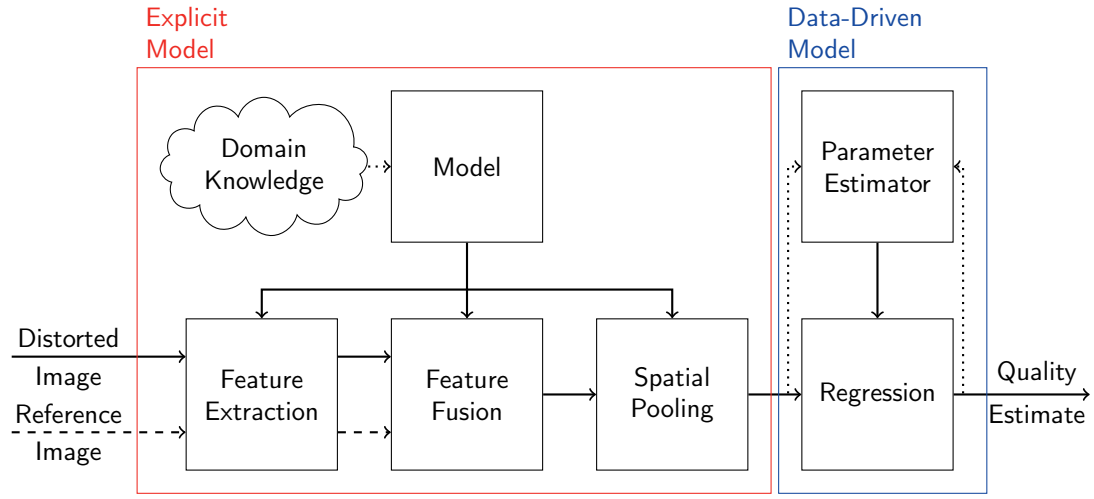
- The formation process of quality judgments is still unknown.
- Reliable quantification of perceived quality in multimedia relies on assessment by humans.
- Psychophysical quality assessment does not provide insights into the internal processes of quality formation.
- Psychophysical quality assessment is flawed by several factors such as subjective and contextual biases, lacking objectiveness and conceptual inconsistencies between individual ratings and pooled quantification of quality.
- Modern neuroimaging techniques are able to reveal and objectify the neural underpinnings of subjective experiences; for the field of quality assessment, EEG is an appropriate technique.
- ERPs and spectral power of EEG have been studied exploratory for quality assessment.
- Rational channel selection and feature extraction methods are not well studied for psychophysiological quality assessment.
- Lack of a common set of stimuli renders comparison of approaches difficult.

3 Computational Estimation of Visual Quality

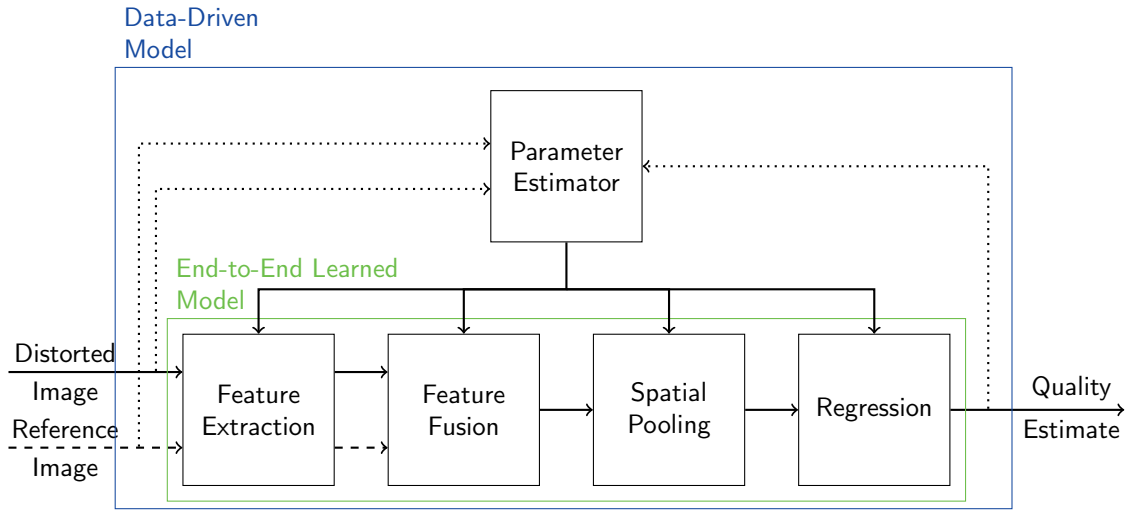
The ultimately decisive criterion for evaluating the quality of a signal or a system is the judgment of a human. However, as discussed in the previous chapter, human quality assessment is cumbersome, expensive and in many application scenarios not accessible. Computational approaches aim at bypassing these problems by estimating the quality of signals without the direct involvement of humans.
This chapter reviews the most relevant computational methods for quality estimation, summarizes different databases of quality annotated images these methods are benchmarked on, and briefly recapitulates the metrics used for performance evaluation.

3.1 Computational Models for Image Quality Estimation

Depending on the information about the reference image available to the algorithm image quality measures (IQMs) correspond to one of three categories each of which has different challenges and application scopes: While FR approaches have access to the full reference image, only the distorted image is available to NR approaches. Reduced-reference (RR) IQMs live in the middle of this spectrum as only a small set of features from the reference image is used for quality estimation [Lin and Kuo, 2011]. Unconstrained NR quality estimation has the notion of being the holy grail of quality estimation as it (ideally) replicates human capabilities. NR approaches are considered to have the broadest scope of applications, but due to the constrained information available the design of reliable models is very difficult. However, conceptually NR is not a feasible approach for some applications. An important example is the encoder control in video compression [Wiegand and Schwarz, 2016]. An unreferenced rate-distortion optimization would steer the encoder towards coding decisions that remove any type of noise or artifact. In some videos, however, noise and artifacts are artistic components that are intentionally introduced in order to evoke a certain emotional response in the viewer. Two popular movies exemplify this: Imagine a video encoder that removes film grain from the Quentin Tarantino movie *The Hateful Eight* or blur and camera shakes form the movie *The Blair Witch Project* due to the use of a NR distortion measure penalizing ”noisy” coding decisions. Such an encoder would undermine the filmmakers' artistic intent and thus might



(a) Model-based image quality estimation



(b) End-to-end data driven image quality estimation

Figure 3.1 – Quality estimation is based on feature extraction, feature fusion, spatial pooling and regression. Information about the reference image (dashed lines) is not available for no reference image quality estimation. Traditional quality estimation employs explicit domain knowledge in the first three stages. Parameters of sigmoidal regression from computational quality scores to perceptual quality estimates are typically trained based on quality annotated images (dotted line). End-to-end data-driven quality estimation does not rely on explicit domain knowledge; all stages of the quality model are trained/learned within one framework. Note that some data-driven quality models apply preprocessing prior to feature extraction.

devalue the viewing experience.

3.1.1 Full-Reference Models

The simplest FR IQM is the mean squared error (MSE) between reference image and distorted image. For having convenient features, it is probably also the widest used IQM , as it *a)* is of low computational complexity, *b)* is memoryless, *c)* qualifies mathematically as a distance metric in $\mathbb{R}^N$ *d)* has a clear physical interpretation as the energy of the error signal *e)* features convexity, symmetry and differentiability, allowing for simple optimization procedures, and *f)* is additive [Wang and Bovik, 2009]. Despite all these advantageous properties the MSE has one crucial disadvantage: As a quality estimator it does not correlate well with visual quality as perceived by humans [Girod, 1993].
This lack of agreement with human perception led scientists and quality engineers to build quality estimators around explicit models that incorporate feature extraction, feature fusion, and/or spatial pooling stages as sketched in Fig. 3.1a. Early approaches exploit processing mechanisms of the human visual system (HVS) or components thereof, such as contrast sensitivity and detection mechanisms [Daly, 1992], different types of masking [Watson et al., 1997], or just noticeable difference (JND) models [Lubin, 1997]. The mean absolute difference (MAD) [Larson and Chandler, 2010] distinguishes between supra- and near-threshold distortions to account for different domains of human quality perception. From an information-theoretic point of view these approaches aim at modelling the *receiver* of visual signals. Although a complete and precise simulation of the HVS most certainly would lead to accurate quality estimators following the receiver modelling line of thinking, the complexity of the architecture of the perceptual system, the non-linearities of its components and the intricacy of its interactions render such a simulation extremely challenging. Note that, as indicated in Chapter 2, also the HVS might encode stimulus features that you the observer is not consciously aware of, which may or may not influence quality, depending on the feature and the definition of quality.
This led researchers in quality estimation to instead model the *transmitter* of visual signals. There, general and in some cases hypothesized abstract properties of the HVS and its functions are assumed in order to identify quality-related features of an image from a signal processing perspective. While in previous receiver model based approaches features from reference and distorted images were mostly fused as feature differences, many of these transmitter model based approaches employ a multiplicative combination of maps of different features.

$$S(x,y) = \prod_i \frac{2 f_{r,i}(x,y) f_{d,i}(x,y)}{f_{r,i}^2(x,y) + f_{d,i}^2(x,y)}. \tag{3.1}$$

These feature maps $f_{r,i}(x,y)$, $f_{d,i}(x,y)$ are extracted from the reference image and the distorted image, respectively, with i indexing different feature types. Typically, $S(x,y)$ is pooled spatially as the arithmetic mean over all sample positions x, y. This similarity structure was introduced with the structural similarity index (SSIM) [Wang et al., 2004]. The SSIM aims at considering the sensitivity of the human visual system towards structural information and pools the luminance similarity (measured by the local luminance), the contrast similarity (measured by the local variance), and the structural similarity (measured as the local covariance between reference

and distorted image) according to Eq. 3.1. Although being heavily criticized [Dosselmann and Yang, 2009] the SSIM is among the most popular IQMs. A multiscale extension led to the multiscale structural similarity index measure (MS-SSIM) [Wang et al., 2003]. Following the similarity framework, the feature similarity index (FSIM) [Zhang et al., 2011] pools two feature maps that are derived from phase coherence and from local gradients. Local gradients and local luminance are used in combination with a contrast masking model [Liu et al., 2012]. The FSIM is significantly simplified and at the same time improved in performance by the Haar wavelet-based perceptual similarity index (HaarPSI) [Reisenhofer et al., 2018]. The visual saliency induced index (VSI) [Zhang et al., 2014b] pools the similarities in gradient magnitude and in estimated visual saliency according to Eq. 3.1 and extends the framework by a local saliency based weighting scheme (see Section 3.1.3). Using an information-theoretic approach to estimate quality it is shown that the mutual information of the wavelet coefficients based on a Gaussian scale mixture model is correlated to perceived quality [Wang et al., 2004].
In all these approaches feature extraction, feature fusion and spatial pooling are informed by rather rigid models based on prior knowledge or assumptions. The regression module however is typically data-driven in a very basic form as the regression parameters for mapping the computational quality scores into the perceptual domain are estimated based on quality annotated ground truth data (see Fig. 3.1a). To account for psychophysical saturation effects, scalar regression is typically implemented by a parameterized sigmoidal function. In Chapter 5 we will discuss this scalar regression process in detail. Multivariate regression combining a set of selected hand-crafted IQMs can improve the performance as shown by using neural network-based regression [Lukin et al., 2015] and support vector regression (SVR) [Lin et al., 2014, Pei and Chen, 2015][1]. In [Pei and Chen, 2015] the output of the multivariate regression is input to a subsequent (and traditional) scalar sigmoid regression.
DeepSim [Gao et al., 2017] extracts feature maps from different layers of a deep convolutional neural network (CNN) that was pre-trained for recognition and employs these feature maps in a similarity framework according to Eq. 3.1. Despite these efforts to apply learning approaches to feature extraction and/or regression consistent end-to-end training as sketched in Fig. 3.1b was not proposed for FR quality estimation.
Novel FR quality models employing end-to-end learning will be proposed in Chapter 4 and Chapter 5.

3.1.2 No-Reference Models

NR image quality estimation predominantly follows the transmitter model approach. In a widely used framework, a parametric statistical image model is assumed and features are extracted from the test image as parameters. Parametric deviations from the underlying statistical model are regressed to quality estimates. As these parameters and their deviations may depend on the distortion type, in a first step the DIIVINE framework [Moorthy and Bovik, 2011] identifies the distortion type affecting an image and employs a distortion-specific

[1] The work of [Lin et al., 2014] was popularized by Netflix under the name video multi-method assessment fusion (VMAF).

regression scheme to estimate the perceived quality in a second step. The statistical features are calculated based on an oriented subband decomposition. BLIINDS-II [Saad et al., 2012] uses a generalized Gaussian density function to model block DCT coefficients of images. BRISQUE [Mittal et al., 2012] proposes a NR quality estimation approach that utilizes an asymmetric generalized Gaussian distribution to model images in the spatial domain. The modeled image features here are differences of spatially neighbored, mean subtracted and contrast normalized image samples. NIQE [Mittal et al., 2013] extracts features based on a multivariate Gaussian model and relates them to perceived quality in an unsupervised manner abandoning explicitly trained regression. In order to cope with more complex and authentic (see Section 3.2.1) distortion types FRIQUEE [Ghadiyaram and Bovik, 2017] employs a deep belief network of 4 layers trained to classify bins of 10 different distortion ranges. Input to the network is a set of handcrafted feature maps. The feature representation on the last hidden layer is extracted to be input to SVR for quality prediction.

Relying on global statistics of full images or full image feature maps these NR approaches do not output spatially localized quality estimates and spatial pooling is done implicitly. However, all these models are still structured based on domain-specific assumption, while the parameters of the models are learned, moving them small steps away from being model-based towards being data-driven.

CORNIA [Ye et al., 2012] relaxes explicit model assumptions and can be considered one of the first purely data-driven NR quality estimators. Here, a codebook is constructed by k-means clustering of luminance and contrast normalized image patches. Soft-encoded distances between visual code-words and patches extracted from distorted images are used as features that are pooled and regressed by SVR for estimating image quality. This approach is extended in the semantic obviousness metric (SOM) [Zhang et al., 2015], where object-like regions are detected and the patches extracted from these regions are input to CORNIA. Like CORNIA, QAF [Zhang et al., 2014a] constructs a codebook, but applies sparse filter learning on patch-wise log-Gabor responses in addition to the contrast and luminance normalized pixel values. As log-Gabor responses are often considered a low-level model of the HVS, conceptually, QAF inherently applies a receiver model. Although these approaches employ much more general models for feature extraction, feature extraction and regression are still treated independently rather than jointly in an end-to-end fashion as sketched in Fig. 3.1b. Motivated by the recent success of CNNs for classification and detection tasks and the notion that the connectivity patterns in these networks resemble those of the ventral stream in the primate visual cortex [Lecun et al., 2015], a shallow CNN consisting of 1 convolutional layer, 1 pooling layer and 2 fully-connected layers, that combines feature extraction and regression is used for quality estimation [Kang et al., 2014]. Quality is estimated on contrast normalized image patches and patch-wise quality is pooled to image-wise quality by simple arithmetic averaging. In order to deal with the scarcity of training data in the form of quality annotated images due to the intricacy of quality assessment, an approach for data augmentation is proposed as BIECON [Kim and Lee, 2017]. CNN-based NR quality estimation is tackled in 2 steps: First, local patch-wise quality is estimated based on normalized image patches employing a CNN of 2 convolutional, 2 pooling and 5 fully-connected layers. This network is trained in order

to replicate a conventional FR IQM such as SSIM or GMSD within a NR framework. Second, mean values and the standard deviations of the extracted patch-wise features are regressed to an image-wise quality estimate employing a perceptron with one hidden layer. Since a FR IQM serves as a proxy, BIECON is not trained in an end-to-end fashion.
Chapter 4 will present novel data-driven NR quality models employing an end-to-end training scheme.

3.1.3 Sensitivity, Saliency and Attention for Image Quality Estimation

Most FR and some NR quality estimators embody a spatially localized representation of quality estimates[2] q_i or local regions i (pixels or patches) that are pooled to a global image-wise quality estimate Q. A classical and widely used method is spatially uniform Minkowski pooling, the l_p norm of the quality estimate maps (or perceptual error estimate map, respectively), where p is typically equal to $p = 1$ or $p = 2$ [Wang and Shang, 2006].
Quality perception is not necessarily spatially uniform and it is intuitively reasonable to assign higher influence to regions *a)* affected by stronger distortions, *b)* of higher saliency, and as such more likely to be attended to, or *c)* carrying scene information, such as objects [Wang and Li, 2011]. A simple and straight-forward method to implement such a spatial non-uniformity is a linear weighting scheme

$$Q = \frac{\sum_i w_i q_i}{\sum_i w_i}, \tag{3.2}$$

where the weight w_i represents the local influence based on a certain model of saliency, attention, scene understanding [Zhang et al., 2016] or distortion magnitude [Wang and Li, 2011]. Despite this intuition and although humans are able to localize visual quality degradation with relative fine spatial granularity, it is so far poorly understood how human observers pool this local judgments to form a global image-wise quality rating (see Section 2.1). For distortion magnitude-based weighting $w_i = w(q_i)$ is often proposed to be a function of q_i itself [Wang and Li, 2011]. By employing a cascade of weighting schemes it is in principle possible to combine different models well.
However, spatial pooling, be it based on saliency, distortion magnitude or any other process, is typically informed by explicit models either transferred from other contexts [Zhang et al., 2016] or based on assumed mechanisms [Wang and Li, 2011]. So far, little work has been done on (end-to-end) data-driven weighting schemes in the context of quality estimation. Saliency models incorporated in IQMs were mostly developed to predict how viewers attend to distinct regions of an image, but not to predict perceived quality explicitly in a joint optimization approach.
Local weights have usually been extracted from the reference image and, as for VSI (see Section 3.1.1), in some cases additionally from the distorted image. Employing saliency-based pooling estimated from the reference image shifts NR quality estimation to the RR domain.

[2] For simplicity we drop the typically used hat symbol for the denotion of estimates $\widehat{q}$, $\widehat{Q}$ in this section.

In Chapter 4 and Chapter 5 patch-wise weighting will be incorporated into data-driven FR and NR quality models and trained in an end-to-end fashion.

3.2 Performance Evaluation

3.2.1 Image Quality Databases

Quality estimators are benchmarked and potentially trained on quality annotated images. There exists a number of publicly available databases of images with quality labels collected in psychometric assessments. Unfortunately, not all databases are compiled according to the relevant recommendations (cf. Section 2.2). This makes the comparison of quality scores from different databases difficult. Note that image quality databases are relatively small in comparison to databases used for object recognition such as ImageNet [Deng et al., 2009]. Furthermore, images in existing databases are mostly of relative low resolution compared to modern image acquisition and display systems.

LIVE Image Quality Database
The LIVE Image Quality Database (LIVE) [Sheikh et al., 2006] database comprises 779 quality annotated images based on 29 source reference images that are subject to 5 different types of distortions at different distortion levels. Most images have a resolution of 768×512 pixels and were upsampled using bicubic interpolation to be presented on 1024×768 screen resolution. Distortion types are JP2K compression, JPEG compression, additive white Gaussian noise, Gaussian blur and a simulated fast fading Rayleigh channel. Quality ratings were collected using a single-stimulus methodology with a hidden, scores from different test sessions were aligned. Resulting DMOS quality ratings lie in the range of $[0,\ 100]$, where a lower score indicates better visual image quality.

Tampere Image Database 2013
The Tampere Image Quality Database 2013 (TID2013) [Ponomarenko et al., 2013] is an extension of the earlier published Tampere Image Quality Database 2008 (TID2008) [Ponomarenko et al., 2009] containing 3000 quality annotated images based on 25 source reference images distorted by 24 different distortion types at 5 distortion levels each. The distortion types cover a wide range from simple Gaussian noise or blur over compression distortions such as JPEG to more exotic distortion types such as non-eccentricity pattern noise. This makes the TID2013 a more challenging database for the evaluation of quality estimators, although the reference images are a subset of those contained in LIVE. The rating procedure differs from the one used for the construction of LIVE, as it employed a competition-like tri-stimulus procedure during which the observers were presented with a reference image and two distorted versions simultaneously. The observer was asked to choose the image of higher visual quality. The chosen image won one point and points assigned to each image were accumulated to the final quality score. Each distorted image was presented in nine comparisons, so the obtained quality scores (named MOS although not complying to the definition of MOS, cf. Section 2.3)

lie in the range [0, 9], where larger values indicate better visual quality. Note that viewing conditions during quality assessment were fairly uncontrolled.

Categorical Subjective Image Quality Database
The Categorical Subjective Image Quality Database (CSIQ) database contains 866 quality annotated images. 30 reference images disjoint to the one in LIVE and TID2013 are distorted by JPEG compression, JP2K compression, Gaussian blur, Gaussian white noise, Gaussian pink noise or contrast change. Image resolution is 512×512 pixels. For quality assessment, subjects were asked to position distorted images horizontally on a monitor array according to their visual quality. After alignment and normalization resulting DMOS values span the range [0, 1], where a lower value indicates better visual quality.

LIVE in the Wild Image Quality Challenge Database
The LIVE In the Wild Image Quality Challenge Database (CLIVE) [Ghadiyaram and Bovik, 2015, Ghadiyaram and Bovik, 2016] comprises 1162 images taken under real life conditions with a large variety of objects and scenes captured under varying luminance conditions using different consumer-grade cameras. In that sense the images are authentically distorted with impairments being the result of a mixture of different distortions, such as over- or under-exposure, blur, grain, or compression. As such, no undistorted reference images are available. Being unreferenced, *quality* might be of more general aesthetic notion that is beyond the scope of this thesis. All images are of the resolution 500×500 pixel. Quality annotations were obtained in the form of MOS in a crowdsourced online study. MOS values lie in the range [0, 100], a higher value indicates higher quality.

3.2.2 Performance Metrics

Quality estimators are typically benchmarked by their prediction accuracy and their prediction monotonicity, where the former is quantified as Pearson Linear Correlation Coefficient (PLCC) and the latter as Spearman rank order coefficient (SROCC) [VQEG, 2004]. The PLCC r_P between datasets $\mathbf{x}$ and $\mathbf{y}$ is defined as

$$r_P = \frac{\sum_{i=0}^{N-1}(x_i - \bar{\mathbf{x}})(y_i - \bar{\mathbf{y}})}{\sqrt{\sum_{i=0}^{N-1}(x_i - \bar{\mathbf{x}})^2 \sum_{i=0}^{N-1}(y_i - \bar{\mathbf{y}})^2}} \tag{3.3}$$

The SROCC r_S between datasets $\mathbf{x}$ and $\mathbf{y}$ is defined as the PLCC of the ranked variables. Note that in contrast to the PLCC the absolute value of the SROCC is not affected by a strictly monotonic mapping such as sigmoid regression functions. For benchmarking both correlation metrics indicate highest performance by an absolute value of 1 and lowest performance by a value of 0.

3.3 Lessons Learned

- The computationally simplest quality estimator is the MSE. It does not correlate well with human perception of visual quality.
- A whole zoo of more sophisticated model-based quality estimators were proposed.
- Little work has been done on end-to-end trained quality estimators.
- Weighted average pooling schemes can be used to improve performance of spatially localized quality estimators.
- Little work has been done on (end-to-end) data-driven pooling schemes
- Publicly available quality annotated image databases are relatively small, rendering data-driven approaches to quality estimation challenging.

4 Data-Driven Estimation of Image Quality

This chapter is based on

Bosse, S., Maniry, D., Müller, K.-R., Wiegand, T., and Samek, W. (2018e). Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on Image Processing*, 27(1):206–219 ©2018 IEEE

4.1 Introduction

Until recently, computer vision has tackeled problems such as image classification and object detection in two steps: (1) designing appropriate features and (2) designing learning algorithms for regression or classification. Although the extracted features were used as input to the learning algorithm, the two steps were mostly independent of each other. More recently, CNNs have outperformed these traditional approaches. One reason is that they allow for joint end-to-end learning of features and regression/classification based on the raw input data, avoiding any hand-engineering [LeCun et al., 1998]. It was further shown that in classification tasks deep CNNs with more layers outperform shallow network architectures [Simonyan and Zisserman, 2014].

This chapter studies the use of a deep CNN in a general quality estimation setting. The CNN is largely inspired by the organization of primate visual cortex, and comprises 10 convolutional layers and 5 pooling layers for feature extraction, and 2 fully connected layers for regression. It is shown that network depth has a significant impact on performance. We start with addressing the problem of FR quality estimation in an end-to-end optimization framework. For that, the concept of Siamese networks known from classification tasks [Bromley et al., 1993, Chopra et al., 2005] is adapted by introducing a feature fusion stage that allows for a joint regression of the features extracted from the reference and the distorted image. Different feature fusion strategies are discussed and evaluated.

As the number of parameters to be trained in deep networks is usually very large, the number of data samples in the training set must be sufficiently large in order to avoid overfitting. The problem of data scarcity discussed in Section 3.2.1 is addressed by artificially augmenting the

datasets, i.e., the network is trained on randomly sampled patches of the quality annotated images. For that, image patches are assigned quality labels from the corresponding annotated images. Unlike most other data-driven quality estimation approaches, patches input to the network are not normalized, which makes the proposed method robust to distortions introduced by luminance and contrast changes. To this end, global image quality is derived by pooling local patch qualities simply by averaging and, for convenience, this method is referred to as Deep Image QuAlity Measure for FR (DIQaM-FR).

As discussed in Section 3.1.3, local quality is not uniformly distributed over an image, and individual image locations vary in how strongly they influence global image quality. This leads to a high amount of label noise in the augmented datasets. Thus, a patch-wise relative weight is assigned to account for the contribution of each image location to the global quality estimate. This is implemented as a simple change to the network architecture that adds two fully connected layers running in parallel to the quality regression layers, combined with a modification of the training strategy. This method is referred to as Weighted Average Deep Image QuAlity Measure for FR (WaDIQaM-FR). This approach allows for a joint optimization of local quality estimation and pooling from local to global quality, formally within the classical framework of saliency weighted distortion pooling.

After establishing our approach within a FR context, one of the feature extraction paths in the Siamese network is abolished. This adaptation allows to apply the network within a NR context as well. Depending on the spatial pooling strategy used, we refer to the NR models as Deep Image QuAlity Measure for NR (DIQaM-NR) and Weighted Average Deep Image QuAlity Measure for NR (WaDIQaM-NR).

Interestingly, by starting with a FR model, our approach facilitates systematic reduction of the amount of information from the reference image necessary for accurate quality prediction. Thus, it helps to close the gap between FR and NR quality estimation. This means that the space of RR quality estimation can be explored from a given FR model without retraining.

The performance of the models trained with the proposed methods are benchmarked on the relevant image quality databases introduced in Section 3.2.1. Because the performance of data-driven approaches depends largely on what data is used for training we evaluate the generalization ability of the proposed methods in cross-database experiments.

This chapter is structured as follows: Section 4.2 develops and details the proposed methods for deep neural network-based FR and NR quality estimation with different patch aggregation methods. In Section 4.3 the presented approaches are evaluated and compared to the state of the art (cf. Section 3.1). Further, weighted average patch aggregation, network depth, and reference reduction are analyzed. The chapter concludes with a discussion and an outlook to future work in Section 4.4.

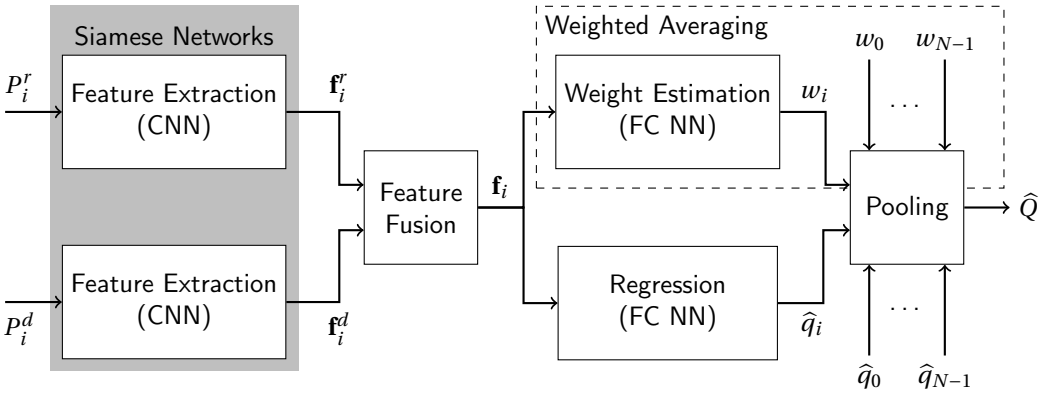


Figure 4.1 – Deep neural network model for FR quality estimation. Features $\mathbf{f}_i^r, \mathbf{f}_i^d$ are extracted from reference and distorted patches P_i^r, P_i^d by a CNN and fused as difference, concatenation or concatenation supplementary with the difference vector to $\mathbf{f}_i$. The fused feature vector is regressed to a patch-wise quality estimate $\widehat{q}_i$. The dashed-boxed branch of the network indicates an optional regression of the feature vector to a patch-wise weight estimate w_i that allows for pooling by weighted average patch aggregation. The output is the image-wise quality estimate $\widehat{Q}$.

4.2 Deep Neural Networks for Image Quality Estimation

4.2.1 Neural network-based FR Quality Estimation

Siamese networks have been used to learn similarity relations between two inputs by processing the inputs in parallel using two networks that share synaptic connection weights. This approach has been applied to signature [Bromley et al., 1993] and face verification [Chopra et al., 2005] tasks, where the inputs are binarily classified as belonging or not belonging to the same category. For FR quality estimation we perform feature extraction with a Siamese network. In order to use the extracted features for the regression problem of quality estimation, feature extraction is followed by a feature fusion step. The fused features are then input to the regression part of the network. The architecture of the proposed network is sketched in Fig. 4.1 and will be further detailed in the following.

Motivated by its superior performance in the 2014 ILSRVC classification challenge [Russakovsky et al., 2015] and its successful adaptation for various computer vision tasks [Girshick, 2015, Long et al., 2015], VGGnet [Simonyan and Zisserman, 2014] was chosen as a basis for the proposed networks. VGGnet has a straight-forward, but deep CNN architecture and was the first neural network to employ cascaded convolutions kernels small as 3×3. The input of the VGG network are images of the size 224×224 pixels. To allow smaller input sizes such as 32×32 pixel-sizes patches, we extend the network by 3 layers (conv3-32, conv3-32, maxpool) placed in front of the original architecture. Our proposed VGGnet-inspired CNN thus comprises 12 weight layers that are organized in a feature extraction module and a regression module. The features are extracted in a series of conv3-32, conv3-32, maxpool, conv3-64, conv3-64, maxpool, conv3-128, conv3-128, maxpool, conv3-256, conv3-256, maxpool, conv3-512, conv3-512,

maxpool layers[1]. The fused features (see Section 4.2.2) are regressed by a sequence of one FC-512 and one FC-1 layer. With 3×3 pixel-size convolution kernels, this produces about 5 million trainable network parameters, 5.2 million if the weight estimates are also included. All convolutional layers are activated through a rectified linear unit (ReLU) activation function $g = \max(0, \sum_i w_i a_i)$, where g, w_i and a_i denote the output, the weight and the input of the ReLU, respectively [Nair and Hinton, 2010]. To obtain an output of the same size as the input, convolutions are applied with zero-padding. All max-pool layers have 2×2 pixel-sized kernels. Dropout regularization with a ratio of 0.5 is applied to the fully-connected layers in order to prevent overfitting [Srivastava et al., 2014].
For our quality estimation approach, reference and distorted images are subdivided into 32×32 sized collocated patches P_i^r, P_i^d that are input to the neural network. Local patch-wise quality estimates $\widehat{q}_i$ are pooled into a global image-wise quality estimate $\widehat{Q}$ by simple or weighted average patch aggregation. The strategy for spatial pooling affects the training of the network and will be explained in more detail in Section 4.2.3.
For convenience we refer to the resulting models as Deep Image QuAlity Measure for FR (DIQaM-FR) and Weighted Average Deep Image QuAlity Measure for FR (WaDIQaM-FR).

4.2.2 Feature Fusion

In order to serve as input to the regression part of the network, the extracted feature vectors $\mathbf{f}_i^r$ and $\mathbf{f}_i^d$ are combined in a feature fusion step. In the FR framework, concatenating $\mathbf{f}_i^r$ and $\mathbf{f}_i^d$ to $\texttt{concat}(\mathbf{f}_i^r, \mathbf{f}_i^d)$ without any further modifications is the simplest way of feature fusion. $\mathbf{f}_i^r$ and $\mathbf{f}_i^d$ are of identical structure, which renders the difference $\mathbf{f}_i^r - \mathbf{f}_i^d$ to be a meaningful representation for distance in feature space. Although the regression module should be able to learn $\mathbf{f}_i^r - \mathbf{f}_i^d$ by itself, the explicit formulation might ease the actual regression task. This allows for two other simple feature fusion strategies, namely the difference $\mathbf{f}_i^r - \mathbf{f}_i^d$, and the concatenation of feature vectors and difference $\texttt{concat}(\mathbf{f}_i^r, \mathbf{f}_i^d, \mathbf{f}_i^r - \mathbf{f}_i^d)$.

4.2.3 Spatial Pooling

4.2.3.1 Pooling by Simple Averaging

The simplest way to pool locally estimated visual qualities $\widehat{q}_i$ to a global image-wise quality estimate $\widehat{Q}$ is to assume identical relative importance of every image region, or, more specifically, of every image patch P_i as

$$\widehat{Q} = \frac{1}{N_p} \sum_i^{N_p} \widehat{q}_i, \tag{4.1}$$

[1] Notation is borrowed from [Simonyan and Zisserman, 2014] where conv⟨receptive field size⟩-⟨number of channels⟩ denotes a convolutional layer and FC⟨number of channels⟩ a fully-connected layer

where N_p denotes the number of patches sampled from the image. For regression tasks, commonly the MSE is used as minimization criterion. However, as simple average quality pooling implicitly assigns the locally perceived quality to be identical to the globally perceived quality Q_p this approach introduces a certain degree of label noise into the training data. Optimization with respect to mean absolute error (MAE) puts less emphasis on outliers and reduces their influence. As our estimation problem is a regression task, we choose MAE as a less outlier sensitive alternative to MSE. The loss function to be minimized is then

$$E_{simple} = \frac{1}{N_p} \sum_{i}^{N_p} |\widehat{q}_i - Q_p|. \tag{4.2}$$

In principle, the number of patches N_p can be chosen arbitrarily. A complete set of all non-overlapping patches would ensure all pixels of the image to be considered and, given the same trained CNN model, be mapped to reproducible scores.

4.2.3.2 Pooling by weighted average patch aggregation

As discussed in Section 3.1.3, the perceived quality in a local region of an image does not necessarily correspond to the global image-wise perceived quality, due to effects such as spatially non-uniformly distributed distortion, summation or saliency effects or combinations thereof. In the pooling-by-average approach described above this is accounted for only very roughly by employing a less outlier-sensitive loss function. However, spatial pooling by averaging local quality estimates does not consider the effect of spatial variability in the perceptual relevance of local quality.
We address this spatial variability of relative image quality by integrating a second branch into the regression module of the network that runs in parallel to the patch-wise quality regression branch (see Fig. 4.1) and shares the same dimensionality. This branch outputs an w_i^* for a patch P_i. By activating w_i^* through a ReLU and adding a small stability term ϵ

$$w_i = \max(0, w_i^*) + \epsilon \tag{4.3}$$

it is guarantied to result in local weights $w_i > 0$ that can be used to weight the estimated quality $\widehat{q}_i$ of the respective patch P_i.
With the normalized weights

$$v_i = \frac{w_i}{\sum_j^{N_p} w_j}. \tag{4.4}$$

the global image quality estimate $\widehat{Q}$ can be calculated as

$$\widehat{Q} = \sum_{i}^{N_p} p_i \widehat{q}_i \; = \; \frac{\sum_i^{N_p} w_i \widehat{q}_i}{\sum_i^{N_p} w_i}. \tag{4.5}$$

As in Eq. 4.2, the number of patches N_p can be set arbitrarily. Comparing Eq. 4.5 to Eq. 3.2 shows that the proposed pooling method implements a weighting technique formally equivalent to the framework of linear saliency weighting as described in Section 3.1.3.

For joint end-to-end training the loss function to be minimized is then

$$E_{weighted} = |\widehat{Q} - Q_p|. \tag{4.6}$$

4.2.4 Network Adaptations for NR Quality Estimation

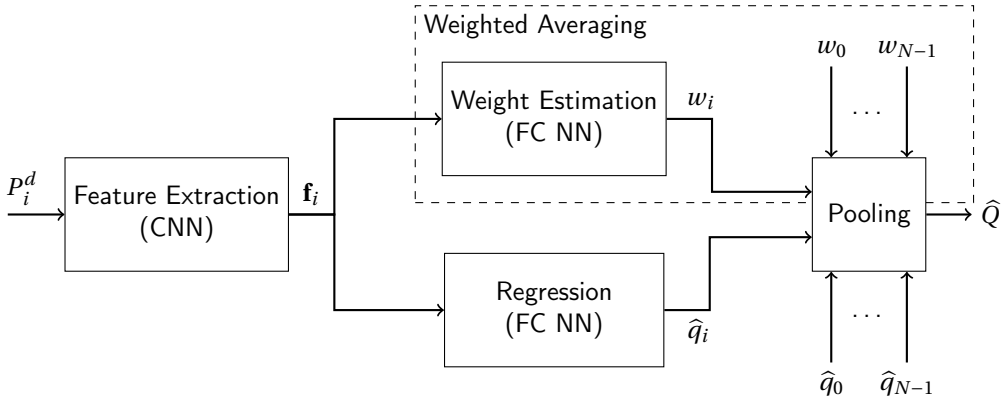


Figure 4.2 – Deep neural network for NR quality estimation. Features are extracted from the distorted patch P_i^d by a CNN. The feature vector $\mathbf{f}_i$ is regressed to a patch-wise quality estimate $\widehat{q}_i$. Patch-wise estimates are aggregated to the global image quality estimate $\widehat{Q}$. The dashed-boxed branch of the network indicates an optional regression of the feature vector to a patch-wise weight estimate w_i that allows for pooling by weighted average patch aggregation.

The proposed deep network can also be used in a NR quality estimation context, by simply abolishing the branch that extracts features from the reference patch using a Siamese network. As features from the reference patch are no longer available, no feature pooling is necessary. However, both spatial pooling methods detailed in Section 4.2.3 are applicable for NR quality estimation as well. The resulting approaches are referred to as Deep Image QuAlity Measure for NR (DIQaM-NR) and Weighted Average Deep Image QuAlity Measure for NR (WaDIQaM-NR) and are subject to the same loss functions as for the FR case. The resulting architecture of the neural network adapted for NR quality estimation is illustrated in Fig. 4.2.

4.2.5 Training

The proposed networks are trained iteratively by backpropagation [LeCun et al., 1998, LeCun et al., 2012] over a number of epochs, where an epoch is defined as a period during which (parts of) each image from the training set has been used once. In each epoch, the training set is divided into mini-batches for batch-wise optimization. Although it is possible to treat each image patch as a separate sample in the case of simple average pooling, for weighting average

pooling image patches of the same image can not be distributed over different mini-batches, as their output is combined when the normalized weights are combined in the last layer. To make the training approach as similar as possible across all methods, each mini-batch contains 4 images, each represented by 32 randomly sampled image patches which leads to the effective batch size of 128 patches. The backpropagated error is the average loss over all images in a mini-batch. For training the FR quality estimation networks, the respective reference patches are included in the mini-batch. Patches are randomly sampled in every epoch to ensure that as many different image patches as possible are used in training.
The learning rate for the batch-wise optimization is controlled per parameter adaptively using the ADAM method [Kingma and Ba, 2014] based on the variance of the gradient. Parameters of ADAM are chosen as recommended in [Kingma and Ba, 2014] as $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$ and $\alpha = 10^{-4}$. The mean loss over all images during validation is computed in evaluation mode (i.e. dropout is replaced with scaling) after each epoch. The 32 random patches for each validation image are only sampled once at the beginning of training in order to avoid noise in the validation loss. The final model used for evaluation is the one with the best validation loss. This amounts to early stopping (see [Prechelt, 2012] for a review), a regularization technique to prevent overfitting.
Note that the two regression branches estimating patch weight and patch quality do not have identical weights, as the update of the network weights is calculated based on gradients with respect to different parameters.

4.3 Experiments and Results

4.3.1 Experimental Setup

For evaluation per database networks are trained and tested either on LIVE, TID2013 or (in the NR case) CLIVE. For cross-validation, databases are randomly split by reference image. This guarantees that no distorted or undistorted version of an image used in testing or validation has been seen by the network during training. For LIVE, the training set is based on 17 reference images, validation and test set on 6 reference images each. TID2013 is split analogously in 15 training, 5 validation and 5 test images. CLIVE does not contain versions of different quality levels of the same image, therefore splitting in sets can be done straightforward on the full set of distorted images. Training set size for CLIVE is 698 images, validation and test set sizes 232 images each. Results reported are based on 10 random splits. Models are trained for 3000 epochs. Even though some models converge after less than 1000 epochs a high number is used to ensure convergence for all models. After training the network has seen ~48M patches in the case of LIVE, ~178M patches in the case of TID2013, and ~67M in the case of CLIVE.
To assess the generalization ability of the proposed methods the CSIQ image database is used for cross-dataset evaluating the models trained either on LIVE or on TID2013. For training the model on LIVE, the dataset is split into 23 reference images for training and 6 for validation; analogously, for training the model on TID2013, the dataset is split into 20 training images and 5 validation images. LIVE and TID2013 share a lot of reference images, thus, tests between

these two are unsuitable for evaluating generalization for unseen images. For cross-distortion evaluation, models trained on LIVE are tested on TID2013 in order to determine how well a model deals with distortions that have not been seen during training and in order to evaluate whether a method is truly non-distortion or just many-distortion specific.
Note that, in contrast to many studies in the literature, we use the full TID2013 database and do not ignore any specific distortion type. To make errors and gradients comparable for different databases, the MOS values of TID2013 and CLIVE and the DMOS values of CSIQ have been linearly mapped to the same range as the DMOS values in LIVE. Note that by this mapping high values of $\widehat{q}_i$ visualized in Section 4.3.3 represent high local distortion, rather than high quality.

4.3.2 Performance Evaluation

Evaluations presented in this subsection are based on image quality estimation considering $N_P = 32$ patches. Other values of N_p will be discussed in Section 4.3.5. Performances of the FR models are reported for features fused by $\texttt{concat}(\mathbf{f}_i^r, \mathbf{f}_i^d, \mathbf{f}_i^r - \mathbf{f}_i^d)$; the influence of the different feature fusion schemes are examined in Section 4.3.6.

Table 4.1 – Performance comparison on LIVE and TID2013 databases ©2018 IEEE

		LIVE		TID2013	
	IQM	LCC	SROCC	LCC	SROCC
Full-Reference	PSNR	0.872	0.876	0.675	0.687
	SSIM [Wang et al., 2004]	0.945	0.948	0.790	0.742
	FSIM_C [Zhang et al., 2011]	0.960	0.963	0.877	0.851
	GMSD [Xue et al., 2014]	0.956	0.958	-	-
	DOG-SSIM [Pei and Chen, 2015]	0.963	0.961	0.919	0.907
	DeepSim [Gao et al., 2017]	0.968	**0.974**	0.872	0.846
	DIQaM-FR (proposed)	0.977	0.966	0.880	0.859
	WaDIQaM-FR (proposed)	**0.980**	0.970	**0.946**	**0.940**
No-Reference	BLIINDS-II[Saad et al., 2012]	0.916	0.912	0.628	0.536
	DIIVINE [Moorthy and Bovik, 2011]	0.923	0.925	0.654	0.549
	BRISQUE [Mittal et al., 2012]	0.942	0.939	0.651	0.573
	NIQE [Mittal et al., 2013]	0.915	0.914	0.426	0.317
	BIECON [Kim and Lee, 2017]	0.962	0.961	-	-
	FRIQUEE [Ghadiyaram and Bovik, 2017]	0.930	0.950	-	-
	CORNIA [Ye et al., 2012]	0.935	0.942	0.613	0.549
	CNN [Kang et al., 2014]	0.956	0.956	-	-
	SOM [Zhang et al., 2015]	0.962	**0.964**	-	-
	DIQaM-NR (proposed)	**0.972**	0.960	**0.855**	**0.835**
	WaDIQaM-NR (proposed)	0.963	0.954	0.787	0.761

Table 4.2 – Performance comparison for different subsets of TID2013 ©2018 IEEE

	Noise	Actual	Simple	Exotic	New	Color
PSNR	0.822	0.825	0.913	0.597	0.618	0.535
SSIM [Wang et al., 2004]	0.757	0.788	0.837	0.632	0.579	0.505
$FSIM_C$ [Zhang et al., 2011]	0.902	0.915	0.947	0.841	0.788	0.775
DOG-SSIM [Pei and Chen, 2015]	0.922	0.933	0.959	0.889	0.908	0.911
DIQaM-FR (proposed)	0.938	0.923	0.885	0.771	0.911	0.899
WaDIQaM-FR (proposed)	**0.969**	**0.970**	**0.971**	**0.925**	**0.941**	**0.934**

4.3.2.1 Full-Reference Image Quality Estimation

The upper part of Table 4.1 summarizes the performance of the proposed FR models in comparison to other state-of-the-art methods on the full LIVE and full TID2013 database in terms of PLCC and SROCC. With any of the two presented spatial pooling methods, the proposed approach obtains superior performance to state-of-the-art on LIVE, except for DeepSim evaluated by SROCC. On TID2013 DIQaM-FR performs better than most evaluated state-of-the-art methods, but is outperformed by DOG-SSIM[2]. Here, employing weighted average patch aggregation clearly improves the performance and WaDIQaM-FR performs better than any other evaluated IQM. This effect can be observed as well in Table 4.2 for the groups of different distortion of TID2013 defined in [Ponomarenko et al., 2013]. While DIQaM-FR performs comparable to state-of-the-art methods, on some groups better, on some worse, WaDIQaM-FR shows superior performance for all grouped distortion types.

4.3.2.2 No-Reference Image Quality Estimation

Table 4.3 – Performance evaluation for NR quality estimation on CLIVE ©2018 IEEE

	PLCC	SROCC
FRIQUEE [Ghadiyaram and Bovik, 2017]	**0.706**	**0.682**
BRISQUE [Mittal et al., 2012]	0.610	0.602
DIIVINE [Moorthy and Bovik, 2011]	0.557	0.509
BLIINDS-II [Saad et al., 2012]	0.449	0.404
NIQE [Mittal et al., 2013]	0.477	0.421
DIQaM-NR (proposed)	0.601	0.606
WaDIQaM-NR (proposed)	0.680	0.671

Performances of the NR IQMs are compared to other state-of-the-art NR quality estimation methods in the lower part of Table 4.1. The proposed model employing simple average pooling (DIQaM-NR) performs best in terms of PLCC among all methods evaluated, and in terms of SROCC performs slightly worse than SOM. Evaluated on TID2013, DIQaM-NR performs superior among all other methods in terms of LCC and SRCC[2]. Although no results were reported

[2] Unfortunately, for many state-of-the-art FR and NR quality estimation methods no results are reported on TID2013.

Table 4.4 – PLCC on selected distortion of TID2013 ©2018 IEEE

	GB	JEPG	J2K	LBDDI
DIQaM-FR	0.884	0.965	0.900	0.634
WaDIQaM-FR	**0.963**	**0.978**	**0.975**	**0.683**
DIQaM-NR	**0.872**	**0.946**	**0.872**	0.479
WaDIQaM-NR	0.618	0.726	0.816	**0.664**

for BIECON [Kim and Lee, 2017] on the TID2013 dataset, this method achieves a relatively high SROCC=0.923 on the older TID2008 database when 5 distortion types (non-eccentricity pattern noise, local block-wise distortions, mean shift, contrast change) are excluded from the analysis. Future investigations will show how BIECON performs on the challenging TID2013 database with all distortions included. In contrast to our FR models in the NR case the weighted average patch aggregation pooling decreases the prediction performance when evaluated on full databases.

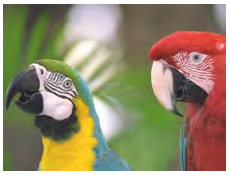
A comparison of performances for the CLIVE database is shown in Table 4.3. Quality estimation on CLIVE is much more difficult than on LIVE or TID2013, thus performances of all methods evaluated are much worse than for the legacy databases. WaDIQaM-NR shows prediction performance superior to most other models, but is clearly outperformed by FRIQUEE. Interestingly and contrasting to the results on LIVE and TID2013, on CLIVE WaDIQaM-NR performs clearly better than DIQaM-NR.

In order to analyze these apparent contradictory results a little deeper, Table 4.4 shows the performance of (Wa)DIQaM-FR and (Wa)DIQaM-NR for four selected distortion types from TID2013 (Gaussian blur, JPEG compression, JP2K compression and local block-wise distortions of different intensity). While for (WA)DIQaM-FR we see the same behavior on single distortion types as on aggregated distortion types, in the NR case weighted patch aggregation pooling decreases performance for GB, JPEG and JP2K, but increases performance for local block-wise distortions of different intensity (LBDDI). We conjecture that for most distortions information from the reference image is crucial to assign local weights for pooling, but if distortions are strongly inhomogeneous as it is the case for LBDDI, the distorted image is sufficient to steer weighting. One of the reasons for CLIVE being so challenging for quality estimation is that distortions and scenes are spatially much more inhomogeneous than in LIVE or TID2013, which the weighted patch aggregation can compensate for. This also explains the huge increase in performance by weighted patch aggregation pooling for the *exotic* subset of the TID2013 in Table 4.4 as this subset contains a larger amount of inhomogeneous distortion types. The resulting local weights will be examined more closely in the next section.

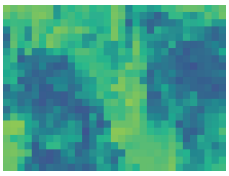
4.3.3 Local Weights

The previous sections showed that the weighted average patch aggregation scheme employed in WaDIQaM-FR/NR has an influence that depends on the distortion type and the availability of a reference.

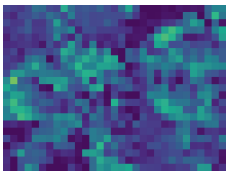
Fig. 4.3 shows the local quality estimates $\hat{q}_i$ and weights w_i for an image subject to JP2K compression from TID2013. The MOS value of the distorted image is 34; the relation between prediction accuracies of the four different models are as expected from the previous evaluations (DIQaM-FR: 54, WaDIQaM-FR: 42, DIQaM-NR:60, WaDIQaM-NR: 70). The left column shows the quality estimate and weight maps computed by the proposed FR models, the right column the maps from the NR models. The DIQaM-FR/NR assign higher distortion values to the background of the image than to the two foreground objects (Figs. 4.3b and 4.3c). In the FR case, the local weights provide some rough image segmentation as higher weights are assigned to image regions containing objects (Fig. 4.3f). This fails in the NR case (Fig. 4.3g), which explains the performance drop from DIQaM-NR to WaDIQaM-NR observed in Section 4.3.2.2.
The local quality estimate and weight maps resulting from an image subject to spatially highly variant distortions, in this example LBDDI from TID2013, is shown in Fig. 4.4. Here, for WaDIQaM-FR as well as for WaDIQaM-NR the network is able to assign higher weights to the distorted image regions and by that improve prediction accuracy compared to the models employing simple average pooling. Note that, as in Fig. 4.3, WaDIQaM-FR is again able to roughly segment the image, whereas WaDIQaM-NR again fails at segmentation. However, for this extreme distortion type the general structure of the image is of less importance.
In Section 4.3.2.2 we conjectured that one reason for WaDIQaM-NR to improve prediction performance over DIQaM-NR for CLIVE, but to decrease performance on LIVE and TID2013 is the higher amount of spatial variance in CLIVE. Fig. 4.5 exemplifies this effect for two images from CLIVE.
The left-most column shows the test images, where the top one (Fig. 4.5a) suffers spatially rather uniformly from underexposure, rendering identification of a certain area of higher impairment difficult, while the bottom one (Fig. 4.5b) contains clear regions of interest that are rather easy to identify against the black background. The lower rows show the corresponding



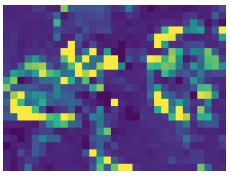
(a) Distorted Image



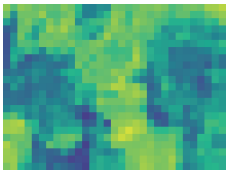
(b) $\hat{q}_i$, DIQaM-FR



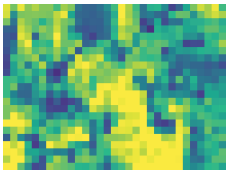
(d) $\hat{q}_i$, WaDIQaM-FR



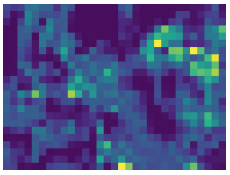
(f) w_i, WaDIQaM-FR



(c) $\hat{q}_i$, DIQaM-NR

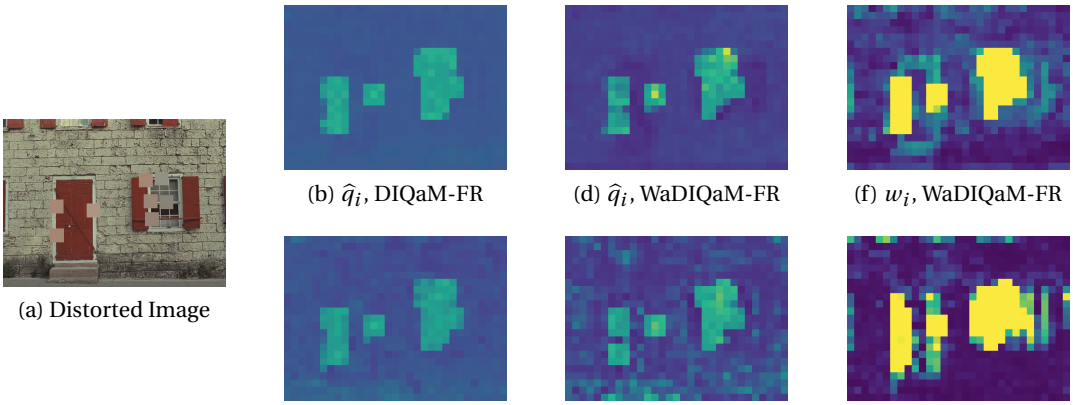


(e) $\hat{q}_i$, WaDIQaM-NR



(g) w_i, WaDIQaM-NR

Figure 4.3 – Local quality estimates $\hat{q}_i$ and weights w_i for a JP2K distorted image from TID2013. Blue indicates low, yellow high values of local distortions and weights, respectively. The MOS value is 34, predicted qualities are 54 by DIQaM-FR, 42 by WaDIQAM-FR, 60 by DIQaM-NR, and 70 by WaDIQam-NR. ©2018 IEEE



(a) Distorted Image

(b) $\widehat{q}_i$, DIQaM-FR

(c) $\widehat{q}_i$, DIQaM-NR

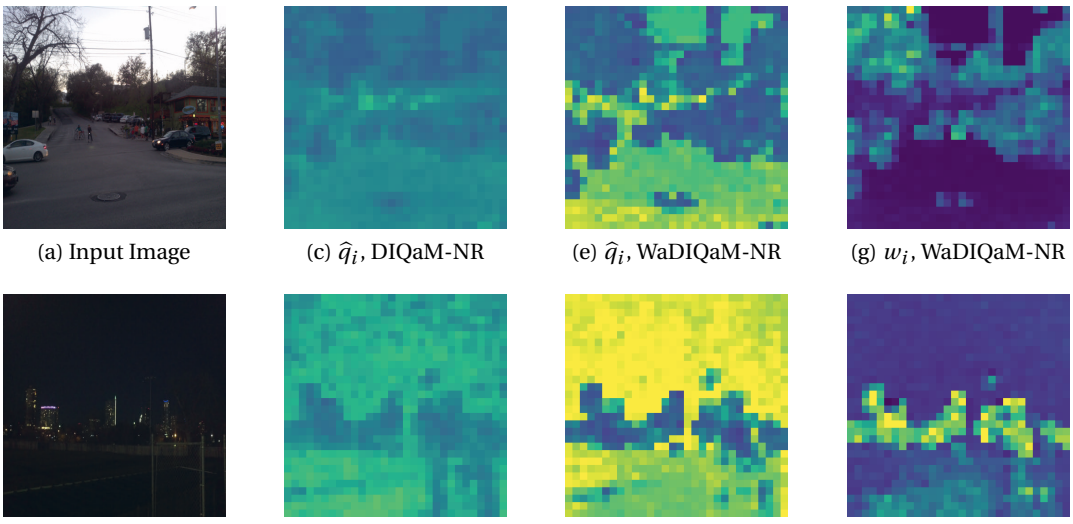
(d) $\widehat{q}_i$, WaDIQaM-FR

(e) $\widehat{q}_i$, WaDIQaM-NR

(f) w_i, WaDIQaM-FR

(g) w_i, WaDIQaM-NR

Figure 4.4 – Local quality estimates $\widehat{q}_i$ and weights w_i for a LBDDI distorted image from TID2013. Blue indicates low, yellow high values of local distortions and weights, respectively. The MOS value is 59, predicted qualities are 30 by DIQaM-FR, 51 by WaDIQAM-FR, 27 by DIQaM-NR, and 53 by WaDIQam-NR. ©2018 IEEE



(a) Input Image

(b) Input Image

(c) $\widehat{q}_i$, DIQaM-NR

(d) $\widehat{q}_i$, DIQaM-NR

(e) $\widehat{q}_i$, WaDIQaM-NR

(f) $\widehat{q}_i$, WaDIQaM-NR

(g) w_i, WaDIQaM-NR

(h) w_i, WaDIQaM-NR

Figure 4.5 – Local quality estimates $\widehat{q}_i$ and weight maps w_i for two image from CLIVE. **Top row**: MOS value is 43, predicted qualities are 42 by DIQaM-NR, 34 by WaDIQam-NR. **Bottom row**: MOS value is 73, predicted qualities are 56 by DIQaM-NR, 66 by WaDIQam-NR. ©2018 IEEE

quality estimate and weight maps. Fig. 4.5h shows that for this spatially highly concentrated scene, WaDIQaM-NR is able to identify the patches contributing the most to the overall image structure. However, as Fig. 4.5g shows, it fails to do so for the homogeneously impaired image. Another important observation from Fig. 4.3, Fig. 4.4 and Fig. 4.5 is that weighted average patch aggregation has an influence also on the quality estimate maps. Thus, the joint optimization

introduces an interaction between $\widehat{q}_i$ and w_i that is adaptive to the specific distortion.

4.3.4 Cross-Database Evaluation

4.3.4.1 Full-Reference Image Quality Estimation

Table 4.5 – SROCC comparison in cross-database evaluation. All models are trained on full LIVE or TID2013, respectively, and tested on either CSIQ, LIVE or TID2013. ©2018 IEEE

Trained on:	LIVE		TID2013	
Tested on:	TID2013	CSIQ	CSIQ	LIVE
DOG-SSIM [Pei and Chen, 2015]	**0.751**	**0.914**	0.925	**0.948**
DIQaM-FR (proposed)	0.437	0.660	0.863	0.796
WaDIQaM-FR (proposed)	**0.751**	0.909	**0.931**	0.936

Table 4.5 shows the results for models trained on LIVE and tested on TID2013 and CSIQ, and for models trained on TID2013 and tested on LIVE and CSIQ. Results are compared to DOG-SSIM, as most other FR quality estimation methods compared to do not rely on training. In all combinations of training and test set the DIQaM-FR model shows insufficient generalization capabilities, while WaDIQaM-FR performs best among the two proposed spatial pooling schemes and comparable to DOG-SSIM. The superior results of the model trained on TID2013 over the model trained on LIVE when tested on CISQ indicate that a larger training set may lead to better generalization.

4.3.4.2 No-Reference Image Quality Estimation

Table 4.6 – SROCC in cross-database evaluation. All models are trained on the full LIVE database and evaluated on CSIQ and TID2013. The subsets of CSIQ and TID2013 contain only the 4 distortions shared with LIVE. ©2018 IEEE

	subset		full	
	CSIQ	TID2013	CSIQ	TID2013
DIIVINE [Moorthy and Bovik, 2011]	-	-	0.596	0.355
BLIINDS-II [Saad et al., 2012]	-	-	0.577	0.393
BRISQUE [Mittal et al., 2012]	0.899	0.882	0.557	0.367
CORNIA [Ye et al., 2012]	0.899	0.892	0.663	0.429
QAF [Zhang et al., 2014a]	-	-	0.701	0.440
CNN [Kang et al., 2014]	-	0.920	-	-
SOM [Zhang et al., 2015]	-	**0.923**	-	-
DIQaM-NR (proposed)	**0.908**	0.867	0.681	0.392
WaDIQaM-FR (proposed)	0.866	0.872	**0.704**	**0.462**

To evaluate the generalization ability of the proposed NR quality estimation models, we extend cross-database experiments from the literature [Zhang et al., 2014a] with our results. For

Table 4.7 – SROCC comparison in cross-database evaluation. All models are trained on the full TID2013 database and evaluated on CSIQ. ©2018 IEEE

Method	CSIQ full
DIIVINE [Moorthy and Bovik, 2011]	0.146
BLIINDS-II [Saad et al., 2012]	0.456
BRISQUE [Mittal et al., 2012]	0.639
CORNIA [Ye et al., 2012]	0.656
DIQaM-NR (proposed)	0.717
WaDIQaM-NR (proposed)	**0.733**

that, a model trained on the full LIVE database is evaluated on subsets of CSIQ and TID2013, containing only the four distortions types shared between the three databases (JPEG, JP2K, Gaussian blur and white noise). Results are shown in Table 4.6. While DIQaM-NR shows superior performance compared to BRISQUE and CORNIA on the CISQ subset, the proposed models are outperformed by the other state-of-the-art methods when cross-evaluated on the subset of TID2013. As for the full CSIQ database, the two unseen distortions (i.e. pink additive noise and contrast change) are considerably different in their visual appearance and statistical structure compared to the ones seen during training. Thus, it is not surprising that all compared methods perform worse in this setting. Despite performing worse on the single database experiments, WaDIQaM-NR seems to be able to adapt better to unseen distortions than DIQaM-NR. This is in line with the results on CLIVE – for CLIVE the specific mixture of distortions of a given image is less likely to be in the training set than e.g. for LIVE. Although being a vague comparison as TID2008 contains less distorted images per distortion type, it is worth noting that BIECON obtains a SROCC=0.923 in a similar experiment (trained on LIVE, tested on the 4 distortions types of the smaller TID2008 and excluding one image).

Given the relatively wide variety of distortions types in TID2013 and with only 4 out of 24 distortions being contained in the training set, a model trained on LIVE can be expected to perform worse if tested on TID2013 than if tested on CSIQ. Unsurprisingly, none of the learning-based methods available for comparison is able to achieve a SROCC over 0.5. These results suggest that learning a truly non-distortion-specific quality estimator using only the examples in the LIVE database is hard or even impossible. Nevertheless, the proposed methods obtain competitive, but still very unsatisfactory results.

Table 4.7 shows the performance in terms of SROCC for models trained on full TID2013 and tested on full CSIQ. Performance of DIIVINE, BLIINDS-II and CORNIA trained on TID2013 is decreased compared to being trained on LIVE, despite TID2013 being the larger and more diverse training set, while BRISQUE and the proposed models are able to make use of the larger training set size. This illustrates the notion that deep neural networks can use larger and more diverse training sets to improve their generalization abilities, while shallow networks cannot.

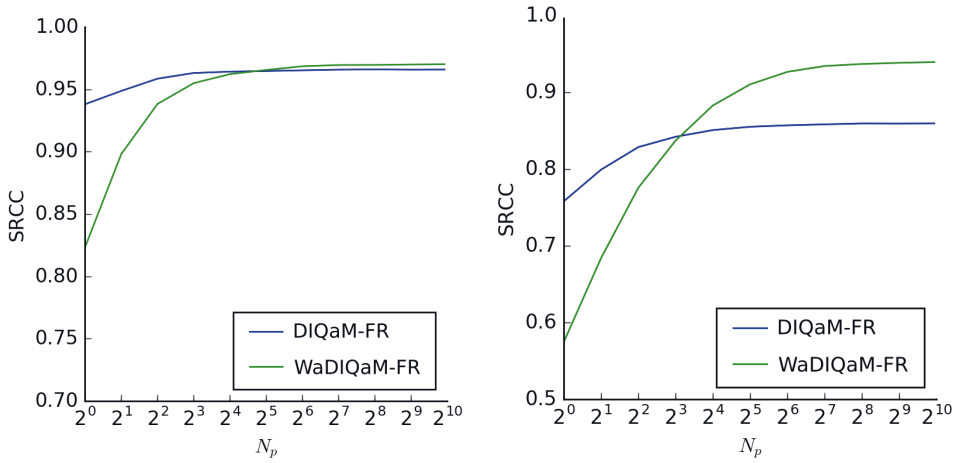
Even though the proposed methods outperform comparable methods, a SROCC of 0.733 on the CSIQ dataset is still far from being satisfactory. Despite having more images in total and

more distortions than LIVE, the TID2013 has 4 reference images fewer. Thus, training on TID2013 has the same short-comings as training on LIVE when it comes to adapting to unseen images.
Note that for the evaluation of learning-based quality estimation methods databases are split into training, testing and (for some approaches) validation sets. Models are then trained and evaluated for a number of different random splits and the performances on these splits need to be pooled for comparison. This renders evaluation a random process. Performances of different models as reported in the literature are obtained based on different split ratios, different numbers of splits and different ways of pooling. This may have a slight influence on the comparison of competing methods.

4.3.5 Convergence Evaluation

Results presented in the previous sections were obtained when $N_p = 32$ patches are considered for quality estimation. However, the prediction performance and accuracy depends on the number of patches used. This subsection examines the influence of N_p and justifies the choice of $N_p = 32$.

4.3.5.1 Full-Reference Image Quality Estimation



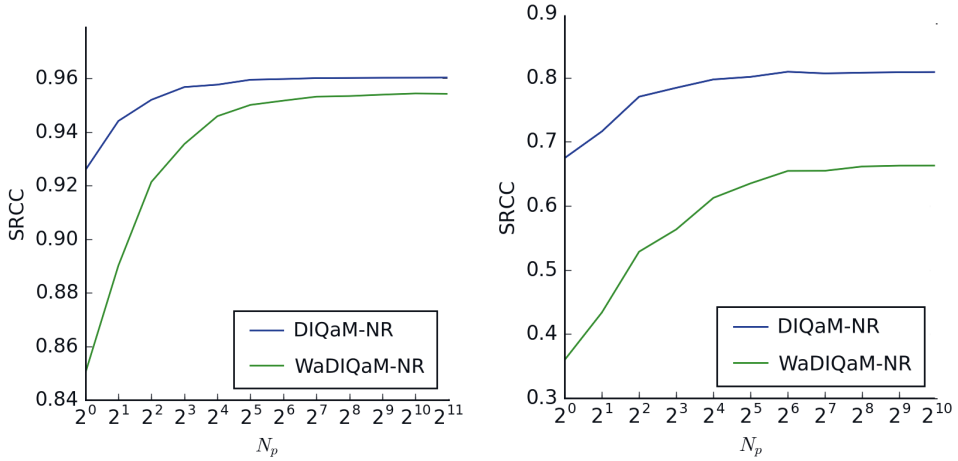
(a) LIVE (b) TID2013

Figure 4.6 – SROCC of the proposed CNN for FR quality estimation in dependence of the number of randomly sampled patches evaluated on LIVE and TID2013. ©2018 IEEE

Fig. 4.6 plots the performance for the models trained and tested on LIVE and TID2013 with varying N_p in terms of SROCC and shows a monotonic increase in performance towards saturation with larger N_p. As noted before, weighted average patch aggregation improves the prediction performance over simple averaging, here we see that this holds when the number of patches large enough, e.g. $N_p > 8$. The WaDIQaM-FR saturates at the maximum performance

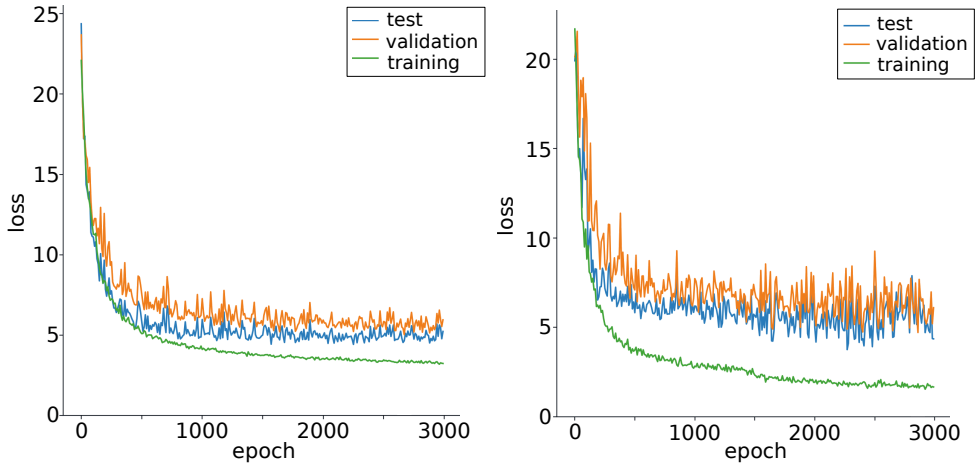
with $N_p \approx 32$, whereas the DIQaM-FR saturates already at $N_p \approx 16$ in all three evaluation metrics.

4.3.5.2 No-Reference Image Quality Estimation



(a) LIVE (b) TID2013

Figure 4.7 – SROCC of the proposed CNN for NR IQA in dependence of the number of randomly sampled patches evaluated on LIVE and TID2013. ©2018 IEEE



(a) DIQaM-NR (b) WaDIQaM-NR

Figure 4.8 – Loss in training, validation and testing vs. number of epochs in training for DIQaM-NR and WaDIQaM-NR. ©2018 IEEE

The influence of the number of patches N_p on the prediction performance of DIQaM-NR and WaDIQaM-NR is plotted in Fig. 4.7. For both pooling methods and on both databases SROCC improves monotonically with increasing number of patches N_p until saturation.
On LIVE, for DIQaM-NR saturation sets in at about $N_p \approx 16$ to reach its maximal performance,

whereas WaDIQaM-NR reaches its maximal performance at $N_p \approx 256$. Over the whole range of N_p the performance of average patch aggregation is superior to the performance of weighted average patch aggregation and the difference is largest for small numbers N_p. This is because the weighted average acts as a filter that ignores patches with lower weights and thereby reduces the number of patches considered in quality estimation. Qualitatively the same results are obtained on TID2013.
Fig. 4.8 shows the course of optimization loss (effectively the MAE) during training, validation and testing in dependence of the number of epochs of training exemplified for DIQaM-NR and WaDIQaM-NR, one random split each, trained on LIVE. For both spatial pooling methods, the loss shows the typical behavior for iterative gradient descent minimization. Interestingly, WaDIQaM-NR achieves a lower loss than DIQaM-NR during training, while losses during validation and testing are approximately equal.

4.3.6 Feature Fusion

Table 4.8 – PLCC Comparison for different feature fusion schemes ©2018 IEEE

Dataset	Method	$\mathbf{f}_i^d - \mathbf{f}_i^r$	concat $(\mathbf{f}_i^r, \mathbf{f}_i^d)$	concat $(\mathbf{f}_i^r, \mathbf{f}_i^d, \mathbf{f}_i^d - \mathbf{f}_i^r)$
LIVE	DIQaM-FR	0.976	0.974	0.976
	WaDIQaM-FR	0.982	0.977	0.982
TID2013	DIQaM-FR	0.908	0.893	0.908
	WaDIQaM-FR	0.962	0.958	0.965

Results presented for the FR models in the previous subsections were obtained employing $\texttt{concat}(\mathbf{f}_i^r, \mathbf{f}_i^d, \mathbf{f}_i^d - \mathbf{f}_i^r)$ as feature fusion scheme. The performances of the three feature fusion schemes are reported for LIVE and TID2013 in Table 4.8. Mere concatenation of both feature vectors does not fail but consistently performs worse than both of the fusion schemes exploiting the explicit difference of both feature vectors. This suggests that while the model is able to learn the relation between the two feature vectors, explicitly providing the relation improves performance. The results do not provide enough evidence for preferring one over the other feature fusion methods, but it appears that adding the original feature vectors explicitly to the representation does not add useful information. Note that the feature fusion scheme might affect the learned features as well, thus, other things equal, it is not guaranteed the extracted features $\mathbf{f}_i^r$ and $\mathbf{f}_i^d$ are equal for different fusion methods.

4.3.7 Network Depth

Comparing the proposed NR approach to CNN-based approaches employing fewer layers [Kang et al., 2014] (see Table 4.1) suggests that the performance of a neural network-based IQM can be increased by making the network deeper. In order to evaluate this observation in a FR context as well, the architecture of the WaDIQaM-FR network was modified by removing several layers and by reducing the intermediate feature vector dimensionality from 512 to 256.

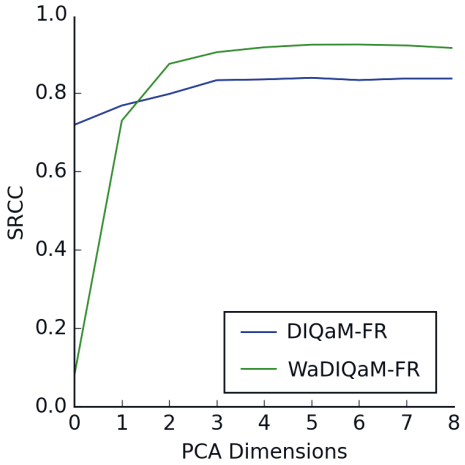


Figure 4.9 – Average performance on TID2013 for patch-wise dimensionality reduced (Wa)DIQaM-FR in terms of SROCC in dependence of the number of principal components of the reference patch feature vector. ($N_p = 32$) ©2018 IEEE

This amounts to the architecture conv3-32, conv3-32, maxpool, conv3-64, maxpool, conv3-128, maxpool, conv3-256, maxpool, FC-256, FC-1 with in total ~790k parameters instead of ~5.2M parameters in the original architecture. When tested on the LIVE database, the smaller model achieves a linear correlation of 0.980, whereas the original architecture achieves 0.984. The same experiment on TID2013 shows a similar result as the shallow model obtains a linear correlation of 0.949, compared to 0.953 obtained by the deeper model. To test whether the decrease in model complexity leads to less overfitting and better generalization, the models trained on TID2013 are additionally tested on CSIQ. The smaller model achieves a SROCC of 0.911, and is outperformed by the original architecture with a SROCC of 0.927. The differences are rather small, but indicate that the deeper and more complex model does lead to a more accurate prediction. However, when computational efficiency is important, small improvements might not justify the five-fold increase in the number of parameters and computations.

4.3.8 Bridging from Full- to No-Reference Image Quality Estimation

If argued strictly, the (Wa)DIQaM-FR as used here is not a FR, but a RR IQM, as only $N_p = 32$ 32×32 patches but not the full image is used for quality estimation. As shown in Fig. 4.6, reference information could be reduced even further by reducing the number of patches N_p considered. This can be done without re-training the model. In the proposed architecture the feature vector extracted from one reference patch is 512-dimensional.
The information available from the reference patch can not only be controlled by steering N_p, but also by reducing the dimensionality of the extracted feature vector. A straight forward approach to do so would be to systematically change the network architecture from (Wa)DIQaM-FR to (Wa)DIQaM-NR by reducing the dimensionality of the number of neurons

in the last layer of reference branch of the feature extraction module. However, this approach would result in a multitude of models, each trained for a specific patch-wise feature dimensionality.
Another approach is to start with a trained FR model and to linearly reduce the dimensionality of the reference patch feature vector $\mathbf{f}_i^r$ using PCA [Hotelling, 1933]. That way, a network trained for FR quality estimation could be used as a NR quality estimation method. In this extreme case the reference feature vector $\mathbf{f}_i^r$ is reduced to the mean of the reference feature vectors observed in the training data.
PCA is estimated based on the feature vectors of 4000 reference patches sampled from the training set and used for patch-wise dimensionality reduction of $\mathbf{f}_i^r$ during testing. Fig. 4.9 shows the performance of this RR IQM on one TID2013 test split for increasing dimensionality of the reference patch feature vectors. While the dimensionality reduced version of DIQaM-FR is still able to make useful predictions even without any reference information, this is not the case for the dimensionality reduced version of WaDIQaM-FR method. This corroborates the previous conjecture that weighted average patch aggregation, i.e. the reliable estimation of the weights, is more dependent on information from the reference image, at least for homogeneous distortions. However, 3 principal components (dimensions) are enough to recover the approximate performance obtained with the 512-dimensional original feature vector, for both DIQaM-FR and WaDIQaM-FR. Note that this is the patch-wise, not the image-wise feature dimensionality.
Although there are studies analyzing the influence of the feature dimensionality on the performance of RR IQM systematically [Soundararajan and Bovik, 2012], we are not aware of any unified framework to study NR to FR approaches. However, albeit being a promising framework, a fair comparison, e.g. to the RRED indices [Soundararajan and Bovik, 2012] would require an analysis of the interaction between patch-wise feature dimensionality and number of patches N_p considered.

4.4 Discussion

This chapter presented a neural network-based approach to FR and NR quality estimation that allows for joint feature and regression learning in an end-to-end framework. To achieve this, novel network architectures were presented, incorporating an optional joint optimization of weighted average patch aggregation that implements a simple yet efficient method for linear pooling of local patch quality estimates to a global image quality estimate. To allow for FR image quality estimation, different feature fusion strategies were studied. The experimental results show that the proposed methods outperform other state-of-the-art approaches for NR as well as for FR quality estimation and achieve generalization capabilities that can compete with state-of-the-art data-driven approaches. However, as for all current data-driven methods generalization performance offers considerable room for improvement. A principle problem for data-driven quality estimation is the relative lack of data. Until larger databases become available networks could undergo unsupervised pre-training by optimizing on reproducing

the quality prediction of a FR IQM, and a pre-trained network employing patch-wise weighting could be refined by end-to-end training.
Even though a relatively generic neural network can achieve high prediction performance, incorporating task specific adaptations to the architecture may lead to further improvements. Our results show that there is room for optimization in terms of feature dimensionality and balancing the ratio between network parameters. Here it is important to study prediction performance and generalization ability. In this work, we optimized based on MAE. However, quality estimation is commonly evaluated in terms of correlation and a different loss function might be beneficial. We sketched how the proposed framework could be used in the RR case. We did not present a full-fledged solution, but believe that the results indicate an interesting starting point.
Local weighting of distortion is not a new concept. Classical approaches usually compute the local weights based on a saliency model from the reference image [Zhang et al., 2016], or the reference and the distorted image [Zhang et al., 2014b]. Selecting an appropriate saliency model is crucial for the success of this strategy and models that excel at predicting saliency are not necessarily best for quality estimation [Zhang et al., 2016]. The proposed weighted average patch aggregation method allows for a joint, end-to-end optimization of local quality estimation and weight assignment.
The equivalence of our weighting scheme to Eq. 3.2 allows us to interpret the two learned maps as a weighting map and a quality estimate map. Thus, formal equivalence with the classical approach of linear distortion weighting suggests that local weights capture local saliency. However, this is not necessarily true, as the optimization criterion is not saliency, but image quality. In fact, we showed that the local weights not only depend on the structural (and potentially semantical) organization of the reference image, but also on the distortion type and the spatial distribution of the distortion. This is a fundamental problem for quality estimation (and as well as for quality assessment) and future work should address the conceptual similarity between the learned weights and saliency models in greater detail.
The proposed network could be adapted for end-to-end learning local weights alone to be used in a weighting scheme used to improve the prediction performance of any given IQM (cf. Section 3.1.3). This could be directly combined with signal-based adaptation to the conventional regression scheme [Bosse et al., 2017c] and will be studied in detail in Chapter 5. Explanation methods [Bach et al., 2015, Montavon et al., 2018] can be applied to better understand what features were actually learned by the network. From an application-oriented perspective the proposed method may be adapted and evaluated for quality estimation of 3D images and 2D and 3D videos.
Although there are still many obstacles and challenges for purely data-driven approaches, the performance of the presented approach, considering its relative simplicity and the fact that no domain knowledge is required, suggests that neural networks used for visual quality estimation have lots of potential and are here to stay.

4.5 Lessons Learned

- Deep CNNs are a feasible tool for FR and NR image quality estimation allowing for joint end-to-end learning of features and regression.
- Linear pooling can be efficiently integrated into CNN quality estimation and for joint optimization.
- For most distortion types the information relevant for linear pooling is contained in the reference image.
- Distorted image carry pooling relevant information for spatially inhomogeneous distributed distortions.
- Deep CNNs show higher quality prediction accuracy than shallow CNNs.
- CNN-based quality estimation allows study the RR space between FR and NR quality estimation systematically.
- Larger databases are necessary to train reliable models.

5 Perceptual Distortion Sensitivity for Quality Estimation

This chapter is based on

Bosse, S., Becker, S., Müller, K.-R., Samek, W., and Wiegand, T. (2018d). Estimation of distortion sensitivity for visual quality prediction using a convolutional neural network. *Digital Signal Processing, submitted*

Bosse, S., Becker, S., Fisches, Z., Samek, W., and Wiegand, T. (2018c). Neural network-based estimation of distortion sensitivity for image quality prediction. In *Proceeding of the IEEE International Conference on Image Processing (ICIP), accepted for publication* ©2018 IEEE

Bosse, S., Siekmann, M., Samek, W., and Wiegand, T. (2017c). A perceptually relevant shearlet-based adaptation of the PSNR. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pages 315–319 ©2017 IEEE

Bosse, S., Helmrich, C., Schwarz, H., Marpe, D., and Wiegand, T. (2017b). Perceptually optimized QP adaptation and associated distortion measure. In *Joint Video Exploration Team (JVET) of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, JVET-H0047*, Macao, China

5.1 Introduction

The previous chapter proposed a fully data-driven, end-to-end trained approach to NR and FR image quality estimation and it was shown that the accuracy of quality estimation benefits from a spatially weighted pooling scheme. This weighting scheme was optimized jointly with the local patch-wise quality estimation.
Also in general FR IQMs benefit from adaptation to the specific content of the images to be tested [Ortiz-Jaramillo et al., 2015], and this adaption is mostly implemented as a formally equivalent weighting scheme. Different models for this weighting have been proposed, e.g. by considering HVS models such as saliency [Zhang et al., 2016] or scale-wise divisive normalization [Laparra et al., 2016], information content [Wang and Li, 2011], conditional probability [Hu et al., 2015] or shearlet-based measurements of local activity [Bosse et al., 2017c]. These

weighing schemes model different aspects of the HVS but relate to the same concept of *distortion sensitivity*, suggesting that distortions measured by a given quality model are more (or less) visible in one image area than in another and that, thus, this image area is more (or less) distortion sensitive than another.
Chapter 4 proposed an end-to-end optimized neural network-based approach to image quality. Although superior prediction performances were achieved, the computational complexity of this approach (as of many other approaches in the literature) render its use infeasible for many time-critical applications.
In this chapter, the concept of distortion sensitivity is introduced to bridge from psychophysical quality assessment over computational quality estimation to perceptually relevant encoder control for image and video compression. Distortion sensitivity is modelled as a feature of the reference image. This is particularly appealing as in combination with a low complex quality model, i.e. the MSE or PSNR, computationally complex processing could be restricted to the reference image only. For time-critical applications such as block-based hybrid video coding, where during mode decision the block-wise distortion is evaluated for every coding mode considered [Wiegand and Schwarz, 2016], this is a crucial property, as complex processing would be gracefully taken out of the search loop of mode decision.
In Section 5.2, a functional definition of distortion sensitivity is derived based on a conceptual and statistical discussion of the parameters of the regression function mapping from the computational to the perceptual domain (cf. Section 3.1). Exemplified for the PSNR, it is shown how a full image-wise compensation of distortion sensitivity significantly improves prediction accuracy and the limits of improvement are explored based on the LIVE database [Sheikh et al., 2006] in an exploration setup. The concept of distortion sensitivity is adapted from a global to a local scale, and in Section 5.3 a neural network-based approach for estimating local distortion sensitivity in an end-to-end trained image quality prediction framework is presented. The performance of the presented approach for neural network-based compensation for distortion sensitivity is evaluated and compared to other relevant approaches on the LIVE [Sheikh et al., 2006] and the TID2013 [Ponomarenko et al., 2013] databases in Section 5.4. In Section 5.5, a short introduction into image and video compression is given in order to derive a distortion sensitivity bit allocation scheme that is evaluated and compared in an image compression framework. Section 5.6 concludes the chapter with a summary and discussion.

5.2 Distortion Sensitivity

5.2.1 Psychometric Relation between Computational and Perceptual Quality

Due to saturation effects in the extreme cases of imperceptible quality loss or strong impairments, subjective image quality ratings typically do not relate linearly to many computational quality measures. As discussed in Section 3.1 and shown in Fig. 3.1a, the relation is commonly linearized by a nonlinear mapping from the computational to the perceptual domain. A widely used function is the 4-parameter generalized logistic function [VQEG, 2004]

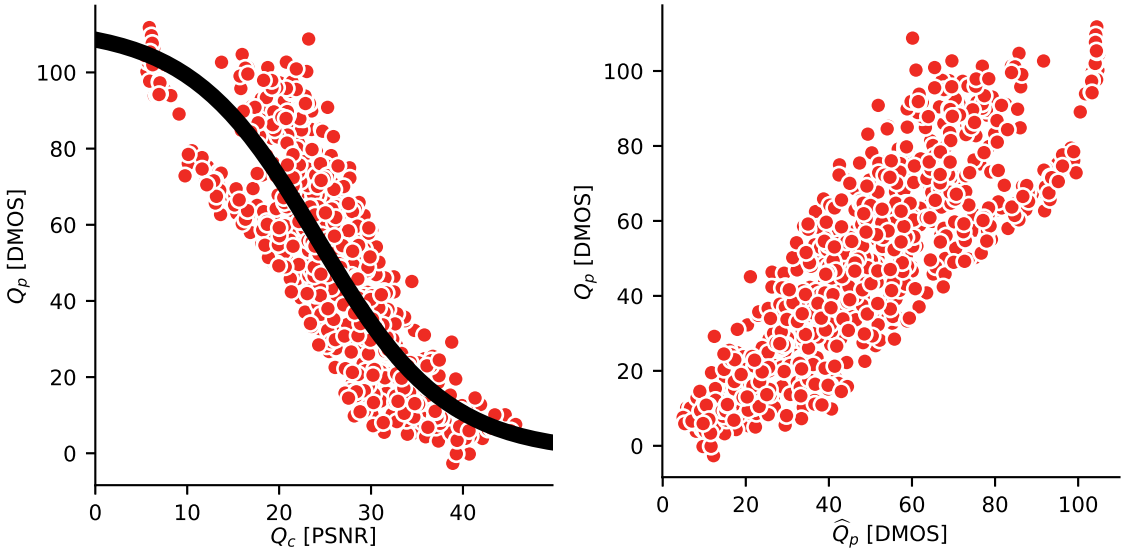


Figure 5.1 – Relation between Q_c, Q_p and $\widehat{Q}$. **Left**: Mapping of Q_c to Q_p for the LIVE database. Red circles indicate Q_c vs. Q_p for individual images, the black curve shows the resulting regression function for the full set. **Right**: Resulting quality predictions $\widehat{Q}_p$ vs. true quality values Q_p.

$$\begin{aligned} Q_p &= f(Q_c;\boldsymbol{\beta}) \\ &= \beta_0 + \frac{\beta_1 - \beta_0}{1 + e^{-\beta_2 \cdot (Q_c - \beta_3)}}. \end{aligned} \tag{5.1}$$

Parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$ are estimated as $\widehat{\boldsymbol{\beta}}$ based on image-wise pairs of computational quality values Q_c, output of a computational quality model, and perceptual quality values Q_p, output of a quality assessment, e.g. a psychophysical test. Resulting estimates of the regression parameters are then used to predict perceptual quality values from computational quality values as

$$\widehat{Q}_p = f(Q_c;\widehat{\boldsymbol{\beta}}). \tag{5.2}$$

Regression parameters[1] $\boldsymbol{\beta}$ are not valid globally, but dependent on the quality assessment procedure used to obtain Q_p and the quality model computing Q_c, where the consistency of the relation between Q_p and Q_c relies on the goodness of the computational quality model. In practice, regression parameters can only be estimated on a limited number of images that need to be sufficiently representative in order to ensure generalization of the prediction to

[1] In order to simplify notation, the $\widehat{\ }$-sign is dropped from now on and estimated regression parameters $\widehat{\boldsymbol{\beta}}$ are referred to as as $\boldsymbol{\beta}$.

unseen images.
Fig. 5.1 exemplifies a typical regression based on Eq. 5.1 from computational to perceptual quality (left) and the resulting prediction of perceptual quality from computational quality (right) with Q_c calculated as PSNR and $\boldsymbol{\beta}$ estimated on the full LIVE database [Sheikh et al., 2006] (see Section 3.2.1). Red circles denote pairs of (Q_p, Q_c) or $(Q_p, \widehat{Q}_p)$ respectively, for individual images. The black line represents the estimated regression function from Q_c to $\widehat{Q}_p$.

Although estimation of regression parameters is typically data-driven, β_0 and β_1 relate directly to the lower and upper bounds of the perceptual quality values. As such, β_0 and β_1 are mainly determined by the range of the perceptual quality scale and, thus, defined by the experimental design of the subjective test and therefore in principle known a-priori. Regression parameter β_3 denotes a horizontal shift of the regression function with respect to Q_c. The slope of the regression, which, with a value of $\frac{\partial \widehat{Q}_p}{\partial Q_c}(Q_c = \beta_3) = \frac{\beta_1 - \beta_0}{4} \cdot \beta_2$, is steepest at $Q_c = \beta_3$, is controlled by β_2 and scaled by the range of β_0 to β_1. Disregarding this scaling, β_2 and β_3 are not depending on the quality scale, but on the relation between the values of a specific quality measure and the ground-truth quality scores for the image set used to estimate the regression parameters. Hence, β_0, β_1 in Eq. 5.1 can be fixed to the lower and upper bound of the rating scale a and b and $\boldsymbol{\beta}$ can be reduced to $\boldsymbol{\beta} = (\beta_2, \beta_3)$.

5.2.2 Distortion Sensitivity as an Image Property

Regression parameters are commonly estimated over a set of images based on an ensemble of reference images that are subject to different distortion types at different distortion magnitudes. However, given enough samples, i.e., impairment levels, regression parameters $\boldsymbol{\beta}^{i,d}$ can also be found per reference image i and distortion type d. Note that in practice this would result in the loss of any generalization ability.
Such a reference image specific estimation of $\boldsymbol{\beta}$ is shown in Fig. 5.2 for JPEG-distorted images from the LIVE database [Sheikh et al., 2006] with Q_c measured as PSNR. The database provides Q_c as DMOS, high values of DMOS denote low subjective quality. Circles denote (Q_c, Q_p) pairs of distorted images and are colored according to the base reference image. Colored dashed curves represent the regression functions estimated for the different reference images, the black curve represents the regression function estimated for the full ensemble. Reference-specific regression curves are widely dispersed around the ensemble-wide regression. This gives raise to the notion of *distortion sensitivity*, as for a given PSNR distorted versions of some reference images exhibit a rather high perceptual quality, while others are reported to appear highly distorted. This is indicated for the extreme cases by vertical black arrows; with regard to the PSNR, the relatively flat image of the sailing boat, represented by green, is perceptually more sensitive towards JPEG distortions than the highly textured image, represented by orange. Based on the previous interpretation of the regression parameters $\boldsymbol{\beta}$ (and as such $\boldsymbol{\beta}^{i,d}$) and the insight that β_0, β_1 are only dependent on the experimental setup for quality assessment, distortion sensitivity can be functionally captured by β_2 and β_3. Hypothetical compensation for the shifting parameter β_3 is sketched for two reference images by dashed black horizontal

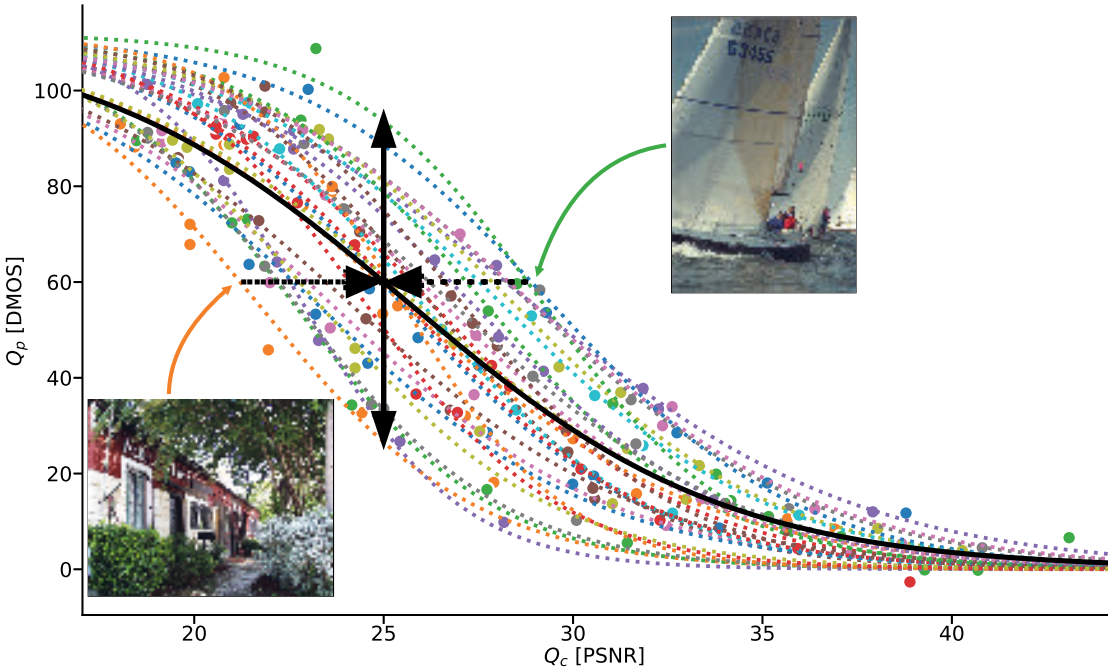


Figure 5.2 – PSNR vs. DMOS for the JPEG-subset of the LIVE database [Sheikh et al., 2006]. High values of DMOS denote low subjective quality. Colored dashed curves and circles indicate regressed and measured DMOS values for individual reference images. The thick black curve shows regressed DMOS values for the whole ensemble. Examples images are given for the two extreme cases of distortion sensitivity.

arrows in Fig. 5.2.

With a functional quantification (for simplicity neglecting different distortion types for the moment) of distortion sensitivity ϑ_0^i and ϑ_1^i of a reference image i such a compensation can be used to adapt a computational quality value Q_c as

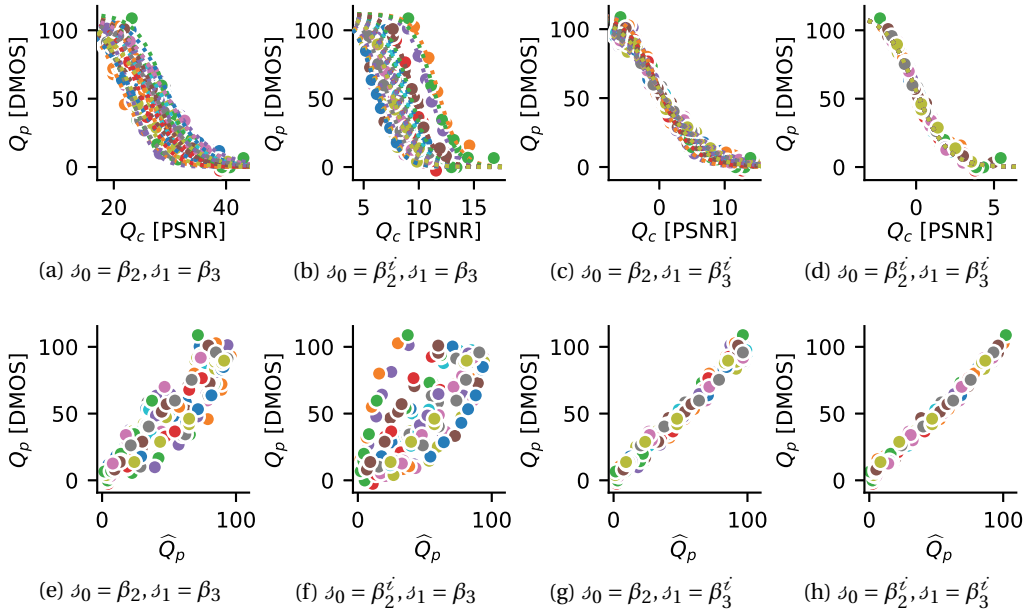
$$Q_{ac} = \vartheta_0 \cdot (Q_c - \vartheta_1). \tag{5.3}$$

Assuming a regression according to Eq. 5.1, β_2^i and β_3^i are optimal predictors of ϑ_0 and ϑ_1. With β_0 and β_1 being the upper and lower bounds a and b of the rating scale, Eq. 5.2 can be rewritten as

$$\begin{aligned} \widehat{Q}_p &= a + \frac{b-a}{1+e^{-\vartheta_0 \cdot (Q_c - \vartheta_1)}} \\ &= a + \frac{b-a}{1+e^{-Q_{ac}}}. \end{aligned} \tag{5.4}$$

Although β_2^i and β_3^i are generally not available in practice, assuming their availability helps to

analyse the influence and limits of full image-wise distortion sensitivity-based compensation in quality estimation. For this we distinguish four different cases in which we assume available *a*) no reference image-specific information: $\vartheta_0 = \beta_2, \vartheta_1 = \beta_3$; *b*) optimal estimation of ϑ_0 only: $\vartheta_0 = \beta_2^i, \vartheta_1 = \beta_3$; *c*) optimal estimation of ϑ_1 only: $\vartheta_0 = \beta_2, \vartheta_1 = \beta_3^i$; and *d*) optimal estimation of ϑ_0 and ϑ_1: $\vartheta_0 = \beta_2^i, \vartheta_1 = \beta_3^i$, where $\beta_{(\cdot)}$, in contrast to $\beta_{(\cdot)}^i$, denotes a parameter estimation over the full ensemble of reference images. Note that, with regard to correlations between Q_p and Q_{ac}, $\vartheta_0 = \beta_2, \vartheta_1 = \beta_3$ and $\vartheta_0 = 1, \vartheta_1 = 0$ are equivalent, but not with regard to the Pearson correlations between Q_p and $\widehat{Q}_p$. Hence, for simplified, yet consistent notation the no adaptation case is represented as $\vartheta_0 = \beta_2, \vartheta_1 = \beta_3$.



(a) $\vartheta_0 = \beta_2, \vartheta_1 = \beta_3$ (b) $\vartheta_0 = \beta_2^i, \vartheta_1 = \beta_3$ (c) $\vartheta_0 = \beta_2, \vartheta_1 = \beta_3^i$ (d) $\vartheta_0 = \beta_2^i, \vartheta_1 = \beta_3^i$

(e) $\vartheta_0 = \beta_2, \vartheta_1 = \beta_3$ (f) $\vartheta_0 = \beta_2^i, \vartheta_1 = \beta_3$ (g) $\vartheta_0 = \beta_2, \vartheta_1 = \beta_3^i$ (h) $\vartheta_0 = \beta_2^i, \vartheta_1 = \beta_3^i$

Figure 5.3 – Influence of compensating the PSNR for distortion sensitivity on JPEG subset of LIVE database. **Top**: Adapted PSNR vs. ground truth DMOS. **Bottom**: Estimated DMOS compensated for distortion sensitivity vs. ground truth DMOS.

Table 5.1 – Correlations between PSNR compensated for distortion sensitivity and ground truth DMOS (Q_{ac} vs. Q_p), and predicted DMOS compensated for distortion sensitivity and ground truth DMOS ($\widehat{Q}_p$ vs. Q_p). All correlations are calculated on the JPEG subset of the LIVE database.

		Q_{ac} vs. Q_p		$\widehat{Q}_p$ vs. Q_p	
		ρ_P	ρ_S	ρ_P	ρ_S
$\vartheta_0 = \beta_2,$	$\vartheta_1 = \beta_3$	-0.88	-0.9	0.9	0.9
$\vartheta_0 = \beta_2^i,$	$\vartheta_1 = \beta_3$	-0.71	-0.72	0.73	0.72
$\vartheta_0 = \beta_2,$	$\vartheta_1 = \beta_3^i$	-0.96	-0.98	0.98	0.98
$\vartheta_0 = \beta_2^i,$	$\vartheta_1 = \beta_3^i$	-0.96	-0.99	0.99	0.99

The effect of compensating the PSNR for distortion sensitivity is shown in Fig. 5.3 for JPEG compressed images from the LIVE database: The top row shows the adapted PSNR (Eq. 5.3)

vs. the ground truth DMOS , the bottom row the predicted DMOS (Eq. 5.4) vs. the true DMOS for previously defined assumptions, i.e. the left hand sided column (Fig. 5.3a and Fig. 5.3e) is equivalent to no adaptation. Fig. 5.3b and Fig. 5.3f suggest that image-wise compensation for the slope disperses the quality estimates even further, while compensating image-wise for the offset (Fig. 5.3b and Fig. 5.3f) and even more a joint compensation for slope and offset (Fig. 5.3d and Fig. 5.3h) achieves a clean alignment of quality estimates.
Corresponding correlations are summarized in Table 5.1 and corroborate this observation. It is noteworthy that a joint compensation for slope and offset achieves only small additional improvement over offset-only compensation.

5.2.3 Distortion Sensitivity and Different Distortion Types

The previous subsection discussed reference image-specific distortion sensitivity subject to a specific distortion type and exemplified this by JPEG distortion. However, different distortion types affect different statistical properties of natural images, hence, also the distortion type may have an influence on distortion sensitivity. This can be accounted for by extending previous considerations and modelling distortion sensitivity not only as a property of a reference image i with respect to a given computational quality measure, but also in dependency of a specific distortion type d.

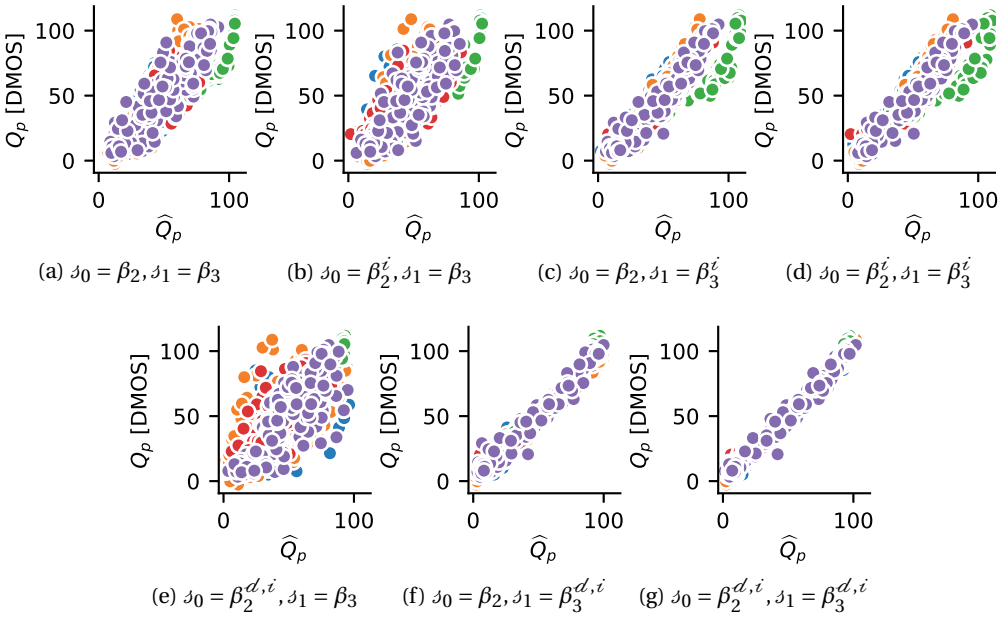


Figure 5.4 – Influence of considering distortion sensitivity on the adapted PSNR for different distortion types. **Top**, from left to right: Distortion type-agnostic consideration of ϑ_0 only, ϑ_1 only, and ϑ_0, ϑ_1 jointly. **Bottom**, from left to right: Distortion type-specific consideration of ϑ_0 only, ϑ_1 only, and ϑ_0, ϑ_1 jointly.

Fig. 5.4 plots the relation between the estimated quality $\widehat{Q}_p$ and the ground truth quality Q_p

Table 5.2 – Correlation between adapted PSNR and true DMOS (Q_{ac} vs. Q_p) and predicted DMOS and true DMOS ($\widehat{Q}_p$ vs. Q_p) for different adaptions of Q_c by considering neither ϑ_0 nor ϑ_0, only ϑ_0, only ϑ_1, both ϑ_0, ϑ_1 when accounting for specific distortion types $\mathscr{d}$ or over the set of all distortion types $\mathscr{D}$.

				Q_{ac} vs. Q_p		$\widehat{Q}_p$ vs. Q_p	
				ρ_P	ρ_S	ρ_P	ρ_S
$\mathscr{d}$	agnostic	$\vartheta_0 = \beta_2$,	$\vartheta_1 = \beta_3$	-0.84	-0.87	0.86	0.87
		$\vartheta_0 = \beta_2^i$,	$\vartheta_1 = \beta_3$	-0.80	-0.83	0.81	0.83
		$\vartheta_0 = \beta_2$,	$\vartheta_1 = \beta_3^i$	-0.88	-0.93	0.90	0.93
		$\vartheta_0 = \beta_2^i$,	$\vartheta_1 = \beta_3^i$	-0.88	-0.94	0.91	0.94
$\mathscr{d}$	specific	$\vartheta_0 = \beta_2^{i,\mathscr{d}}$,	$\vartheta_1 = \beta_3$	-0.52	-0.50	0.77	0.77
		$\vartheta_0 = \beta_2$,	$\vartheta_1 = \beta_3^{i,\mathscr{d}}$	-0.93	-0.96	0.98	0.99
		$\vartheta_0 = \beta_2^{i,\mathscr{d}}$,	$\vartheta_1 = \beta_3^{i,\mathscr{d}}$	-0.96	-0.99	0.99	0.99

for different distortion sensitivity compensation schemes, where again $\beta_{(\cdot)}$ (without superscript) denotes a parameter estimated over the full dataset, $\beta_{(\cdot)}^i$ denotes a reference image-wise estimation over all distortion types in the database, and $\beta_{(\cdot)}^{\mathscr{d},i}$ denotes parameter estimation per reference image i and distortion type $\mathscr{d}$. Clearly, a joint compensation of distortion type $\mathscr{d}$ and reference image i can improve the prediction accuracy. Corresponding correlations are summarized in Table 5.2. Interestingly, as observed previously in Table 5.1 for the single distortion case, a compensation solely based on the slope parameter decreases prediction accuracy also in the multi-distortion case, be it estimated per reference image over all distortions type ($\vartheta_0 = \beta_2^i, \vartheta_1 = \beta_3$) or per reference image and distortion type ($\vartheta_0 = \beta_2^{\mathscr{d},i}, \vartheta_1 = \beta_3$). Compensation for the offset over all distortions per reference image ($\vartheta_0 = \beta_2, \vartheta_1 = \beta_3^i$) improves prediction accuracy, also considering the distortion type ($\vartheta_0 = \beta_2, \vartheta_1 = \beta_3^{\mathscr{d},i}$) further improves the quality estimation. However, a joint consideration of slope and offset ($\vartheta_0 = \beta_2^i, \vartheta_1 = \beta_3^i$ and $\vartheta_0 = \beta_2^{\mathscr{d},i}, \vartheta_1 = \beta_3^{\mathscr{d},i}$) achieves only little additional improvement.

The discussion and findings presented in this section suggest distortion sensitivity can be efficiently modelled as a feature of a reference image and functionally captured based on the shifting parameter of the 4-parameter generalized logistic function. Additional image-wise compensation for the slope parameter achieves only little further improvements in prediction accuracy. Hence, in the following only the shifting parameter will be considered as a functional representation of distortion sensitivity. For simplified notation β_2 is replaced by c. This modifies Eq. 5.3 and Eq. 5.4 to

$$Q_{ac} = Q_c - \vartheta. \tag{5.5}$$

and

$$\begin{aligned}\widehat{Q}_p &= a + \frac{b-a}{1+e^{-c\cdot(Q_c-\vartheta)}} \\ &= a + \frac{b-a}{1+e^{-cQ_{ac}}},\end{aligned} \tag{5.6}$$

where c is estimated over full datasets and distortion sensitivity is denoted as ϑ.

5.2.4 Localized Distortion Sensitivity

Previous considerations studied distortion sensitivity as a full image feature. However, statistics of natural images are locally structured and spatially highly non-stationary [Ruderman, 1994, Bell and Sejnowski, 1997] so that distortion sensitivity not only varies globally across different images, but also spatially within a given image.
Although in principle applicable to any computation distortion measure, the PSNR allows for a very simple consideration of local distortion sensitivity. According to Eq. 5.5 the PSNR is compensated for distortion sensitivity and the perceptually imagewise adapted PSNR (paPSNR_I) written as

$$\begin{aligned}\text{paPSNR}_I &= \text{PSNR} - \vartheta_I \\ &= 10\cdot\log_{10}\frac{C^2}{10^{\frac{\vartheta_I}{10}}\text{MSE}},\end{aligned} \tag{5.7}$$

with ϑ_I denoting the image-wise distortion sensitivity and C the maximum (peak) sample value of the given signal class, e.g. for 8-bit SDR images $C = 255$. While PSNR and paPSNR_I do not allow for a direct local weighting, the MSE can be adopted image-wise to the perceptually adapted MSE (paMSE_I)

$$\text{paMSE}_I = 10^{\frac{\vartheta_I}{10}}\cdot\text{MSE}. \tag{5.8}$$

By localizing distortion sensitivity to a pixel position (x,y) as $\vartheta(x,y)$, with $s(x,y)$ being the reference and $\tilde{s}(x,y)$ the distorted image samples, we define the perceptually adapted MSE (paMSE) as

$$\text{paMSE} = \frac{1}{M\cdot N}\sum_{x=0}^{M-1}\sum_{y=0}^{N-1}10^{\frac{\vartheta(x,y)}{10}}(s(x,y)-\tilde{s}(x,y))^2 \tag{5.9}$$

leading directly to a perceptually adapted PSNR (paPSNR)

$$\text{paPSNR} = 10\cdot\log_{10}\frac{C^2}{\text{paMSE}}. \tag{5.10}$$

Note that when distortion sensitivity is available only globally for a full image, with $\vartheta(x,y) = \vartheta_I$ then Eq. 5.10 simplifies to Eq. 5.7.

The resulting compensation for local distortion sensitivity is very similar to the weighting scheme introduced in Chapter 4, but does not employ a image-wise normalization of the weights (see also Section 3.1.3).
Due to the scarcity of samples, i.e. distortion levels per reference image, no performance limits can be derived for the local compensation of distortion sensitivity.

5.3 Estimating of Distortion Sensitivity using Neural Networks

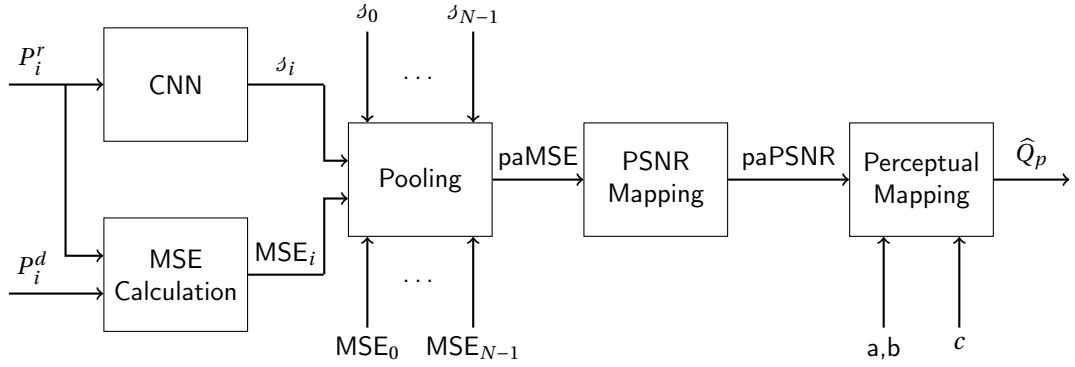


Figure 5.5 – CNN-based compensation of the PSNR for distortion sensitivity. Distortion sensitivity ϑ_i is estimated by the CNN from the reference patch P_i^r. The image-wise paPSNR is calculated from the sensitivity-weighted MSE of all image patches and mapped into the perceptual domain on the quality estimate $\widehat{Q}_p$.

The neural network used for end-to-end trained image quality estimation in Chapter 4 is re-used for the estimation of patch-wise distortion sensitivity. Input to the network are 32×32 pixel-sized patches of the gray-scale converted reference image. The proposed CNN comprises 12 weight layers that are used to estimate the distortion sensitivity ϑ_i of a given reference image patch P_i^r. The network is organized as a series of conv3-32, conv3-32, maxpool, conv3-64, conv3- 64, maxpool, conv3-128, conv3-128, maxpool, conv3-256,conv3-256, maxpool, conv3-512, conv3-512, maxpool layers, followed by FC-512, FC-1 layers. Convolutional layers are activated through a Leaky Rectified Linear Unit (LReLU) activation function [He et al., 2015] with a leakyness of 0.2.
To allow for the estimation of distortion sensitivity for patch sizes other than 32×32 pixels, the network architecture is adapted for the processing of patches of *a)* 8×8 pixels by removing the first two pooling layers; *b)* 16×16 pixels by removing the first pooling layer; *c)* 64×64 pixels by introducing an additional pooling layer succeeding the 7th convolution layer; and *d)* 128×128 pixels by introducing two additional pooling layers succeeding the 7th and the 9th convolution layer.
Analogous to Section 5.2.4, the distortion sensitivity estimate ϑ_i output of the network is used to weight the patch-wise MSE_i, measured between a reference image patch P_i^r and the collocated image patch P_i^d from the distorted image. The resulting image-wise paMSE from Eq. 5.9 leads with Eq. 5.10 directly to the image-wise paPSNR. The image-wise paPSNR is

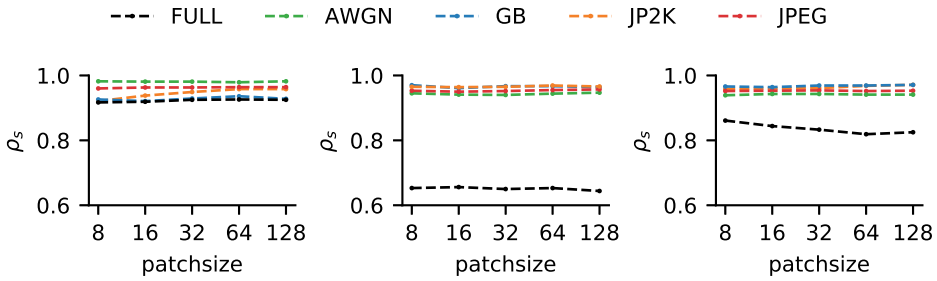
mapped into the perceptual domain by Eq. 5.6. Based on previous considerations, parameters a, b are fixed as the lower and upper value of the quality scale used in the psychophysical quality assessment; an additional parallel branch consisting of only 1 weight with a constant input of 1 is used for estimating a global value of c. The overall architecture is sketched in Fig. 5.5.
As in Chapter 4, the network is optimized by minimizing the MAE between reported and predicted perceptual quality

$$E = |\widehat{Q}_p - Q_p|. \tag{5.11}$$

5.4 Experiments and Results

5.4.1 Experimental Setup

As in Chapter 4, for single database-evaluation networks are trained and tested either on LIVE, TID2013, or CSIQ, cf. Section 3.2.1. Databases are randomly split in training, validation and test set. To guarantee that no distorted or undistorted version of an image used in testing or validation has been seen by the network during training, the datasets are split by reference image. For each database validation and test set each contain 6 reference images, whereas the training set consists of 17, 13 and 18 reference images for LIVE, TID2013 and CSIQ. Results are reported as the average over 30 random splits. Models are trained for 150 epochs after which the model with the lowest validation loss is selected and tested; this amounts to early stopping [Prechelt, 2012]. Training and validation of models with an input patch size of 32×32 pixels is based on 32 patches, randomly sampled from the images in each iteration. To keep the amount of data seen by the neural network in each training iteration constant for different patch sizes, the number of sampled patches per image is scaled inversely proportionally with the square of the patch size, i.e. 512 patches of 8×8 pixel, 128 patches of 16×16 pixels, 8 patches of 64×64 pixels and 2 patches of 128×128 pixels. Different to Chapter 4, patches are densely sampled, i.e. the full image is considered, for testing.
To assess the generalization ability of the proposed methods the CSIQ image database is used for cross-dataset evaluating the models trained either on LIVE or on TID2013 and models trained for single database evaluation were reused. LIVE and TID2013 share a lot of reference images, thus, tests between these two are unsuitable for evaluating generalization for unseen images. For cross-distortion evaluation, models trained on LIVE are tested on TID2013 in order to determine how well a model deals with distortions that have not been seen during training and in order to evaluate whether a method is truly non-distortion or just many-distortion specific.
Note that, in contrast to many results reported in the literature, if not explicitly stated differently, we use the full TID2013 database and do not ignore any specific distortion type.



(b) LIVE (c) TID2013 (d) CSIQ

Figure 5.6 – Influence of the patch-size on the prediction performance measured as SROCC on LIVE, TID2013 and CSIQ evaluated for selected distortion types (Gaussian blur, white Gaussian noise, JP2K and JPEG compression) and over the full databases.

5.4.2 Influence of Patch Size

In a first evaluation, the influence of the patch size is investigated for distortion types that are shared among LIVE, TID2013 and CSIQ and for the full databases. SROCC obtained with the proposed method is plotted with regard to the patch-size on which distortion sensitivity is estimated in Fig. 5.6. The prediction monotonicity is surprisingly little affected by the size of the patch on which distortion sensitivity is estimated. In accordance with the patch size used in Chapter 4, further results in this section are achieved based on distortion sensitivity estimation on 32×32 pixel sized patches.

5.4.3 Performance Evaluation

The performance of the presented paPSNR-based quality estimation is summarized and compared to related methods for selected distortion types on LIVE, TID2013 and CSIQ in terms of SROCC in Table 5.3. The proposed method clearly outperforms the PSNR for almost all distortion types and databases. An exception that is observable in all databases is additive white Gaussian noise (AWGN), for which the original PSNR is already a very good predictor and thus difficult to improve. Although applying complex processing on the reference image only, the SROCC of the proposed method is in general close to methods that perform complex processing on the distorted image as well.
Table 5.5 presents a comparison of the proposed method to state-of-the-art methods, evaluated on the full LIVE and TID2013 databases. Although the proposed method ($\text{paPSNR}_{\gamma=1}$) outperforms the PSNR on LIVE, its prediction accuracy is clearly inferior to all other approaches. Here, the distinction of the proposed approach from methods employing complex processing on the distorted image is important to note; the computational advantage of the proposed approach will be discussed in detail in later. In contrast to the single distortion results shown in Table 5.5, on TID2013 the proposed approach not only performs inferior to other sophisticated state-of-the-art approaches, but even worse as compared to the PSNR. This can be explained by the distortion type dependency of distortion sensitivity analyzed in

Table 5.3 – Average SROCC over 20 runs of the proposed method for the distortion types of LIVE and CSIQ databases and the *actual* subset of TID2013 in comparison to PSNR, SSIM [Wang et al., 2004], MS-SSIM [Wang et al., 2003], FSIM [Zhang et al., 2011] and HaarPSI [Reisenhofer et al., 2018].

		PSNR	SSIM	MS-SSIM	FSIM	HaarPSI	**paPSNR**
LIVE	JP2K	0.895	0.961	0.962	0.972	0.968	0.949
	JPEG	0.881	0.976	0.981	0.984	0.983	0.963
	AWGN	0.985	0.969	0.973	0.972	0.985	0.981
	GB	0.782	0.952	0.954	0.971	0.967	0.929
	FF	0.891	0.956	0.947	0.952	0.951	0.941
TID2013	AWGN	0.929	0.865	0.865	0.91	0.937	0.94
	SCN	0.92	0.852	0.854	0.89	0.931	0.944
	MN	0.832	0.777	0.807	0.809	0.786	0.856
	HFN	0.914	0.863	0.86	0.904	0.907	0.948
	IN	0.897	0.75	0.763	0.825	0.867	0.916
	GB	0.915	0.967	0.967	0.955	0.912	0.967
	DEN	0.948	0.925	0.927	0.933	0.947	0.943
	JPEG	0.919	0.92	0.927	0.934	0.951	0.952
	JP2K	0.884	0.947	0.95	0.959	0.97	0.965
	MGN	0.891	0.78	0.779	0.857	0.89	0.934
	LCNI	0.915	0.906	0.907	0.949	0.962	0.963
CSIQ	AWGN	0.936	0.897	0.947	0.936	0.967	0.943
	JPEG	0.888	0.955	0.963	0.966	0.97	0.954
	JP2K	0.936	0.961	0.968	0.97	0.982	0.961
	GPN	0.934	0.892	0.933	0.937	0.954	0.939
	GB	0.929	0.961	0.971	0.973	0.978	0.969
	CTRST	0.862	0.792	0.952	0.944	0.945	0.901

Section 5.2.3.
This distortion type dependency can be effectively approximated by simple linear scaling of $\mathfrak{s}$ with a distortion type-specific factor γ [Bosse et al., 2017c]. The scaling is incorporated as an additional trainable parameter into the sensitivity estimation described in Section 5.3 and γ is distortion type-specific jointly optimized with all other distortion type-agnostic parameters of the network. The resulting performance over the full dataset is referred to as $\text{paPSNR}_{\gamma=\gamma^*}$ in Table 5.5. Note that this evaluation relies on the (for most applications reasonable) assumption that the distortion type by which the test image is affected is known. As Table 5.5, considering distortion type dependency increases the prediction performance substantially, especially when tested on TID2013 containing a multitude of different distortion types.

5.4.4 Local Weights

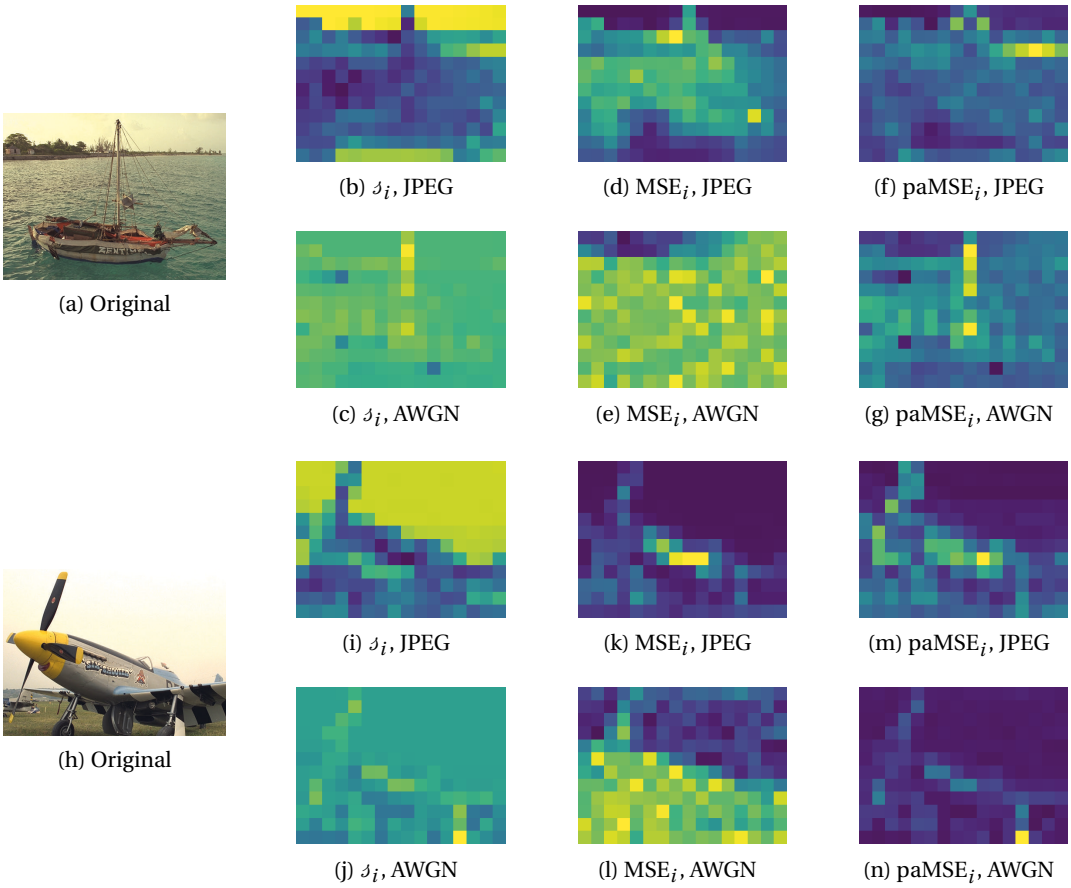
The spatial distribution of patch-wise estimated distortion sensitivity $\mathfrak{s}_i$ and the resulting distortion sensitive MSE is exemplified in Fig. 5.7 for two reference images and two distortion

Table 5.4 – Comparison of the proposed method to the state-of-the-art FR image quality estimation models based on the LIVE and TID2013 databases. The highest PLCC and SROCC are set in bold. The reported correlation are achieved on the test sets of 30 random train-test splits.

Table 5.5 – Performance Comparison on LIVE and TID2013 Databases

			LIVE		TID2013	
			LCC	SROCC	LCC	SROCC
Complex processing on distorted image	Yes	SSIM [Wang et al., 2004]	0.945	0.948	0.790	0.742
		$FSIM_C$ [Zhang et al., 2011]	0.960	0.963	0.877	0.851
		GMSD [Xue et al., 2014]	0.956	0.958	-	-
		DOG-SSIM [Pei and Chen, 2015]	0.963	0.961	0.919	0.907
		DeepSim [Gao et al., 2017]	0.968	**0.974**	0.872	0.846
		HaarPSI [Reisenhofer et al., 2018]	0.967	0.900	0.87	0.863
		DIQaM-FR (cf. Chapter 4)	0.977	0.966	0.88	0.859
		WaDIQaM-FR (cf. Chapter 4)	**0.980**	0.97	**0.946**	**0.940**
	No	PSNR	0.872	0.876	0.675	0.687
		$paPSNR_{\gamma=1}$ (proposed)	0.904	0.925	0.588	0.65
		$paPSNR_{\gamma=\gamma^*}$ (proposed)	0.938	0.943	0.863	0.876

types, namely JPEG compression and additive white Gaussian noise. Original images are presented in Fig. 5.7a and Fig. 5.7h, corresponding sensitivity maps for JPEG compression distorions in Fig. 5.7b and Fig. 5.7i and those for AWGN in Fig. 5.7c and Fig. 5.7j. Examples for patch-wise MSE maps are visualized in the second from right column, resulting paMSE maps in the right column of Fig. 5.7. Distortion sensitivity maps are presented in the same color scale representing values of ϑ_i from 21 to 34, thus are directly comparable. Local distortion sensitivities values lie in a range expected from Fig. 5.2. Color scales differ between the visualizations of different MSE and paMSE maps in order to use full ranges for each map.
Comparing the distortion sensitivity maps shows that for the case of JPEG distortions, local distortion sensitivity varies largely within the images. While low values of sensitivity are assigned to textured regions of the images, high values of sensitivity are estimated for rather flat areas, e.g. the sky in Fig. 5.7a and Fig. 5.7h. This is expected as distortions in textured regions are subject to masking effects, whereas JPEG-specific distortions such as blocking are highly visible in flat areas.
For the case of additive white Gaussian noise, local values of ϑ_i do not show this wide range of variation, but are relatively uniformly distributed over image. This suggests that, disregarding a global shift, the (unadapted) PSNR already is a good quality predictor for images affected by additive white Gaussian noise. This is in line with the numerical results presented in Section 5.4.3.



(a) Original (b) ϑ_i, JPEG (c) ϑ_i, AWGN (d) MSE_i, JPEG (e) MSE_i, AWGN (f) paMSE_i, JPEG (g) paMSE_i, AWGN (h) Original (i) ϑ_i, JPEG (j) ϑ_i, AWGN (k) MSE_i, JPEG (l) MSE_i, AWGN (m) paMSE_i, JPEG (n) paMSE_i, AWGN

Figure 5.7 – Examples of local distortion sensitivity for two reference images and two distortion types. The left-most column show the reference images from which patch-wise distortion sensitivity is estimated. The second from left column shows the resulting maps of distortion sensitivity for JPEG compression and AWGN distortions. In the second from right column the patch-wise MSE is shown, the perceptually adapted MSE resulting from patch-wise MSE and patch-wise distortion sensitivity is shown in the right-most column. Low values are represented by blue, high values by yellow. For comparability, colors are aligned for the distortion sensitivity maps.

Table 5.6 – Average SROCC over 100 runs of paPSNR trained and tested on different databases for selected distortion types and over full databases.

Trained on	LIVE		TID2013	
Tested on	TID2013	CSIQ	LIVE	CSIQ
JP2K	0.96	0.962	0.945	0.956
JPEG	0.923	0.958	0.949	0.935
AWGN	0.932	0.95	0.983	0.932
GB	0.906	0.97	0.893	0.959
FULL	0.637	0.815	0.897	0.815

5.4.5 Cross-Database Evaluation

The generalization ability of the neural network-based adaptation of the PSNR is studied in a cross-database evaluation for selected distortions and over full databases. For cross-database evaluation on the full database, no knowledge about the distortion type is assumed, i.e. $\gamma = 1$. The results are presented in terms of SROCC in Table 5.6.
High generalization ability is achieved for the single distortion case. Given the large amount of reference images shared between LIVE and TID2013, this is not surprising. For single distortions the approach also generalizes well for images unseen during training in CSIQ. Cross-database evaluation over full image databases results in low prediction accuracies. As shown in Section 5.4.3, the proposed method does not perform well without consideration of the distortion type; hence, high accuracies can neither be expected for distortion-type agnostic cross-database evaluation.

5.4.6 Weight Estimation on Distorted Images

Table 5.7 – Performance comparison on LIVE and TID2013 databases with models trained on the distorted image instead of the reference image.

	LIVE		TID2013	
	LCC	SROCC	LCC	SROCC
$\text{paPSNR}^{\text{dst}}_{\gamma=1}$	0.971	0.971	0.739	0.741
$\text{paPSNR}^{\text{dst}}_{\gamma^*}$	0.972	0.971	0.898	0.902

Although it does not follow the previously derived concept of distortion sensitivity and gives away the advantage of graceful distribution of complex processing to the reference image only, local weights can in principle also be estimated from the distorted image. The resulting prediction performance is presented in Table 5.7. The results show that adaptation of the PSNR based on the distorted image achieves higher prediction accuracy compared to adaptation based on the reference image both in terms of Pearson linear correlation coefficient (LCC) and SROCC. From the perspective of distortion sensitivity this is very surprising. However, it was shown in Chapter 4 that a neural network can learn to extract quality related information from the distorted image only; such an information is not available from a reference image. Further, the distorted image contains information about the distortion type [Moorthy and Bovik, 2011] that can be exploited by the network to improve prediction accuracy.
It can be hypothesised that a network trained on the distorted image in fact learns a different representation compared to a network trained on the references. The inferior performance obtained by predicting 'distortion sensitivity' from a undistorted image by a network trained on distorted images (LCC: 0.877, SROCC: 0.921) and predicting 'distortion sensitivity' from an distorted image by a network trained on undistorted images (LCC: 0.79, SROCC: 0.807) corroborates this conjecture.

5.5 Perceptually Distortion Sensitive Video Compression

Lossy data compression builds on two principles: Redundancy reduction and irrelevance reduction [Wiegand and Schwarz, 2016]. While redundancy reduction exploits the statistical regularities of a signal for compact representation and is inherently lossless, irrelevance reduction takes advantage of the limited capacity of the information sink and removes information that is irrelevant, e.g. imperceptible, for the receiver in order to reduce the amount of data. Removed information is generally not recoverable, thus irrelevance reduction is a lossy technique. Hence, a reliable and accurate measure of irrelevance is crucial. Although perceptual models are widely used to guide irrelevance reduction in audio and speech compression [Brandenburg et al., 2013], only few perceptual properties are successfully exploited in image or video compression, e.g. by chroma subsampling [Wiegand and Schwarz, 2016].
The previous sections showed how the concept of distortion sensitivity significantly improves the perceptual relevance of a simple quality measure such as the PSNR. This section shows how distortion sensitivity can be directly used to guide the irrelevance reduction and by that the bit allocation in video compression.

5.5.1 Block-Based Hybrid Video Coding

In block-based hybrid video coding algorithms such as H.264|MPEG-4 AVC [Wiegand et al., 2003] or H.265|MPEG-H HEVC [Sullivan et al., 2012], spatial and temporal redundancies are exploited by block-based prediction from spatially or temporally correlated blocks.
With $\mathbf{s}_k$ and $\widehat{\mathbf{s}}_k$ being the original and the predicted signal of an image block[2] B_k the resulting prediction error signal (or residual signal) calculates as

$$\mathbf{u}_k = \mathbf{s}_k - \widehat{\mathbf{s}}_k, \tag{5.12}$$

that is transformed by $T(\cdot)$ to the transform coefficients

$$\mathbf{c}_k = T(\mathbf{u}_k). \tag{5.13}$$

Transform coefficients $\mathbf{c}_k$ are mapped to quantization indices that are entropy coded and sent to the receiver [Wiegand and Schwarz, 2016]. After entropy decoding, quantization indices are mapped to the quantized transform coefficients $\tilde{\mathbf{c}}_k$. Quantization is a lossy process, thus generally $\mathbf{c}_k \neq \tilde{\mathbf{c}}_k$; irrelevance reduction takes place here. Inverse transform gives the reconstructed residual signals

$$\tilde{\mathbf{u}}_k = T^{-1}(\tilde{\mathbf{c}}_k), \tag{5.14}$$

[2] Note the conceptual similarity to the previously used term *patch*. Here, the difference is only that blocks are always non-overlapping and densely sampled. In the following this is indicated by the use of B_k instead of P_i.

leading to the reconstructed signal

$$\tilde{\mathbf{s}}_k = \tilde{\mathbf{u}}_k + \widehat{\mathbf{s}}_k. \tag{5.15}$$

5.5.1.1 Encoder Control by Lagrangian Optimization

Modern codecs such as H.265|MPEG-H HEVC [Sullivan et al., 2012] provide a multitude of tools and corresponding coding parameters, such as block subdivision parameters, reference picture indexes, motion data, quantization parameters, and transform coefficient levels. Based on the available tools and their parameters, the encoder has to determine the set of coding parameters $\mathbf{p}$ to encode a video such that the distortion $D(\mathbf{s}, \tilde{\mathbf{s}}(\mathbf{p}))$ between the original video signal $\mathbf{s}$ and the reconstructed video signal $\tilde{\mathbf{s}}(\mathbf{p})$ is minimized, while the required number of bits $R(\mathbf{p})$ does not exceed a bit budget R_B. The resulting constrained optimization

$$\min_{\mathbf{p}} \; D(\mathbf{p}) \quad \text{subject to} \quad R(\mathbf{p}) \leq R_B. \tag{5.16}$$

can be transformed, based on the concept of Lagrangian multipliers, into an unconstrained optimization problem

$$\min_{\mathbf{p}} \; D(\mathbf{p}) + \lambda \cdot R(\mathbf{p}), \tag{5.17}$$

with $\lambda \geq 0$ being the *Lagrange multiplier*.
Lagrangian optimization can be applied to the allocation of restricted resources among multiple entities [Everett, 1963], such as the allocation of bits among different blocks. The optimal coding decisions $\{\mathbf{p}_0, \mathbf{p}_1, \ldots\} = \{\mathbf{p}_k\}$ over a set of blocks $\{B_0, B_1, \ldots\} = \{B_k\}$ that form a picture are

$$\min_{\{\mathbf{p}_k\}} \; D_{pic}(\{\mathbf{p}_k\}) + \lambda \cdot R_{pic}(\{\mathbf{p}_k\}), \tag{5.18}$$

and with a block-wise additive distortion measure D, simplified to

$$\min_{\{\mathbf{p}_k\}} \; \sum_{\forall k} D_k(\{\mathbf{p}_k\}) + \lambda \cdot \sum_{\forall k} R_k(\{\mathbf{p}_k\}). \tag{5.19}$$

If we ignore the influence of the current coding decision on future decisions, the optimal solution of this minimization can also be found by k separate minimizations [Everett, 1963]

$$\forall k \; \min_{\mathbf{p}_k} \; D_k(\mathbf{p}_k) + \lambda \cdot R_k(\mathbf{p}_k), \tag{5.20}$$

where $R_k(\mathbf{p}_k)$ represents the bitrate required to transmit all coding parameters $\mathbf{p}_k$ for a block B_k and $D_k(\mathbf{p}_k)$ denotes the resulting distortion for the block B_k. The Lagrange multiplier λ is typically constant over all blocks of a frame.

The sum of squared errors (SSE) is a block-wise additive distortion measure $D_k = D_k^{\mathrm{SSE}}$ that,

with the exception of motion estimation, is widely used in video coding. This means that the encoder is basically optimized with respect to the MSE[3].

5.5.1.2 Quantization

Quantization determines the quantization indexes $\mathbf{q}_k$ that approximate the residual signal $\mathbf{u}_k$ in the transform domain. Note that the distortion of the reconstructed signal $\tilde{\mathbf{s}} = \tilde{\mathbf{u}} + \hat{\mathbf{s}}$ is equal to the distortion of the reconstructed residual $\tilde{\mathbf{u}}$. If a l_2-norm based distortion measure such as SSE is used and the inverse transform is unitary, the distortion D_k^{SSE} in a block B_k can conveniently be calculated in the transform domain as

$$D^{\mathrm{SSE}} = \sum_{i,j} (\mathbf{c}_k(i,j) - \tilde{\mathbf{c}}_k(i,j))^2 \tag{5.21}$$

with i, j denoting the positions of the coefficients in the transform block.
In general, the encoder could choose any set of transform coefficient levels $\mathbf{q}_k$ for a block B_k. In the simplest case, in practical video coding standards such as H.264|MPEG-4 AVC [Wiegand et al., 2003] or H.265|MPEG-H HEVC [Sullivan et al., 2012], uniform reconstruction quantizers are used with the inverse quantizer mapping

$$\tilde{\mathbf{c}}_k(i,j) = \Delta \cdot \mathbf{c}_k(i,j), \tag{5.22}$$

where the quantization step size Δ is determined by the QP, with the approximate relationship[4]

$$\Delta \propto 2^{\frac{\mathrm{QP}}{6}}. \tag{5.23}$$

Given a quantization step size Δ, rounding the transform coefficients to the nearest integer according to

$$q(i,j) = \mathrm{sgn}\left(c\left(i,j\right)\right) \left\lfloor \frac{|c(i,j)|}{\Delta} + \frac{1}{2} \right\rfloor \tag{5.24}$$

minimizes the distortion $D(q)$. Note that for conceptual simplicity we disregard rate distortion optimized quantization (RDOQ) [Wiegand and Schwarz, 2016] here.

Although conceptually the operational point of an encoder is determined by λ, and the quantization parameter QP_k of a block B_k can be considered as an element of its coding parameter vector $\mathbf{p}_k$, for speed reasons practical encoders are controlled by setting QP_k to a pre-selected frame-wise value QP [Wiegand and Schwarz, 2016].
Using high-rate approximations [Wiegand and Schwarz, 2016] for the SSE distortion and the operational rate-distortion curve of a block, the relationship between the Lagrange multiplier

[3] Due to the typical process of standardization one could further argue that actually whole compression algorithms are optimized with regard to the MSE

[4] Note that for conceptual simplicity, we disregard any scaling here.

λ and the associated quantization step size Δ can be derived as

$$\lambda \propto \Delta^2. \tag{5.25}$$

Plugging Eq. 5.23 into Eq. 5.25 yields the relation between QP and λ

$$\mathrm{QP} - 3\log_2 \lambda = \text{const.} \tag{5.26}$$

5.5.2 Distortion Sensitive Lagrangian Bit Allocation

The concept of distortion sensitivity allows for direct perceptual adaptation of the Lagrangian bit allocation scheme.

For a simplified notation we define a local weight $w_k = 10^{\frac{\delta_k}{10}}$ and write the perceptually adapted SSE of a full picture as

$$\begin{aligned} \mathrm{paSSE}_{pic} &= \sum_k \mathrm{paSSE}_k && (5.27)\\ &= \sum_k w_k \cdot \mathrm{SSE}_k. && (5.28) \end{aligned}$$

For the conventional SSE distortion measure $D_k^{\mathrm{SSE}} = \mathrm{SSE}_k$, the coding decision for each block is

$$\min_{\mathbf{p}_k} \; D_k^{\mathrm{SSE}}(\mathbf{p}_k) + \lambda_{pic} \cdot R_k(\mathbf{p}_k). \tag{5.29}$$

If the perceptually adapted distortion measure paSSE is used, the minimization changes to

$$\min_{\mathbf{p}_k} \; w_k D_k^{\mathrm{SSE}}(\mathbf{p}_k) + \lambda_{pic} \cdot R_k(\mathbf{p}_k), \tag{5.30}$$

which is equivalent to

$$\min_{\mathbf{p}_k} \; D_k^{\mathrm{SSE}}(\mathbf{p}_k) + \lambda_k \cdot R_k(\mathbf{p}_k), \quad \text{with} \quad \lambda_k = \frac{\lambda_{pic}}{w_k}. \tag{5.31}$$

Consequently, for each block B_k, we have the same optimization problem as with the conventional SSE distortion, but only the Lagrange multiplier λ_k is perceptually modified on a block basis in dependence of its distortion sensitivity.
Conceptually, this perceptual adaption technique directly follows from the functional notion of distortion sensitivity introduced in the beginning of this chapter. Practically, this approach is particularly attractive for video encoder control as it preserves the advantages of the SSE distortion, such as direct calculation in the transform domain, straightforward quantization and low complexity, as the increase of complexity inherent to the calculation of w_k is outside the mode decision loop.

5.5.2.1 Distortion Sensitive Adapted QP-Selection

Distortion sensitivity can be directly incorporated into the approximate relation between QP and λ. The picture-wise relation between QP and λ is given by Eq. 5.26 and can be refined for a block B_k as

$$\begin{aligned} \mathrm{QP}_k - 3\log_2 \lambda_k &= \text{const.} \\ \mathrm{QP}_k - 3\log_2 \lambda w_k &= \text{const.} \\ \mathrm{QP}_k + 3\log_2 w_k - 3\log_2 \lambda &= \text{const.} \end{aligned} \tag{5.32}$$

where the comparison of Eq. 5.26 and Eq. 5.32 yields the perceptual QP assignment

$$\mathrm{QP}_k = \mathrm{QP}_{pic} - 3\log_2 w_k \tag{5.33}$$

5.5.3 Experiments

5.5.3.1 Perceptual Image Compression

(a) BQTerrace (b) Cactus (c) Kimono

(d) MarketPlace (e) ParkScene (f) Wisley

Figure 5.8 – Originals of images used in test.

Distortion sensitive bit allocation is evaluated experimentally for still image coding using High Efficiency Video Coding (HEVC) using *intra only* settings as defined in the JCT-VC common test conditions [Bossen, 2013]. Bit allocation is controlled by CTU-wise adaptation of the Lagrangian multiplier λ and the QP according to Eq. 5.31 and Eq. 5.33.
Local weights w_k are estimated based on the reference image by a neural network trained for a block size of 64×64 pixels (cf. Section 5.3). Given the lack of quality annotated HEVC compressed images or videos, the model was trained on the JPEG-subset of the LIVE database

Table 5.8 – MOS gains and bit rate savings for different bit allocation schemes.

	NN-based QPA			σ-based QPA		
	ΔQ	ΔR[%]	CI[%]	ΔQ	ΔR[%]	CI[%]
BQTerrace	0.42 [-0.17; 1.01]	-28.0 [19.9; -50.4]	74.6	0.31 [-0.31; 0.91]	-21.2 [19.9; -47.3]	73.3
Cactus	-0.13 [-0.73; 0.47]	7.4 [24.4; -19.2]	92.8	0.03 [-0.53; 0.56]	1.2 [24.4; -22.4]	93.1
Kimono	0.16 [-0.45; 0.77]	-9.1 [35.8; -31.8]	100.0	-0.03 [-0.67; 0.61]	1.3 [35.8; -25.9]	100.0
MarketPlace	0.13 [-0.44; 0.70]	-9.0 [37.1; -36.5]	93.8	0.03 [-0.52; 0.58]	-2.3 [37.1; -32.6]	92.8
ParkScene	0.55 [-0.08; 1.18]	-33.9 [39.9; -55.6]	81.2	0.22 [-0.40; 0.83]	-14.9 [39.9; -44.3]	82.0
Wisley	0.34 [-0.23; 0.95]	-22.4 [12.8; -46.0]	94.9	0.20 [-0.35; 0.77]	-15.5 [12.8; -39.0]	96.4
Overall	0.24 [-0.35; 0.85]	-15.9 [28.3; -39.9]	89.6	0.12 [-0.47; 0.71]	-8.6 [28.3; -35.2]	89.6

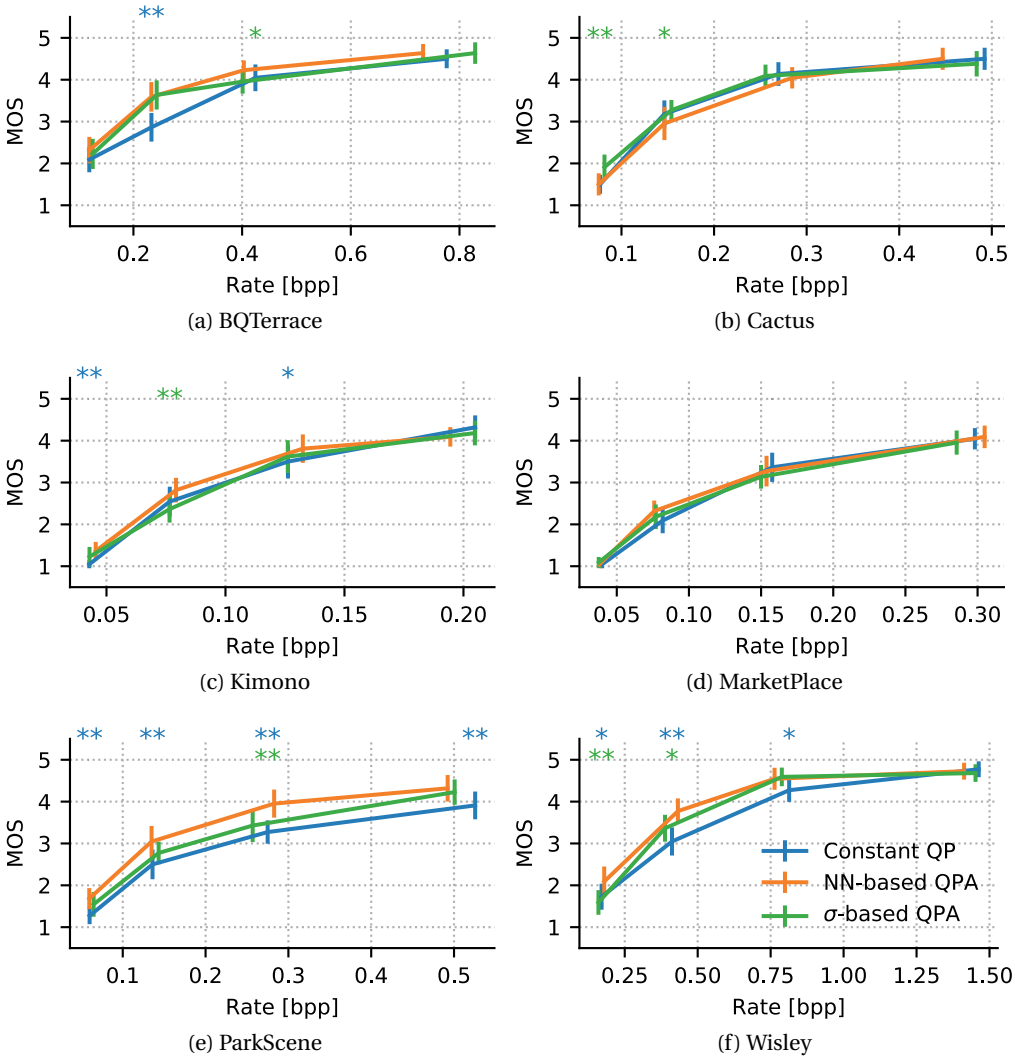
[Sheikh et al., 2006].
The compression performance of neural network-based estimation of distortion sensitivity is compared to a standard deviation-based approach presented in [Bosse et al., 2017b]. Both approaches are referenced to a conventional constant QP bit allocation scheme.
Evaluation is based on four rate points and 6 reference images, shown in Fig. 5.8. QPs for adaptive approaches are selected reference image-specific in order to achieve bit rates closest to the ones obtained by conventional compression at QP $\in [27, 32, 37, 42]$.

5.5.3.2 Quality Assessment

Quality is assessed in a psychometric test, employing a DCR procedure as described in Section 2.2 using a 5-grade rating scale. Stimuli were presented on a Sony KD-65XE9305 UHD television set side-by-side in original resolution with the reference shown on the left-hand side and the distorted image shown on the right-hand side. The non-active area of the display was set to black. The 48 pixel wide area between the images was set to 50%-mid gray. In accordance with the recommendations [ITU-T Rec. P.910, 2008], observers were seated at a distance of $3 \cdot H$ (H being the active screen height) in front of the screen. Up to three participants, seated side-by-side, assessed the images simultaneously in one session. Pairs of reference and processed images were presented in random order for 10 s. Two hidden references were injected randomly. After each presentation, observers were asked to note down the quality rating on a piece of paper. The first 3 image pairs presented were chosen to represent the quality range in order to make the observers familiar with the task and the quality range. The ratings thereof are not considered for further analysis. 22 subjects (4 female, 18 male, average age: 31.1 ± 4.43 years) participated in the experiment. Most of the participants reported to have experience in the field of video compression.

5.5.4 Results

Responses given to the presentation of the hidden references were used for a simple screening; ratings of one subject were discarded for further analysis as the average rating given to the hidden references was below 4. MOSs were calculated from the data of the remaining 21 subjects according to Eq. 2.1.
The quality of the compressed images is shown over the bit rate in Fig. 5.9 for the three different



(a) BQTerrace (b) Cactus

(c) Kimono (d) MarketPlace

(e) ParkScene (f) Wisley

Figure 5.9 – Compression performance for different bit allocation schemes. Vertical bars indicate the 95% confidence intervall. Asterisks denote statistical significant differences (*: $0.05 \leq p < 0.1$, **: $p < 0.05$) between NN-based QPA and constant QP (blue asterisks) and between NN-based QPA and σ-based QPA.

bit allocation schemes *a)* constant QP, *b)* neural network-based QP adaptation, and *c)* standard deviation-based QP adaptation. Vertical bars indicate the 95%-confidence interval for the MOS. Statistical significance (tested using a pair-wise Student's t-test [Howell, 2013]) of the differences in subjective ratings between the proposed neural network-based QP adaptation scheme and constant QP are denoted by blue asterisks, statistical significance between neural network-based QP adaptation and standard deviation-based QP adaptation by green asterisks respectively; * indicates a significance level of $0.05 \leq p < 0.01$ and ** indicates a significance level of $p < 0.05$. Note that, as it is difficult to encode images at identical bit rates, the rate points at which statistical significance is tested are are not exactly identical.

For all images, except *Cactus*, both QP adaptions achieve superior or at least comparable quality scores over the full bit rate range, with the neural network-based adaptation performing mostly better compared to the standard deviation-based. For *Cactus* the standard deviation-based approach performs significantly better than the neural network-based approach for the two lower rate points. For most images, significant quality gains are achieved by the QP adaptation schemes only at bit rates below the saturation of quality. This can be expected as lossy image compression with a bit budget that is already large enough for high visual quality (high rate points) cannot benefit significantly from an improved bit allocation scheme. Although typically not relevant for applications, it would be interesting to evaluate statistical significance in bit rate regimes closer to the saturation at the lower end of the quality scale.
In Table 5.8 the compression performance with regard to the MOS is summarized in a psychophysical adaptation of the Bjøntjegaard model [Hanhart and Ebrahimi, 2014]. As, in contrast to the PSNR, the MOS is a statistical measurement, not only the mean estimates of the gains in MOS (ΔQ) and bit rate reductions (ΔR) of the two adaptation schemes over constant QP assignment are reported, but also the minimal and maximal estimates based on the 95%-confidence interval. The confidence interval (CI) attempts to quantify if the quality range is sufficiently covered for comparison [Hanhart and Ebrahimi, 2014]. Again, with exception of *Cactus*, the proposed neural network-based QP adaptation scheme shows to be clearly superior over the standard deviation-based scheme. The reason for the inferior performance of the proposed method (as well as of the standard deviation-based method) for *Cactus* is probably the presence of letters and numbers in the image; these these high-level semantics are not captured by the distortion sensitivity models. However, the averaged over all test images (Overall), the performance of the proposed method is clearly superior to the standard deviation-based QP adaption scheme.

5.6 Discussion

Based on an analysis of the non-linear mapping of computational quality values Q_c to subjective quality scores Q_p and a discussion of the parameters of this mapping function this chapter derived a concept of distortion sensitivity and showed that the shift parameter in the psychometric mapping function can serve efficiently as functional definition thereof. Distortion sensitivity was modelled as a distortion type-dependent property of an image, and it was shown that compensating for it can efficiently improve the prediction performance of a given computational quality model. Limits of such approaches were explored quantitatively. A neural network-based method for patch-wise estimation of distortion sensitivity within an image quality estimation framework was presented that significantly improves the quality estimation accuracy of the base quality model, i.e. the PSNR.
The derivation of a distortion sensitive PSNR led to a local weighting scheme for a perceptual adaptation of the MSE. This weighting scheme was incorporated into the bit allocation in hybrid block-based video compression. The perceptually superior performance of the resulting bit allocation scheme was experimentally validated for image compression in an

psychophysical quality assessment study. Hence, this chapter bridged from psychometric properties to bit allocation for perceptual image compression.

The presented definition of distortion sensitivity and the proposed framework for estimation thereof can be easily adapted to other quality models than the PSNR and extended to other signal modalities such as videos, assuming the availability of quality annotated data.
The neural network-based patch-wise compensation for distortion sensitivity significantly improves the performance of the PSNR. However, comparing the achieved performance with the limits determined by (hypothetical) optimal image-wise compensation shows that the method still has further potential for improvement. The sub-optimality indicates that there is some room for improving the generalization ability of the model with regard to unseen images.
Weights used for spatial pooling are commonly normalized (cf. Section 3.1.3)). Different to the weighted average patch aggregation in Chapter 4, the weighting scheme derived from distortion sensitivity does not comprise any normalization. This also explains the independence of the SROCC from patch-size as non-normalized weights are capable of capturing a global image property (cf. Section 5.2.2). Imagine one image of high and spatially uniform distortion sensitivity and another image of low and spatially uniform distortion sensitivity. While a non-normalized weighting scheme could differentiate between high and low sensitivity, this information would be lost by normalization of the weights. However, in future work, differences between normalized and non-normalized weighting can be studied within the presented framework. This potentially also brings better understanding on how humans spatially pool perceptual visual quality.
The proposed method works better if local weights are estimated from the distorted images rather than from the reference images. This does not follow the concept of distortion sensitivity that was presented as a property of the reference image and, thus, appears surprising. It is however not unexpected, since neural networks, as shown in Chapter 4, are able to predict quality relatively accurately from the distorted image alone as well. More insight into the nature of distortion sensitivity and relevant features driving distortion perception might be gained by investigating differences in the internal representations in networks trained based on the original and distorted images using explaining methods [Bach et al., 2015, Montavon et al., 2017, Montavon et al., 2018]. Also note that the reference image-based models were trained on a smaller sample size regarding the input signal compared to the distortion image-based models, while the number of quality labels is identical. At this point it is not clear how this imbalance impacts the training. However, although achieving higher prediction accuracy, estimating quality based on weights extracted from the distorted image forfeits the crucial advantage of performing complex computations on the reference image only.
The discussion of the limits of the proposed shows that, as concluded previously for the approach presented in Chapter 4, the availability of quality annotated images and videos is crucial for the success of data-driven approaches to quality assessment. This is especially important for an application such as distortion sensitive bit allocation as most databases do not consider modern compression algorithms such as HEVC as a distortion type and typically

only contain images and videos of resolutions that are practically not of highest relevance any more. Given the flaws of conventional approaches (cf. Chapter 2) the next chapter will investigate a novel approach to psychophysiological image quality assessment. However, the main advantage and technical motivation of the proposed distortion sensitive quality assessment is not a remarkable high accuracy, but the allocation of computational complex processing to the reference image only. Hence, until larger database are available, the method could be trained on images annotated by quality models that are computationally less graceful, but more accurate.
Incorporating the proposed approach into a distortion sensitive bit allocations scheme obtained superior coding gains compared to an explicit model-based state-of-the-art method. This is especially promising, as the data-driven approach was trained for JPEG distortions on images of a resolution in the order of about 600×600 pixels, but tested in a HEVC context on HD content. Even better results might be achievable with models trained on more appropriate data. Future work should evaluate if the proposed approach works as efficiently on videos as well. For this, distortion sensitive bit allocation could also directly employ models trained on video data.
Distortion sensitive adaptation of the block-wise QP was controlled on a CTU-level (64×64). HEVC allows for QP adaptation on smaller block sizes as well; this might offer potential for higher perceptual compression gains.

5.7 Lessons Learned

- Distortion sensitivity can be modelled as a distortion type-dependent property of an image.
- The shift parameter of the psychometric function that maps the output of computational quality models into the perceptual domain represents an efficient functional definition of distortion sensitivity.
- Compensating a given quality model for distortion sensitivity increases its quality estimation accuracy.
- Local distortion sensitivity can be estimated within a quality estimation framework by a neural network.
- The concept of distortion sensitivity can be directly incorporated into a perceptually improved bit allocation scheme for image and video compression.
- The performance of data-driven estimation of distortion sensitivity and its application for image and video compression is limited by the scarcity of training data.

6 Image Quality Assessment Using Steady-State Visual Evoked Potentials

This chapter is based on

Bosse, S., Acqualagna, L., Samek, W., Porbadnigk, A. K., Curio, G., Blankertz, B., Müller, K.-R., and Wiegand, T. (2017a). Assessing perceived image quality using steady-state visual evoked potentials and spatio-spectral decomposition. *IEEE Transactions on Circuits and Systems for Video Technology*, 8215(c):1–1

6.1 Introduction

Chapter 4 and Chapter 5 investigated two different approaches to end-to-end trained data-driven image quality estimation; results showed that in order to train models that achieve satisfactory generalization ability, databases larger than the currently available are necessary. For obtaining more efficient models for bit allocation as described in Section 5.5 quality annotated images and videos of more relevant resolutions and distortion types are beneficial. Therefore, this chapter shifts the perspective on visual quality from its estimation to its assessment. Driven by the insights from the discussion on the limitations of psychophysical quality assessment in Section 2.2, a novel neurophysiological approach to image quality assessment based on SSVEPs (cf. Section 2.3.2.1) is investigated.

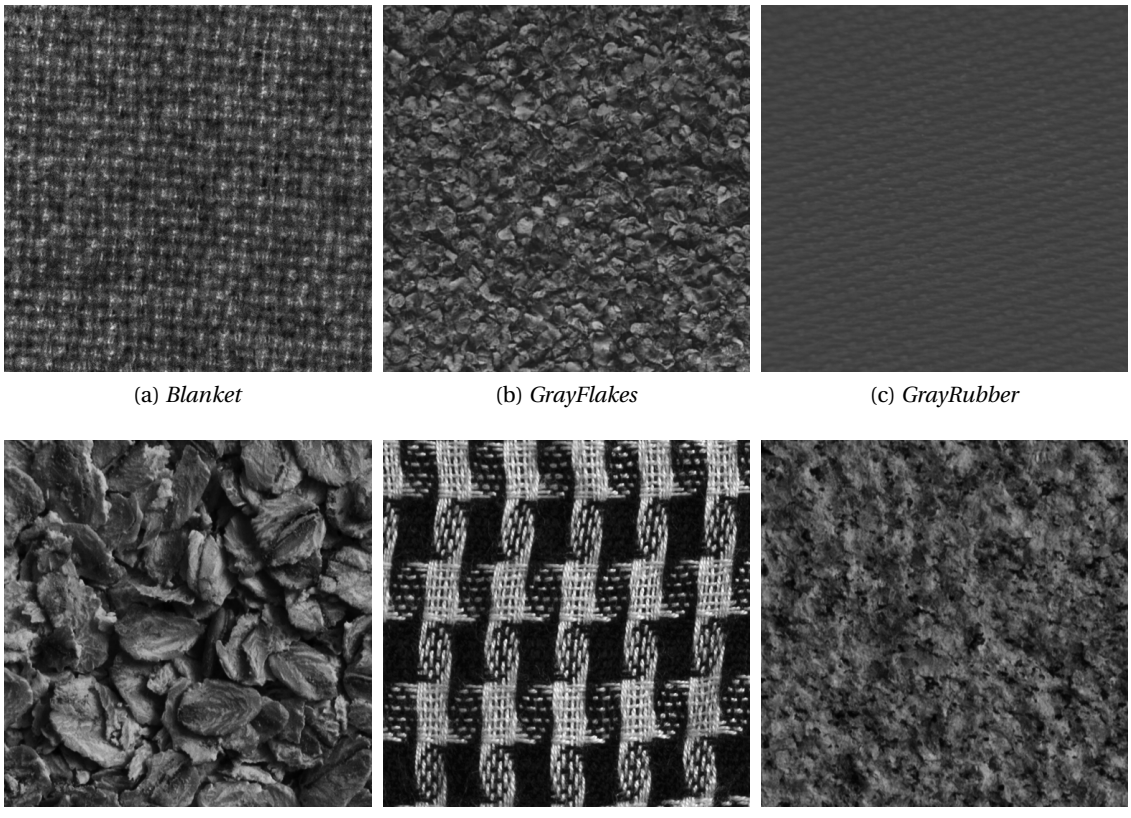
The data corpus this chapter works with was used earlier to study the neurally informed detection of quality degradations in images using EEG [Acqualagna et al., 2015]. In this work, the neural signal was classified according to the presentation of distorted or undistorted images, disregarding any information about the distortion magnitude and, as such, neglecting any quantification of image quality, e.g. represented as MOS. For spatial dimensionality reduction, the common spatial pattern (CSP) technique [Blankertz et al., 2008] was employed and it was shown that classification accuracy depends on distortion magnitude. While the study provides strong evidence for the general feasibility of SSVEP-based image quality assessment, it does not allow for the assessment of perceived quality because different image quality levels were not preserved in the analysis. Moreover, CSP is a supervised dimensionality reduction

method that requires labeled training data as it maximizes the variance ratio between two known classes. In quality assessment, quality ratings are typically not given a-priori, but rather the outcome of the experiment and can thus not be used as labels. This makes unsupervised preferable over supervised methods in quality assessment studies.
This chapter addresses the limitations of the previous study by evaluating the correlation between brain responses and MOS values using SSD [Nikulin et al., 2011] for dimensionality reduction, extending previous work on SSVEP-based image quality assessment (IQA) [Bosse et al., 2014, Bosse et al., 2015]. SSD is an unsupervised method that aims at finding a spatial filter that maximizes SNR assuming disjoint spectral distributions of signal and noise. The SNR properties of the SSVEP (cf. Section 2.3.2.1) makes SSD a natural choice for channel decomposition and dimensionality reduction. Thus, in this chapter SSD is re-formulated in the frequency domain, rendering it particularly suited for the SSVEP framework. In contrast to most previous work on psychophysiological quality assessment (cf. Section 2.2), SSD provides a tool for rational channel selection that does not rely on neuroanatomical knowledge nor labeled data. This advantage means that neural components extracted from the EEG-signal will reflect the sensory processing underlying image quality perception so that significant correlations between these extracted neural features and MOS can be achieved. A simple screening method, based on the angular distribution of the extracted SSD components, is proposed for rejecting those subjects for which SSD fails. Finally, the feasibility of the proposed approach is shown by predicting the MOS from the neural responses on subject-level, leading to prediction accuracies comparable to psychophysical methods.
The chapter starts with a description of the experimental setup. Section 6.3 details the signal processing methods for data analysis and re-formulates SSD in the frequency domain. Results are presented in Section 6.4. Section 6.5 concludes the chapter with a discussion.

6.2 Experimental Setup

6.2.1 Stimuli

In order to reduce the influence of visual saliency and to make the measurement less dependent on nuisance factors such as local variances in image statistics and the current gaze position, six spatio-statistically roughly stationary gray-level texture images, shown in Fig. 6.1, were chosen from two texture image databases [Ojala et al., 2002, Kylberg, 2011] as the basis for stimulus generation. The images have a size of 512×512 pixel and were shifted to equal values of mean luminance.
Visual quality of the base images was degraded to six different quality levels. Distortions were introduced as compression artifacts by coding the original images using the HM10.0 test model [JCT-VC, 2014] of the High Efficiency Video Coding standard (HEVC) [Sullivan et al., 2012] using *intra only* settings as defined in the JCTVC common test conditions [Bossen, 2013]. HEVC offers a flexible quad-tree structure for prediction and transform. Statistical redundancies are exploited by block-wise temporal (for video signals) and spatial linear prediction. The residual signal is transformed block-wise, and coefficients are quantized in the transform



(a) *Blanket* (b) *GrayFlakes* (c) *GrayRubber*

(d) *Oatmeal* (e) *Scarf* (f) *Stone*

Figure 6.1 – Texture images used in experiment from Kylberg Texture dataset [Kylberg, 2011] and Outex dataset [Ojala et al., 2002]. ©2017 IEEE

(a) QP=14, MOS=8.5 (b) QP=43, MOS=3.8 (c) QP=51, MOS=1.5

Figure 6.2 – Perceptual quality of distorted images in experiment exemplified for texture *Oatmeal*. ©2017 IEEE

domain. Coding artifacts, which are perceived by the human observer as a loss of visual quality, are introduced by the quantization of the transform coefficients [Wiegand and Schwarz,

2016]. Quantization step-size is controlled by the QP [Sullivan et al., 2012]. Distortion levels, mediated by the QP, have been estimated in a pilot study in order to correspond to roughly similar perceptual qualities with two conditions per texture above the assumed perception threshold at MOS ≈ 8, one condition per texture close to the perception threshold and three conditions per texture distributed below the perception threshold. Fig. 6.2 gives an impression of the resulting quality of the stimuli, exemplified for the image *GrayFlakes*.

6.2.2 Participants

Sixteen participants (seven females and nine males, in the age group 21-46) took part in the experiment. All had normal or corrected-to-normal vision and none of them had a history of neurological diseases. They were all native German speakers or at least with a level of German comprehension of five, on the six level scale of competence laid down by the Common European Framework of reference for Languages [Little, 2007]. All of them were naïve in respect of visual quality assessment studies and were paid for their participation. Each subject was briefed individually about the purpose of the experiment. The study was performed in accordance with the declaration of Helsinki [WMA General Assembly, 2013] and all participants gave written informed consent.

6.2.3 EEG Data Acquisition

EEG was recorded with a sampling frequency of 1000 Hz using BrainAmp amplifiers and an ActiCap active electrode system with 64 sensors (both by Brain Products, Munich, Germany). Electrodes were positioned at Fp1,2, AF3,4,7,8, Fz, F1-10, FCz, FC1-6, FT7,8, Cz, C1-6, T7, CPz, CP1-6, TP7,8, Pz, P1-10, POz, PO3,4,7,8, Oz, O1,2. The electrode that in the standard EEG montage, see Fig. 2.4, is placed at T8 was replaced under the right eye and used to measure eye movements. All electrodes were referenced to the left mastoid, using a forehead ground. For offline analyses, electrodes were re-referenced to linked mastoids. All impedances were kept below 10 kΩ.

6.2.4 Stimulus Presentation

As outlined in Section 2.2 and Section 2.3, conventional psychophysical quality assessment and neurophysiological assessment each use highly distinct stimulus presentation regimes. However, the general setup for stimulus presentation was identical for both types of assessment in the experiment: Stimuli were shown on a 23" screen (Dell U2311H) with a native resolution of 1920 × 1080 pixels at a refresh rate of 60 Hz. The screen was normalized according to the recommended values [ITU-R Rec. BT.500-13, 2012]. Stimuli were presented without scaling or interpolation at native resolution. The inactive part of the screen was set to mid grey. Viewing distance was set to 110 cm in compliance with the recommendations [ITU-T Rec. P.910, 2008], with a stimulus resolution of 512 × 512 pixels (128 × 128 mm) leading each

image to span a visual angle of approximately 6.6°. Subjects sat in front of the display in a dimly lit room. Between the two parts of the experiment, subjects had a short rest and were provided small snacks and drinks.

6.2.4.1 Behavioral Part

In the psychophysical assessment part of the experiment, participants evaluated the perceived quality of the images following a Degradation Category Rating (DCR) procedure with Simultaneous Presentation (SP) [ITU-R Rec. BT.500-13, 2012] (cf. Section 2.2). Each image pair was presented for 10 s. Quality ratings were given by the participants following stimulus presentation using a mouse-controlled slider on a nine-grade degradation scale (cf. Section 2.2). In this scale, grade 8 is considered as to approximate psychophysical perception threshold of the impairment [ITU-T Rec. P.910, 2008]. Each stimulus pair was presented 3 times in order to obtain reliable measurements. To reduce learning effects, the psychophysical assessment started with the presentation of 12 training stimuli that were not included in the further analysis.

6.2.4.2 Neurophysiological Part

In the neurophysiological assessment part of the experiment stimuli were presented in trials consisting of 6 consecutive texture groups. Each texture group was based on one of the texture images shown in Fig. 6.1, and each texture group started with the presentation of an undistorted texture for 8/3 s, which served as a base image. After presentation of the undistorted base image, distorted and undistorted images were presented periodically alternating at a constant rate of 3 Hz, corresponding to a stimulation frequency (equivalent to full cycles per second) of $f_{stim} = 1.5$ Hz, in distortion level-wise blocks. 4 alternations from distorted to undistorted image were presented per distortion level. With 6 distortion levels (cf. Section 6.2.1) this results in a total presentation duration of 18 2/3 s per texture block. Six texture groups were presented in each trial, with six distortion levels for each group. In order to avoid memory or adaptation effects, both the order of texture groups for each base image and the order of distortion levels presented for each texture group was permuted randomly. An example of the temporal structure of one trial is visualized in Fig. 6.3.
A fixation cross was presented in the beginning of each trial. EEG data was recorded in 3 blocks, consisting of 20, 15, and 16, respectively, trials. This amounts to 51 trials per condition. The 3 blocks were divided by short breaks of about 10 min during which participants were provided drinks and small snacks.

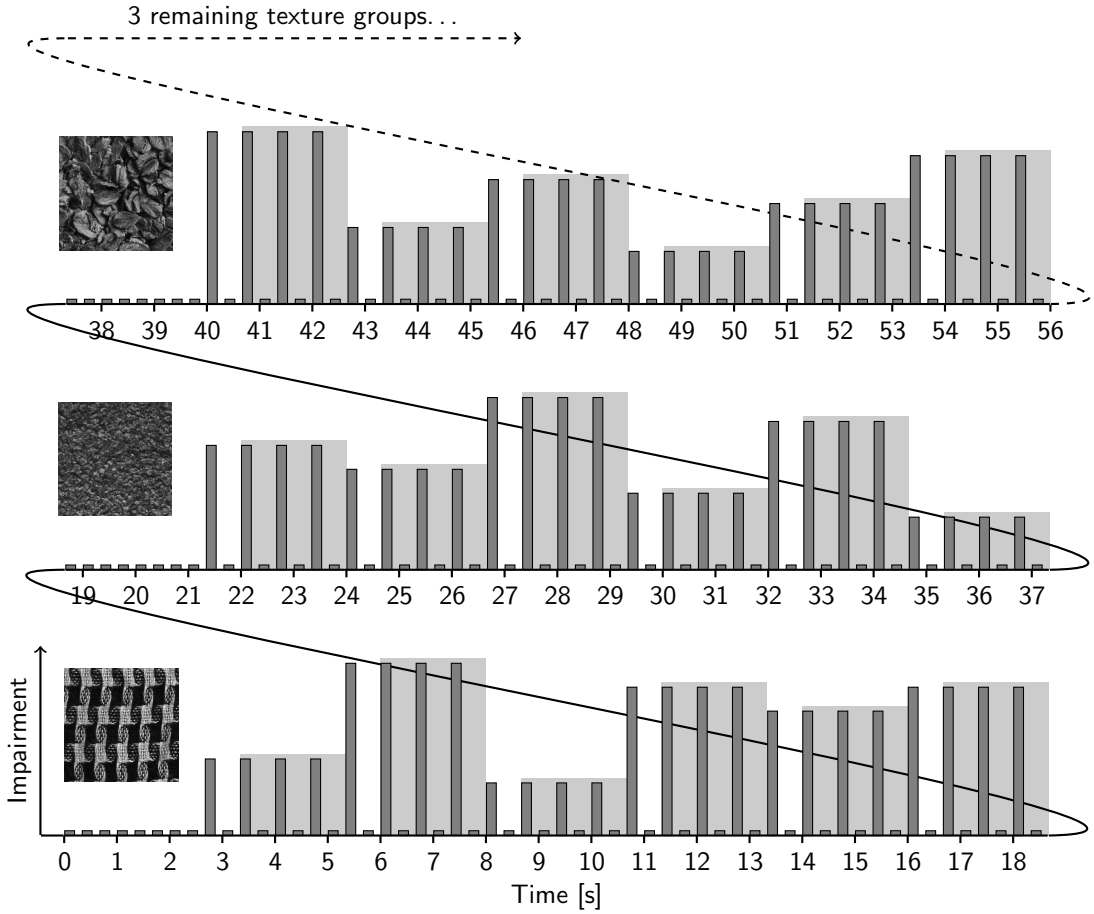


Figure 6.3 – Temporal structure of one trial of stimulus presentation during EEG recording exemplified for 3 base image groups. Stimulus presentation is grouped by base image. Each base image group is introduced with the presentation of the corresponding undistorted image for 8/3 s. Distorted and undistorted images are presented periodically alternating at a constant rate of 3 Hz (amounting to a stimulation frequency $f_{stim} = 1.5$ Hz). For each impairment 4 cycles were presented. The presentation order of the 6 impairment levels is randomized for each base image group. Also the presentation order with respect to the base image is randomized. For further processing, recorded EEG-data is epoched condition-wise, neglecting the first alternation cycle as indicated by the light gray shadings.

6.3 Methods and Data Analysis

6.3.1 Analysis of Psychophysical Data

In psychophysical tests, some observers might give inconsistent responses that can distort the result of the test, cf. Section 2.2.2. Those observers are identified by screening and excluded from further analysis following the recommendations [ITU-R Rec. BT.500-13, 2012]. MOS are obtained by averaging condition-wise over the ratings reported by individual observers

according to Eq. 2.1.

6.3.2 Preprocessing of EEG Data

EEG signal is bandpass filtered from 0.5 Hz to 40 Hz using a zero-phase Chebyshev filter of order 8 (3 dB of ripple in the passband and 40 dB of attenuation in the stopbands) and downsampled to 90 Hz. After eye-movement regression, the EEG is re-referenced to the common average of all electrodes.

6.3.2.1 Eye-Movement Regression

Let $x_k(t)$ denote the EEG signal recorded at sensor k and $\mathbf{x}(t) \in \mathbb{R}^K$ the recorded signal of K sensors at a time point t. Horizontal eye movements are captured by the difference signal of the sensors F9 and F10 $x_{hor}(t) = x_{F9}(t) - x_{F10}(t)$, vertical eye movements and blinks by the difference between signals measured at electrodes Fp2 and EOG $x_{ver}(t) = x_{Fp2}(t) - x_{EOG}(t)$. By combining $x_{hor}(t)$ and $x_{ver}(t)$ to $\mathbf{x}_{eye}(t) = [x_{hor}(t), x_{ver}(t)]^T$ we define Σ_{eye} as the covariance matrix of $\mathbf{x}_{eye}(t)$, Σ_x as the covariance matrix of $\mathbf{x}(t)$ and $\Sigma_{x,eye}$ as the cross-covariance of $\mathbf{x}_{eye}(t)$ and $\mathbf{x}(t)$. This leads to a backward model relating the sensor activity to the underlying originating sources $\mathbf{W} = \Sigma_x^{-1}\Sigma_{x,eye}$ [Haufe et al., 2014b]. The forward model, relating the source activity to the observed sensor activities is then given as $\mathbf{A} = \Sigma_x \mathbf{W} \Sigma_{eye}^{-1}$ [Haufe et al., 2014b], where Σ_x and Σ_x^{-1} (from $\mathbf{W}$) cancel out. Interferences of eye motion can now be regressed out from the recorded signal as $\tilde{\mathbf{x}}(t) = \mathbf{x}(t) - \mathbf{A}\mathbf{A}^{\#}\mathbf{x}(t)$ [Parra et al., 2005, Müller et al., 2003], where $^{\#}$ denotes the pseudo-inverse of a matrix. For further processing and analysis, data recorded at EOG is neglected.
For the sake of readability, although eye artifacts are regressed out, in the following recorded data $\mathbf{x}(t)$ is assumed to be free of eye movement artifacts. Therefore, $\tilde{\mathbf{x}}(t)$ is denoted as $\mathbf{x}(t)$.

6.3.2.2 Epoching

EEG data recorded for each subject is subdivided into epochs of 2 s length from $\frac{2}{3}$ s to $2\frac{2}{3}$ s relative to the beginning of a texture/distortion level-wise block as denoted by the shadings in Fig. 6.3. Hence, the epoched EEG data relates to the presentation of 3 cycles of alternations from distorted to undistorted images. Neglecting the first cycle reduces potential influence of transient components to the stimulus onset.

6.3.2.3 Artifact Rejection

EEG epochs that contained a large percentage (more than 20%) of data samples exceeding a threshold of 25 µV are excluded as artifacts. Typically, these epochs are associated with strong eye movements, blinks or other body movement that could not be regressed out.

6.3.3 Feature Extraction

Fourier transform is applied epoch-wise to the recorded EEG data. As the sampling frequency ($f_s = 90\,\text{Hz}$) is an integer multiple of the stimulation frequency ($f_{stim} = 1.5\,\text{Hz}$) and the epoch length of 2 s allows for an integer number of stimulation periods per epoch (in this case 3 periods), no windowing is applied to the data in order to avoid side-bands [Bach and Meigen, 1999].

6.3.4 Dimensionality Reduction

A common method for analysing EEG data and dimensionality reduction is to find a spatial filter $\mathbf{W}$ that projects the sensor-wise measurement $\mathbf{x}(t)$ to a new subspace containing the spatial components $\mathbf{y}(t) = \mathbf{W}^\top \mathbf{x}(t)$ [Haufe et al., 2014b, Haufe et al., 2014a]. $\mathbf{W}$ is found by optimizing $\mathbf{y}(t)$, given $\mathbf{x}(t)$, with regard to a specific criterion. The column $\mathbf{w}_i \in \mathbb{R}^K$ of $\mathbf{W} \in \mathbb{R}^{K\times K}$ contains the filter that project $\mathbf{x}(t)$ onto the components i of the subspace. Thus, the time course of the ith spatial component is calculated as $y_i(t) = \mathbf{w}_i^\top \mathbf{x}(t)$. Accordingly, with $\mathbf{A} = \left(\mathbf{W}^{-1}\right)^\top$, the column $\mathbf{a}_i \in \mathbb{R}^K$ of $\mathbf{A} \in \mathbb{R}^{K\times K}$ contains the spatial activity patterns[1] of the respective component i [Haufe et al., 2014b].

If the signal and noise components $\mathbf{x}_s(t)$ and $\mathbf{x}_n(t)$ of an EEG recording are assumed to be spectrally disjoint, the SNR can be used as an optimization criterion for finding $\mathbf{W}$ [Nikulin et al., 2011]. For this, SSD extracts components of neural oscillations by maximizing the power in one frequency band and, simultaneously, minimizing the power in another frequency band. For a sensor i the SNR is defined as

$$\text{SNR}(x_i) = \frac{P_{s,i}}{P_{n,i}}, \tag{6.1}$$

where $P_{s,i}$ and $P_{n,i}$ are the power contained in $x_{s,i}(t)$ and $x_{n,i}(t)$, respectively. Signal and noise components $\mathbf{x}_s(t)$ and $\mathbf{x}_n(t)$ are obtained by bandpass filtering the recorded signal $\mathbf{x}(t)$ according to the assumed disjoint frequency bands of signal and noise.
With a spatial filter $\mathbf{W}$ and the projection $\mathbf{y}(t) = \mathbf{W}^\top \mathbf{x}(t)$, the SNR of the component $y_i(t)$ can be found as

$$\begin{aligned} \text{SNR}(y_i) =& \text{SNR}(\mathbf{w}_i^\top \mathbf{x}_i) && (6.2)\\ =& \frac{\mathbf{w}_i^\top \Sigma_s \mathbf{w}_i}{\mathbf{w}_i^\top \Sigma_n \mathbf{w}_i} && (6.3) \end{aligned}$$

with Σ_s and Σ_n being the covariance matrices of the bandpass filtered signals $\mathbf{x}_s(t)$ and $\mathbf{x}_n(t)$. Considering all components, maximizing the SNR for $\mathbf{W}$ leads to the generalized eigenvalue

[1] Note that $\mathbf{A} = \left(\mathbf{W}^{-1}\right)^\top$ only holds in the square case. Generally, $\mathbf{A} = \Sigma_\mathbf{x}\mathbf{W}\Sigma_\mathbf{y}$ [Haufe et al., 2014b].

problem

$$\Sigma_s \mathbf{W} = \mathbf{D}\Sigma_n \mathbf{W}, \tag{6.4}$$

where the entries of **D** contain the generalized eigenvalues and can be interpreted as the amount of SNR projected to a specific component.

The narrowband property of SSVEP (cf. Section 2.3.2.1) can be directly utilized for SNR optimization and a reformulation of SSD allows for convenient application in the frequency domain. As the signal of SSVEP is confined to the frequency bins centered at the harmonic frequencies $n \cdot f_{stim}$ of the stimulation frequency f_{stim}, its noise is typically estimated as a linear approximation from the content in the spectrally neighbored frequency bins centered at $n \cdot f_{stim} \pm \Delta f$, with Δf denoting the frequency resolution. This leads to a sensor-wise estimation of the SNR per frequency bin centered at frequency f as

$$\mathrm{SNR}(f) = \frac{P_s(f)}{P_n(f)} \approx \frac{P(f)}{0.5\big(P(f-\Delta f) + P(f+\Delta f)\big)}, \tag{6.5}$$

where $P(f)$ denotes the power of the spectral component at f and $P(f \pm \Delta f)$ denotes the power of the spectrally neighbored components.
Exploiting the unitary property of the Fourier transform, with $X_k(f;e)$ being the Fourier transform of the EEG signal within an epoch e at a sensor position k and with K sensors in total, we find the covariance matrices in Eq. 6.4 as

$$\Sigma_s = \begin{bmatrix} C^s_{0,0} & C^s_{0,1} & \cdots & C^s_{0,K-1} \\ C^s_{1,0} & C^s_{1,1} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ C^s_{K-1,0} & \cdots & \cdots & C^s_{K-1,K-1} \end{bmatrix} \quad \text{and} \quad \Sigma_n = \begin{bmatrix} C^n_{0,0} & C^n_{0,1} & \cdots & C^n_{0,K-1} \\ C^n_{1,0} & C^n_{1,1} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ C^n_{K-1,0} & \cdots & \cdots & C^n_{K-1,K-1} \end{bmatrix} \tag{6.6}$$

with

$$C^s_{i,j} = \sum_e X_i(f;e)X^*_j(f;e) + X_i(-f;e)X^*_j(-f;e) \tag{6.7}$$

$$C^n_{i,j} = \sum_e X_i(f-\Delta f;e)X^*_j(f-\Delta f;e) + X_i(-f+\Delta f;e)X^*_j(-f+\Delta f;e) \tag{6.8}$$
$$+ X_i(f+\Delta f;e)X^*_j(f+\Delta f;e) + X_i(-f-\Delta f;e)X^*_j(-f-\Delta f;e).$$

A spatial filter **W** can be found for every harmonic component with $f = n \cdot f_{stim}$. After solving Eq. 6.4 **W** is normalized column-wise.

6.4 Results

6.4.1 Behavioral Data

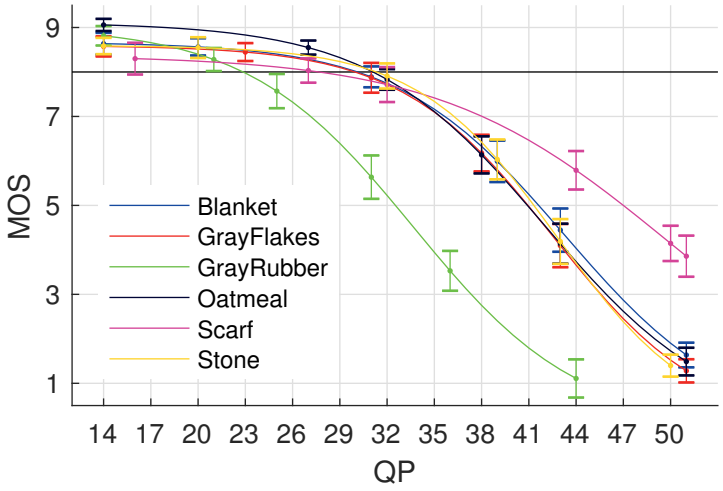
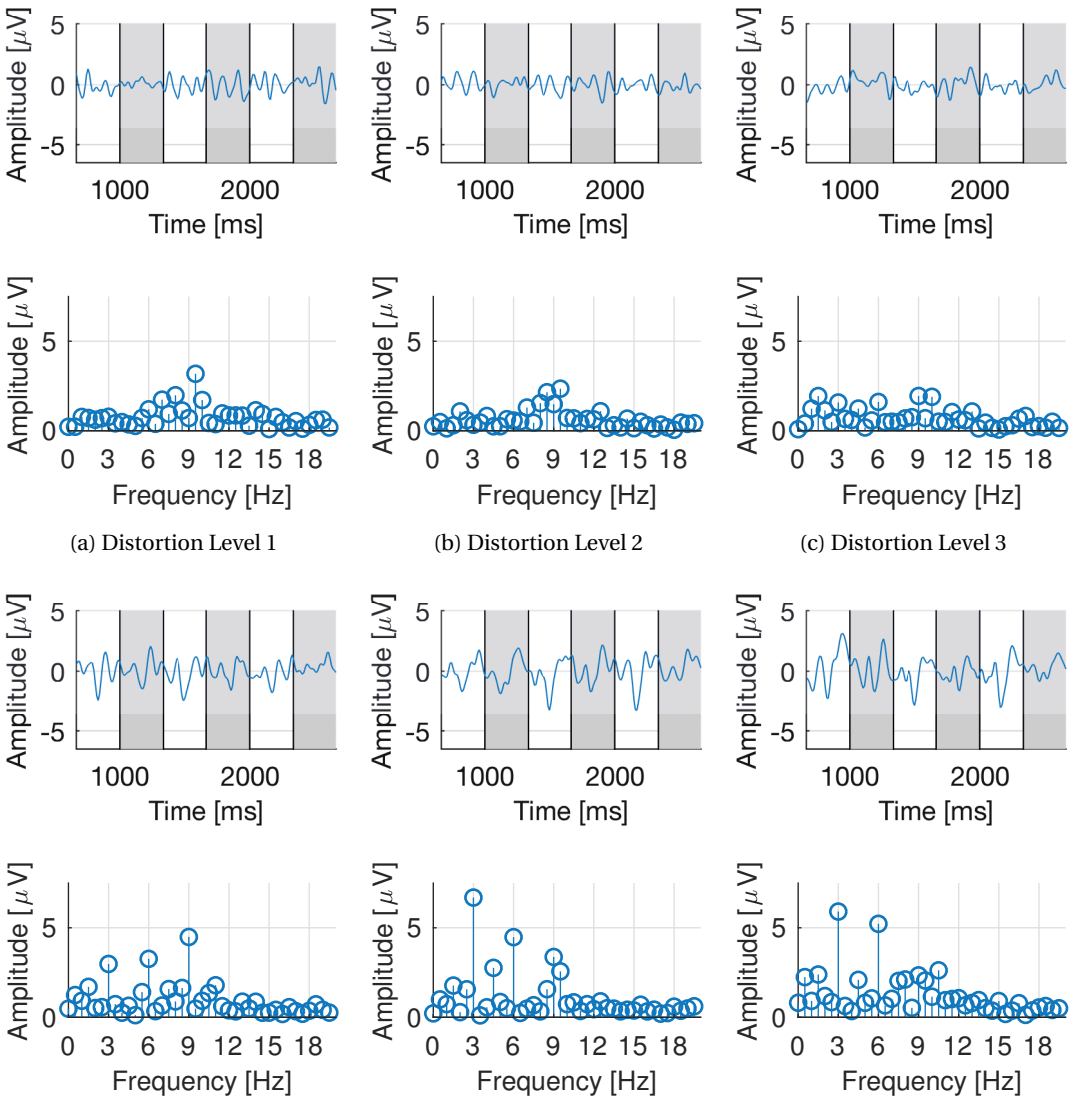


Figure 6.4 – MOS values obtained in the behavioral part of the experiment for all 6 textures. The commonly assumed distortion perception threshold around MOS = 8 is indicated by the horizontal black line. Vertical errorbars denote 95% confidence intervals of MOS values of the specific condition. ©2017 IEEE

Based on screening of the behavioral data, cf. Section 2.2.2, no participant had to be rejected. Fig. 6.4 shows the resulting MOS in dependence of QP. Vertical bars represent the 95% confidence interval of the MOS. Solid lines show interpolated values based on a 3-parameter logistic function $\mathrm{MOS}(\mathrm{QP}) = \frac{\beta_1}{1+e^{-\beta_2 \cdot (\mathrm{QP}-\beta_3)}}$. The horizontal black line indicates the commonly assumed perception threshold at MOS $\approx$ 8 [ITU-R Rec. BT.500-13, 2012]. The two highest quality levels are above the perception threshold for each texture.
As intended, quality levels can be considered as being perceptually approximately equal for different texture images; the resulting MOS are mostly very close and the confidence levels overlap for most instances of equal quality levels. However, the maximal QP in HEVC is QP=51, therefore the quality of *Scarf* could not be reduced any further. Note the extreme cases of images *Scarf* and *GrayRubber*: The horizontally and vertically oriented structure of *Scarf* allows for a relatively good representation by separable DCT and the high contrast is able to mask quantization noise while the rather flat diagonal structure of *GrayRubber* can not be captured by the DCT and structure vanishes due to quantization [Wiegand and Schwarz, 2016].

6.4.2 Neurophysiological Data

An example for the neural response during stimulus presentation is shown in Fig. 6.5 for subject VPik. The plots show the time course and the amplitude spectra measured at Oz for different distortion levels averaged over all trials and textures. An increase of distortion



(a) Distortion Level 1 (b) Distortion Level 2 (c) Distortion Level 3

(d) Distortion Level 4 (e) Distortion Level 5 (f) Distortion Level 6

Figure 6.5 – Epoched neural signals of participant VPik, measured at electrode Oz and averaged over all textures and trials at different distortion levels. **Top/second from bottom**: Amplitude time courses averaged over trials, gray shadings indicate the presentation of a distorted image. **Second from top/bottom**: Amplitude spectra of trial-wise averaged time course. ©2017 IEEE

magnitude triggers a concomitant increase in neural processing at f_{stim} and its harmonics. Thus, as shown by the time courses in Fig. 6.5, the EEG signal becomes increasingly modulated. This modulation can be quantified directly in the spectral domain, where the modulation is represented by increased amplitudes of the spectral components in the frequency bins at the harmonics of f_{stim}. Although we would expect to see this effect at all harmonics, for the signal recorded at Oz this behavior is most evident at the even harmonics $2f_{stim} = 3\,\text{Hz}$ and $4f_{stim} = 6\,\text{Hz}$. Note that it becomes less conclusive for $6f_{stim} = 9\,\text{Hz}$, likely because the SSVEP is buried

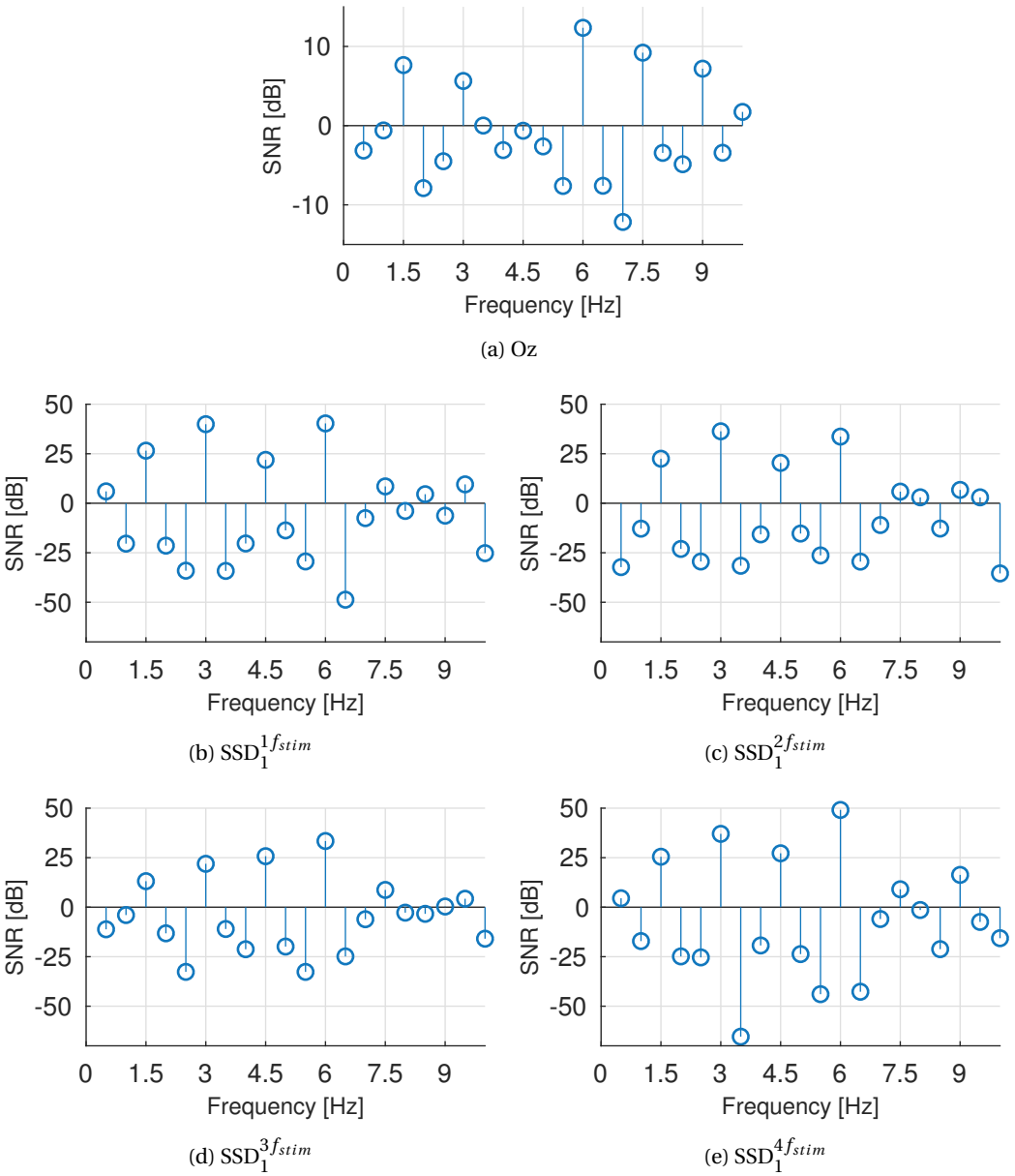
in the alpha band of the EEG. Although SSVEPs are expected to be elicited predominantly at electrode positions covering the visual cortex, brain anatomy varies among individuals and cap positions can not be perfectly aligned between experiments. Thus, optimal (and among subjects comparable) electrode positions are generally unknown for each measurement. This will be addressed in the next subsection.

6.4.3 Spatio-Spectral Decomposition

In order to reduce the spatial dimensionality of the EEG data and to overcome the problem of channel selection, SSD [Nikulin et al., 2011] is used to find a linear combination of sensors that maximizes the SNR . The set of neural sources processing distortion-relevant visual properties is assumed to be dependent on the distortion type (in the experiment HEVC compression), the subject, and the spectral harmonic of the stimulation frequency, but not on distortion magnitude or reference image. Spatial filters are estimated per subject and harmonic over all textures and distortion levels; the first 4 harmonics are considered.
Epoching EEG data to segments of length of 2000 ms yields a frequency resolution of $\Delta f = 0.5$ Hz. The resulting SNR is exemplified for subject VPih and distortion level IV in Fig. 6.6, where the SNR at electrode position Oz is compared to the SNR in the first components of the SSD optimized to different harmonics. Spatial filtering increases the SNR significantly over the SNR at Oz. Surprisingly, the increase of SNR can not only be observed at the harmonic for which the SSD was optimized on (e.g. 3 Hz in Fig. 6.6c, or 4.5 Hz in Fig. 6.6d), but also at other harmonics (1.5 Hz,3 Hz and 6 Hz in Fig. 6.6d). Note that this general increase in SNR cannot be observed for the spectral components buried in the alpha band.
The resulting activation pattern (the 1st component of the SSD optimized for one of the first 4 harmonics) are given in Fig. 6.7 for the subjects VPid and VPih. Note that the signs of the eigenvectors of a matrix are ambiguous, causing different, (approximately) opposing directions of $\mathbf{a}_1(3f_{stim})$ and $\mathbf{a}_1(4f_{stim})$ for VPid and VPih. However, the activation pattern show the neurophysiological plausibility of the obtained filters as the highest activation is found in the area covering the visual cortex, most pronounced around electrode Oz, but also stretching to other electrode locations. The activity on the even harmonics occur more concentrated at occipital channel positions, while on the odd harmonics activity also appears at parietal channel locations. This might be explained by different neural mechanisms sensitive to the presented stimuli and was reported for the 1st and 2nd harmonic previously [Norcia et al., 2014]. However, it is surprising that different neural sources processing symmetric (even harmonics) and asymmetric (odd harmonics) responses do not lead to more distinct spectral profiles of the SNR in Fig. 6.6 for SSD optimized on even vs. odd harmonics.

6.4.4 Relating Neurophysiological to Behavioral Data

The relation between the amplitudes of coherently averaged SSVEP responses and the MOS is shown for the first 4 harmonics in Fig. 6.8 in terms of PLCC for individual subjects (VPal-VPir) and for the grand average (GA) over all subjects. Considered neural signals are *a*) amplitude



(a) Oz

(b) $SSD_1^{1f_{stim}}$

(c) $SSD_1^{2f_{stim}}$

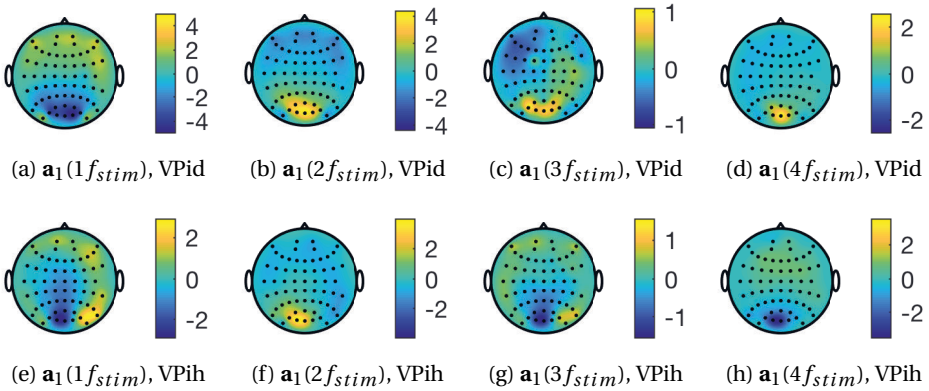
(d) $SSD_1^{3f_{stim}}$

(e) $SSD_1^{4f_{stim}}$

Figure 6.6 – Influence of SSD on SNR for subject VPih: (a) SNR measured at Oz. (b) SNR measured at 1st SSD component optimized on $1f_{stim}$. (c) SNR measured at 2nd SSD component optimized on $2f_{stim}$. (d) SNR measured at 3rd SSD component optimized on $3f_{stim}$. (e) SNR measured at 4th SSD component optimized on $4f_{stim}$. ©2017 IEEE

of SSVEP response at Oz (Oz); *b*) amplitude of SSVEP response in 1st SSD component (SSD_1); and *c*) mean of amplitudes of SSVEP responses in 1st and 2nd SSD components ($SSD_1 + SSD_2$); Significant correlations ($p < 0.05$) are denoted by a filled, non-significant correlations by an empty square. For even harmonics, correlations obtained from the response at Oz (Fig. 6.8b and Fig. 6.8d) are significant for all subjects. For most subjects, the correlation is increased[2]

[2] In absolute values.



(a) $\mathbf{a}_1(1f_{stim})$, VPid (b) $\mathbf{a}_1(2f_{stim})$, VPid (c) $\mathbf{a}_1(3f_{stim})$, VPid (d) $\mathbf{a}_1(4f_{stim})$, VPid

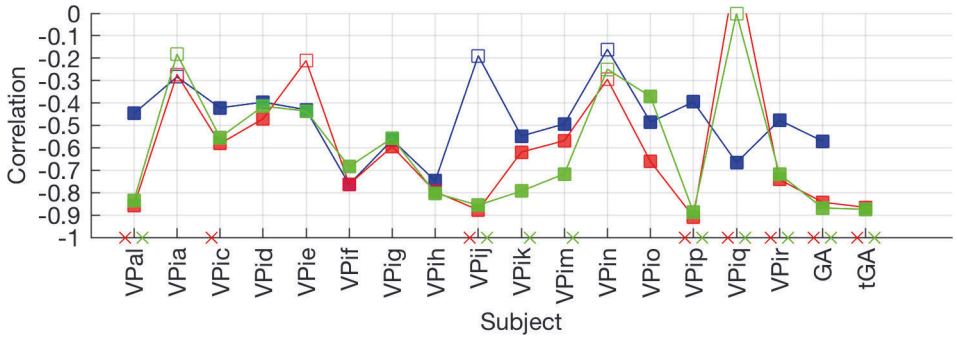
(e) $\mathbf{a}_1(1f_{stim})$, VPih (f) $\mathbf{a}_1(2f_{stim})$, VPih (g) $\mathbf{a}_1(3f_{stim})$, VPih (h) $\mathbf{a}_1(4f_{stim})$, VPih

Figure 6.7 – Activation patterns of the 1st SSD components optimized on the first four harmonics (from left to right). **Top**: For subject VPid. **Bottom**: For subject VPih. ©2017 IEEE

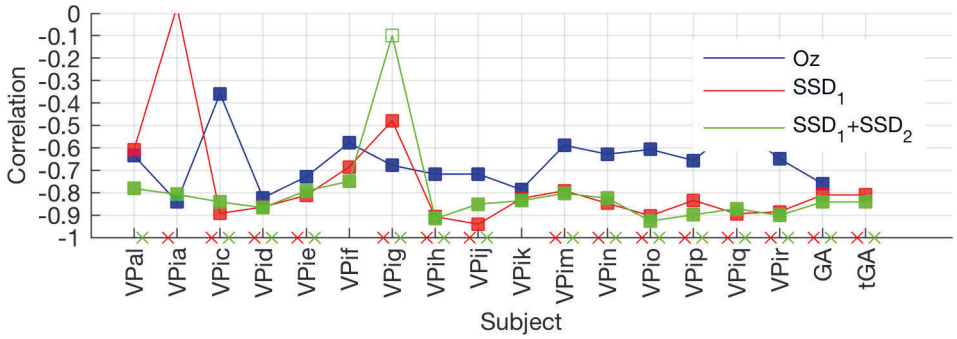
on the 1st SSD component, for some subjects (VPia and VPig on $2f_{stim}$ and VPir on $4f_{stim}$), however, correlations decrease and significance is lost. The statistical significance of the increase in correlation (with regard to the correlation obtained from Oz) is tested using Steiger's Z-test [Howell, 2013]. Significant differences in correlation are denoted by symbol × ($p < 0.05$) above the subject codes in Fig. 6.8. For 12 out of 16 subjects for the 2nd harmonic and 10 out of 16 for the 4th harmonic the increase in correlation by SSD is significant. While for both even harmonics the correlation of the grand average (GA) on the 1st SSD component as well as on the sum of the first two SSD components is increased, this increase is statistically significant only on the 2nd harmonic, but not on the 4th harmonic.
Correlations obtained at the odd harmonics on Oz are in general lower as compared to the correlations on the even harmonics (Fig. 6.8a and Fig. 6.8c) and for fewer subjects significant. Also, SSD is less efficient in increasing the correlation from the one obtained on Oz for single subjects. For the grand average, however, SSD obtains significantly higher correlations on the odd harmonics with gains much higher than on the even harmonics.

6.4.5 Differences between Subjects

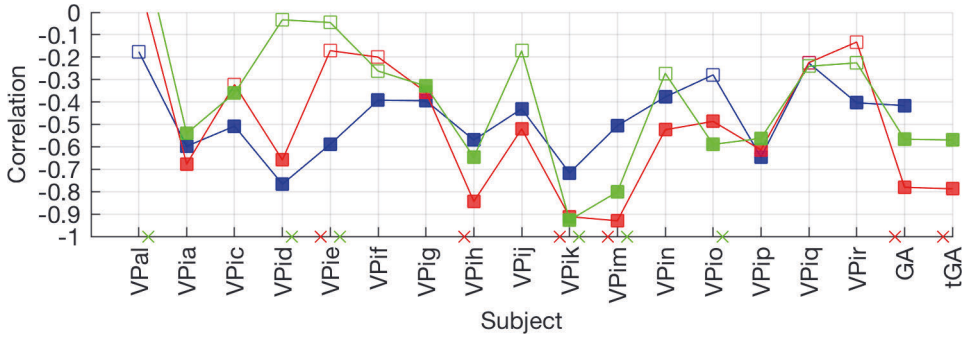
The results (e.g. for subjects VPie and VPiq on $1f_{stim}$, VPia and VPig on $2f_{stim}$, VPie and VPif on $3f_{stim}$, and VPig and VPir on $4f_{stim}$) in Fig. 6.8 show that SSD, is not always successful in extracting the spatial components related to perceived quality. An important reason for this is that SSD is inherently unsupervised. Cases like that are to be expected as they reflect the biological variance among participants. For some subjects (e.g. VPie on $1f_{stim}$, VPia on $2f_{stim}$, or VPig on $4f_{stim}$) the correlation can be re-increased by taking the 2nd SSD component into account as well. Activation patterns for some subjects for that SSD fails are shown in Fig. 6.9. Evidently and in contrast to the results shown in Fig. 6.7, the activation patterns of the 1st component do not focus on the activation in the visual cortex (see e.g. Fig. 6.9a, Fig. 6.9c Fig. 6.9e, Fig. 6.9i, Fig. 6.9m, Fig. 6.9o). This explains the drop in correlation when the



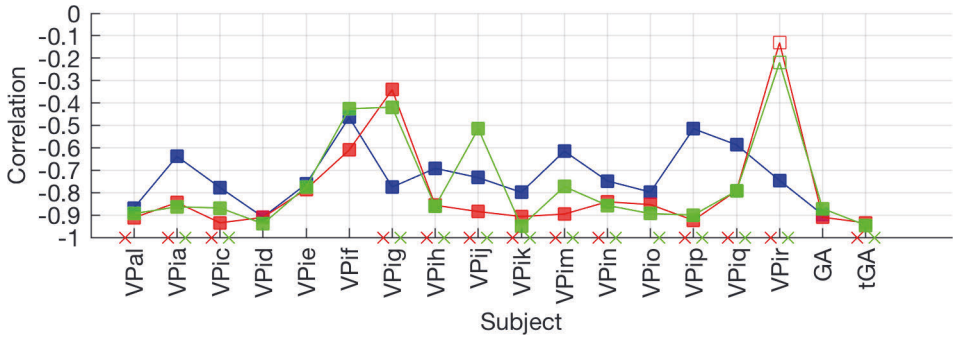
(a) PLCC at $1 f_{stim}$



(b) PLCC at $2 f_{stim}$

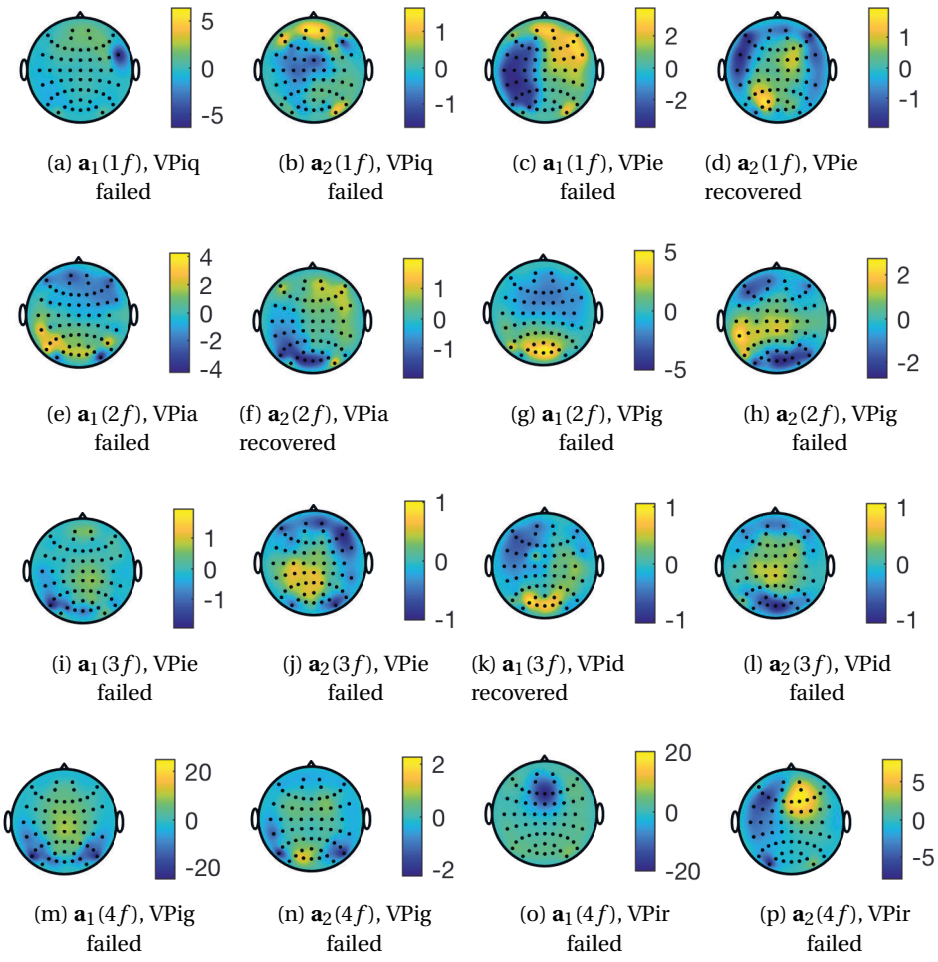


(c) PLCC at $3 f_{stim}$



(d) PLCC at $4 f_{stim}$

Figure 6.8 – Correlations between MOS values and texture-wise averaged neural signal for all subjects (VPal-VPir), averaged over all subjects (GA) and averaged over all remaining subjects after screening by thresholding (tGA). Significant correlations are indicated by a filled square.



(a) $\mathbf{a}_1(1f)$, VPiq failed
(b) $\mathbf{a}_2(1f)$, VPiq failed
(c) $\mathbf{a}_1(1f)$, VPie failed
(d) $\mathbf{a}_2(1f)$, VPie recovered
(e) $\mathbf{a}_1(2f)$, VPia failed
(f) $\mathbf{a}_2(2f)$, VPia recovered
(g) $\mathbf{a}_1(2f)$, VPig failed
(h) $\mathbf{a}_2(2f)$, VPig failed
(i) $\mathbf{a}_1(3f)$, VPie failed
(j) $\mathbf{a}_2(3f)$, VPie failed
(k) $\mathbf{a}_1(3f)$, VPid recovered
(l) $\mathbf{a}_2(3f)$, VPid failed
(m) $\mathbf{a}_1(4f)$, VPig failed
(n) $\mathbf{a}_2(4f)$, VPig failed
(o) $\mathbf{a}_1(4f)$, VPir failed
(p) $\mathbf{a}_2(4f)$, VPir failed

Figure 6.9 – Examples of failing SSD on different harmonics. ©2017 IEEE

amplitude of the SSD components are used as a neural marker of quality. For some subjects, e.g. VPie on $1f_{stim}$ (Fig. 6.9d) or VPia on $2f_{stim}$ (Fig. 6.9f), activity in the visual cortex appears to be captured by the 2nd component. This explains the improved correlation achieved by taking into account the 2nd SSD component as well that can be observed for these subjects (see Fig. 6.8). For other subjects, e.g. VPiq on $1f_{stim}$ (Fig. 6.9b), VPie on $3f_{stim}$ (Fig. 6.9j), or VPir on $4f_{stim}$ (Fig. 6.9p), also the 2nd SSD component fails at extracting physiologically meaningful components and its consideration cannot improve the prediction performance. Note that e.g. for VPig on $2f_{stim}$ extracted activity appears plausible, but SSD fails in terms of correlation to MOS. As e.g. VPid on $3f_{stim}$ shows (Fig. 6.9j) considering the amplitude of the 2nd SSD component can render the success obtained with the 1st component void. Thus, as also observable in Fig. 6.8, it is not advisable to generally use the 2nd SSD component.
As shown in Fig. 6.7 and Fig. 6.9, in many cases activation pattern resulting from successful SSD differ from those resulting from failing SSD. Previously, an outlier detection based on the power contained in the SSD components was proposed for the 4th harmonic [Bosse et al.,

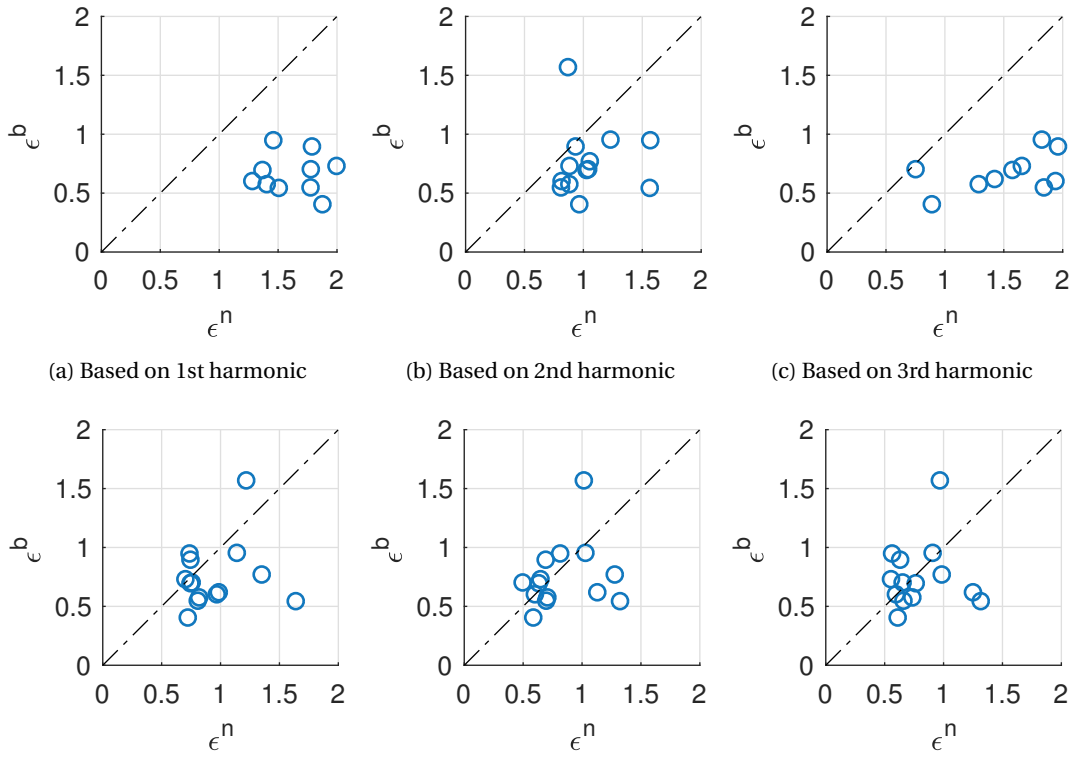
2017a]. However, this approach misses many failing SSD on the first 3 harmonics. A more efficient harmonic-wise outlier detection method is based on the angle between the activation pattern of the 1st SSD components of different subjects: For each harmonic 2 prototypical SSD activation pattern are selected as angular inliers [Harmeling et al., 2006, Krauledat et al., 2008]; for each subject and harmonic, the minimal angular distance to these prototypes is calculated. This allows to define a simple statistical measure for screening subjects with regard to successful SSD, employing a variation of the upper outer fence based on the interquartile range [Tukey, 1977]: With the threshold $\tau = 3 \cdot (Q_{75\%} + Q_{25\%})$ and $Q_{75\%}$ and $Q_{25\%}$ being the first and third quartile of all subject's minimal angular prototype distance, we reject subjects for which the minimal angular prototype distance exceeds τ. Excluding these subjects from the calculation of the grand average further increases the correlation on the 1st and 4th harmonic, as shown in Fig. 6.8 with the outlier-excluded grand average denoted as tGA. In contrast to the correlation obtained on the grand average when considering all subjects, the increase of correlation for the screened grand average is statistically significant with $p < 0.05$ also on the 4th harmonic.

6.4.6 Predicting the MOS from the Neural Signal

As discussed in Section 2.3.3, correlations between neural responses and MOS are reported in several studies on psychophysiological quality assessment. However, for moving towards applicability of neurophysiological methods in quality assessment it is crucial to investigate the predictive power of neurophysiological approaches.
In order to account for subject-wise differences in amplitude ranges [Strasburger et al., 1988], caused e.g. by anatomical differences among the subjects, the amplitudes of the neural signals are normalized subject-wise over all source images and distortion levels to a range between 0 and 1.
We evaluate the prediction performance subject-wise based on a linear model

$$y = \begin{bmatrix} \beta_0 & \cdots & \beta_N \end{bmatrix} \cdot \begin{bmatrix} 1 \\ \vdots \\ x_{N-1} \end{bmatrix} + \epsilon_R \tag{6.9}$$

with $\mathbf{x} = [1 \cdots x_{N-1}]^\top$ denoting the features extracted from the EEG signal, y being MOS values, $\boldsymbol{\beta} = [\beta_0 \cdots \beta_N]$ the regression coefficients and ϵ_R the prediction error.
For subject S_i prediction model parameters $\boldsymbol{\beta}^i$ are estimated based on the $\mathrm{MOS}_{j \neq i}$, obtained by pooling over the ratings of all subjects except S_i, and the EEG features from all subjects except S_i. Obtained parameters $\boldsymbol{\beta}^i$ are used to predict the MOS as $\widehat{\mathrm{MOS}}_i$ from the EEG features of subject S_i.
Prediction is evaluated and compared for 6 different sets of features: 1. the subject-wise averaged amplitude at the 1st harmonic (only 1 feature), 2. the subject-wise averaged amplitude at the 2nd harmonic (only 1 feature), 3. the subject-wise averaged amplitude at the 3rd harmonic (only 1 feature), 4. the subject-wise averaged amplitude at the 4th harmonic (only 1 feature),



(a) Based on 1st harmonic (b) Based on 2nd harmonic (c) Based on 3rd harmonic

(d) Based on 4th harmonic (e) Based on first 4 harmonics (f) Based on first 4 harmonics and interactions

Figure 6.10 – Comparison of accuracies of predictions of the MOS from self-reported behavioral responses (ϵ^{b}) and from different sets of SSVEPs (ϵ^{n}). Each circle represents the MAD between prediction and true MOS for one subject.

5. subject-wise averaged amplitudes at first 4 harmonics (4 features), 6. subject-wise averaged amplitudes at first 4 harmonics and their multiplicative interactions (10 features).
Prediction performance is quantified subject-wise as the MAD between true value and prediction $\epsilon_i^{\mathrm{n}} = |\mathrm{MOS} - \widehat{\mathrm{MOS}}_i|$. For comparison and as behavioral counterpart, the prediction performance of the individual subjects condition-wise rating OS_i for the MOS is quantified as $\epsilon_i^{\mathrm{b}} = |\mathrm{MOS} - \mathrm{OS}_i|$.
Fig. 6.10 scatters the performances of the two prediction schemes for the different sets of neural features. Each circle represents the prediction performances of one subject over all conditions, the dashed line indicates identity of prediction performances. The mean of ϵ_i^b over all subjects is 0.75. Prediction solely based on the amplitude of one of the first 3 harmonics (upper row in Fig. 6.10) is clearly inferior to prediction from behavioral responses. Predicting the MOS solely from the amplitude on the 4^{th} harmonic, a MAD over all subjects of 0.95 is achieved. Although the MAD from the neural prediction is higher than the MAD from the behavioral prediction (also visible as most points are located slightly below the dashed line, and by that suggesting slightly higher accuracy of the behavioral approach), assuming Gaussian distributed ϵ^n and ϵ^b and employing a t-test [Howell, 2013] reveals that the means of ϵ^n

and ϵ^b are statistically equal ($p < 0.05$). Dropping the assumption of Gaussianity, signed-rank testing [Howell, 2013] also shows statistical equality of the medians of ϵ^n and ϵ^b ($p < 0.05$). Joint prediction from amplitudes of all first 4 harmonics obtains a higher prediction accuracy (MAD over subjects: 0.83), additionally considering the multiplicative interactions as well (MAD over subjects: 0.8) further increases accuracy of the prediction from the EEG signal.

6.5 Discussion

This chapter presented a neurophysiological approach to image quality assessment based on SSVEP. SSVEP were elicited by visual stimulation with periodic alternation between reference image and distorted image at a cycle rate of $f_{stim} = 1.5$. Dimensionality of the recorded EEG data was reduced using SSD. For that, SSD was specifically adapted to be used in the frequency domain. This allows for a direct application to SSVEP as SSVEP are naturally represented in the frequency domain. It was shown that for most subjects SSD is able to extract physiologically meaningful components from EEG data recorded in an image quality assessment setup. The neural signal in the extracted components shows a significantly higher correlation to perceived quality than the signal recorded at Oz. Using the first SSD component as a neural marker for quality overcomes the problem of channel selection in SSVEP-based image quality assessment. This is especially favorable as the optimal channel might be different for different subject (due to anatomical variance), different test sessions (due to deviations in electrode positions) and the harmonic component of the SSVEP studied. This chapter did not present a final solution for objective and reliable assessment of video quality, but showed that with the presented method high correlations of the extracted neural signal with MOS values are achieved and that, using a linear model, the proposed method is feasible to predict MOS values from single subject responses with an accuracy comparable to behavioral approaches.

The SSVEP approach is able to achieve a significantly higher SNR than ERP-based approaches [Norcia et al., 2015], and also the number of trials collected per time is much higher compared to ERP-based approaches as no stimulation-free interstimulus period is required. However, as discussed in Section 2.3, no common stimulus dataset for benchmarking neurophysiological quality assessment is available and, thus, studies evaluating SSVEP and ERP for image quality assessment have used different sets of stimuli; this renders a final conclusion difficult. To understand the differences of the two paradigms, it will be important to establish a similar set of stimuli and then conduct further experiments to allow for a precise comparison and identify strengths and weaknesses of the two approaches.
In order to arrive at a real-world solution to quality assessment, this paper is limited in following respects and raises several challenges for future work:
Most important, subjects sensitive to photonic flicker might suffer not only from headache, but even seizures could in principle be evoked by the presentation of a flickering stimulus if the subject is suffering from epilepsy [Fisher et al., 2005]. In an SSVEP-based quality assessment study this must be prevented by identifying and excluding affected subjects from experiments. Eliciting SSVEPs relies on temporally highly precise and alternating stimulus presentation.

This is a clear limitation of the proposed approach and renders SSVEP-based quality assessment 'in the wild' a very hard challenge. However, controlled in-lab studies, e.g. aiming at building huge databases of quality annotated images for testing, designing or training models for quality estimation (cf. Chapter 4 and Chapter 5), are not affected by this conceptual drawback. Here, the linear prediction model was estimated on the data of all other subjects. However, for application scenarios it would be beneficial to identify a subject-wise model that does not rely on other subjects' responses to predict MOS values directly from individual neural responses.

As for all ERP-based approaches as well, the presented SSVEP focuses on the perception of quality change (introduced by the alternation between undistorted and distorted images) rather than quality perception per se. This bounds the approach to the full reference domain and might limit its applicability to only certain real-world applications.

Although our results indicate clear feasibility of the proposed approach, future work should consider to systematically study the influence of low level image statistics such as luminance or contrast on the prediction performance.

In this study presented in this chapter, subjects for which only a low correlation between neural signal and MOS values was obtained (e.g. VPif, see Fig. 6.8) were not identified by conventional screening [ITU-R Rec. BT.500-13, 2012] based on overtly reported quality ratings. Different to psychophysical quality assessment in psychophysiological quality assessment also the neuroanatomy influences the recorded quality related signal. On the recorded data, a simple screening method based on interquartile ranges [Tukey, 1977] was shown to be useful for identifying subjects for which SSD failed. By that the performance of the proposed method in terms of PLCC is statistically significantly increased. Replicability of this screening approach will have to be evaluated on other recordings, other subjects and for paradigms other than SSVEP. Thus, in order to allow for applications of neurophysiological quality assessment methods, analogously to psychophysical methods, appropriate screening techniques will need to be studied further. Identifying and excluding people for which BCI methods fail [Suk et al., 2014, Blankertz et al., 2010, Hammer et al., 2012] is able to boost performance of neurophysiological methods. As an example, [Acqualagna et al., 2015] shows a negative relation between EEG-based distortion detection and the power in the α-band (7.5 Hz to 12.5 Hz) and argues that the α-activity interferes with the processing of the visual information, while a state of high cortical excitability is reflected by decreased α-activity. For future work, this observation can serve as a starting point to study screening methods for EEG-based quality assessment. Also, identifying 'high performing' subjects may reduce the number of subjects necessary for EEG-based quality assessment studies.

Results presented in this study are based on averages across trials. Besides identifying the number of subjects, for real world applications it is crucial to identify the number of trials necessary for reliable quality assessment studies and in the optimal case allow for single trial quality assessment. The question regarding the number of subjects and the number of trials can be treated analogously to psychophysical approaches [ITU-R Rec. BT.500-13, 2012, Pinson et al., 2012].

However, in order to move towards applying the proposed method generically in image quality

assessment studies, parameters of the experimental design need to be optimized. Also here a common dataset allowing for comparable results would be beneficial. For assessment studies based on SSVEP, these factors potentially driving the performance of the approach include the stimulation frequency used for eliciting the SSVEP and the dimensionality reduction method. In the presented study, the stimulation frequency was set to $f_{stim} = 1.5\,\text{Hz}$. It is known that for specific cognitive tasks there are optimal stimulation frequencies [Alonso-Prieto et al., 2013, Won et al., 2015]. Future work should evaluate if such an optimal stimulation frequency also exists for image quality assessment. By using an optimal stimulation frequency, the duration of the experiments might be shortened as the SNR might be increased as less trials may be necessary or due to the presentation of more trials in the same time if a higher stimulation frequency is used. First investigations study the relation between the SNR as a proxy for prediction accuracy or correlation and the stimulation frequency in image quality assessment [Bosse et al., 2018b]. It is not clear, however, if the SNR really translates into higher correlations with perceived quality.

It was shown that different harmonic components represent different neural processing [Liu-Shuang et al., 2015, Norcia et al., 2014]. Activation pattern obtained by SSD indicate that distinct neural mechanisms underlying the processing of distorted images are captured by odd and even harmonics. A deeper understanding of the neural sources driving quality perception might help to improve the experimental design.

For some subjects, SSD failed to increase the correlation to behavioral responses and to extract neurophysiological plausible components. One reason for that (in contrast to CSP in [Acqualagna et al., 2015]) is that SSD is an unsupervised channel decomposition technique. Although a strategy for identifying subjects for which SSD fails was proposed, it would be beneficial to enhance the robustness of dimensionality reduction methods e.g. by divergence methods that allow a higher resistance to outlier trials or other noise contamination [Samek et al., 2014].

Future studies may aim at distortion levels close to the perception threshold as this is a desirable operational point for image communication systems. Here, a neurally informed quality assessment procedure might help to complement conventional behavioral methods, taking into account the heteroskedastic noise characteristics at the edge of perception [Porbadnigk et al., 2015].

The scope of this chapter is the assessment of image quality and for facilitating the experimental setup, source reference content was restricted to texture images. Conceptually there is no reason to limit the proposed approach to this class of stimuli. Thus, it will be interesting to study experimentally whether the proposed approach is also a feasible method to assess perceived quality of complex natural images. SSVEPs have also been used to assess motion perception [Norcia et al., 2015]. Following this line, the feasibility of SSVEP to assess video quality could be evaluated in an extension of the presented experimental setup.

The extra preparation time ($\approx \frac{1}{2}\,\text{h}$) required for the setup of the EEG system might eliminate the benefits of EEG measurements, but a new generation of dry electrode-based EEG caps has the potential to shorten the preparation time drastically. It is recommended for psychophysical experiments not to last longer than 30 minutes, in order to prevent the subjects from

becoming unreliable in their behavioral responses due to fatigue or boredom (cf. Section 2.2. In EEG-based experiments in contrast, no response has to be given by the subjects and it is not known yet what the limits in terms of duration are; in cognitive neuroscience length of EEG-based experiments can range between 2-3 hours.
We evaluated and quantified quality related neural correlates based on an SSVEP paradigm. Clearly, several aspects of the presented method need further evaluations and improvements, but we showed that neural signals that significantly correlate to perceived quality are elicited and that spatial filtering using SSD increases the correlation for most of the subjects. By this, potentially a less biased and more objective measure of quality perception than obtained with conventional behavioral methods can be established.

6.6 Lessons Learned

- SSVEP elicited at a stimulation frequency of $f_{stim} = 1.5\,\text{Hz}$ show a high correlation with overtly reported quality perception.
- SSD can be reformulated in the Fourier domain and efficiently applied to natively represented SSVEP.
- The first SSD component is neurophysiologically plausible and extracts a quality related neural signal from the SSVEP.
- The angular distance between the activity patterns obtained by SSD can be used to detect subjects for which SSVEP-based quality assessment fails.
- An individual subject's neural signal is a reliable predictor for overtly reported quality of a disjoint population.

7 Conclusion

Perceptual signal quality is a central aspect of modern multimedia technology. In this thesis, quality was explored from the perspective of estimation, from the perspective of video compression and from the perspective of assessment.

7.1 Where are we now?

Two novel data-driven approaches to image quality estimation were presented. In Chapter 4, an end-to-end trained deep neural network-based method for NR and FR image quality estimation was presented. To efficiently address the problem of scarcity of quality annotated images, this neural network was designed to estimate local patch-wise qualities that were averaged to a global, image-wise quality estimate. The trained model achieved prediction performance comparable or superior to state-of-the-art methods. A spatially weighted average aggregation scheme was proposed in order to account for label noise that is inherent to the image-wise quality labels and increased by assuming global image-wise quality labels as proxies for local patch-wise quality. Local weight and quality estimation were optimized jointly, in a purely data-driven manner. For FR image quality assessment, this locally weighted pooling further increased prediction performance.
In order to reduce the computational complexity of quality models, Chapter 4 derived a functional definition of distortion sensitivity and showed how this concept leads to a weighted pooling scheme that is very similar to the one introduced in Chapter 5. Distortion sensitivity was modelled as a distortion type-dependent property of the reference image and it was shown (exemplified for the PSNR) that an image-wise consideration of distortion sensitivity drastically improves the prediction accuracy of computational quality models if knowledge about the type of distortion is available. After these conceptual discussions, a neural network was trained to estimate distortion sensitivity patch-wise in an image quality assessment framework based on the PSNR. The achieved prediction performance improved over the PSNR, although the conceptually derived limits could not be reached.
While localizing distortion sensitivity for the PSNR is straight forward and conveniently leads to a weighting scheme for the MSE, as a pitfall for other more sophisticated quality models, a

local distortion sensitive adaptation might be much more difficult to realize.
The crucial advantage of the distortion sensitive PSNR in time critical systems such as video encoding is the restriction of complex processing to the reference image only. Hence, the concept of distortion sensitivity was transferred to rate-distortion theory and a distortion sensitive bit allocation scheme for block-based video coding was proposed. The proposed approach was evaluated in an image compression experiment and significant bit rate savings at identical perceptual quality were achieved. The concept of distortion sensitivity, with its functional definition based on the psychometric regression function, thus bridges directly from quality assessment over (data-driven) quality estimation to perceptual video compression.
The success of data-data driven quality estimation and its application in compression crucially depends on the availability of training data, as Chapter 4 and Chapter 5 suggest. Considering the flaws of conventional psychophysical quality assessment discussed in Chapter 2, as a potential remedy Chapter 6 proposed a novel neurophysiological method for image quality estimation. Subjects were presented with periodic alternations between distorted and undistorted images. The amplitude of the elicited SSVEP were shown to be highly correlated with perceived visual quality of the distorted images. A reformulation of SSD in the frequency domain permitted a convenient dimensionality reduction that can be applied to native representations of the SSVEP. This provided a rational criterion for extracting quality related spatial components. Based on a simple linear model, the proposed method accurately predicted behaviorally reported quality.
It is important to note that SSVEP quality assessment is inherently bound to the full reference domain. The same holds for ERP-based approaches (cf. Chapter 2). This makes an application in a general assessment of quality of experience challenging, if not impossible. Hence, these approaches might be limited to applications excluding real-world and real-time quality assessment in natural viewing environments.

7.2 Outlook

Data-driven quality estimation and neurophysiological quality assessment are new topics in multimedia technology. Several important aspects that were not covered by this dissertation should be investigated in future research.
As already indicated, a crucial factor for robust data-driven quality models is the availability of larger databases of quality annotated images and videos. When creating these databases, it would be beneficial to concentrate on distortion types and resolutions that are of practical relevance. Until such databases are available, machine learning approaches to quality estimation could be studied by using proxy labels, e.g. output of existing and sufficiently trustworthy quality models. Such labels could also be used for pre-training data-driven models, that are then fine-tuned based on real labels.
The data-driven approaches presented can directly be applied to other media types as well, particularly interesting is the application to video signals. Also the extension of spatial distortion sensitivity to spatio-temporal distortion sensitivity is promising.

Models that are optimized on more relevant resolutions and distortion types also present a promising avenue for improving perceptual compression schemes. Beyond that, it is important to study the extent to which models that have been proposed and evaluated for image compression can be generalized to video compression and, eventually, to guide bit allocation using future models that take spatio-temporal distortion sensitivity into account.
The concept of distortion sensitivity may be captured directly in psychophysiological quality assessment with sweep SSVEPs, that have been used previously e.g. to measure thresholds of face detection [Ales et al., 2012]. Open questions for the SSVEP-based assessment of image quality include the identification of a potentially optimal stimulation frequency, more reliable feature extraction and dimensionality reduction methods, and the adaptation and evaluation of the proposed method for video signals.
Currently, one of the biggest obstacles in neurophysiological quality assessment is the lack of comparability of different approaches due to the use of different stimulus material. An important first step towards solving this problem has been taken by the VQEG by publishing a test plan for psychophysiological quality assessment [Bosse et al., 2018a] comprising a dataset of distorted videos available to other researchers.

Bibliography

[Acqualagna et al., 2015] Acqualagna, L., Bosse, S., Porbadnigk, A. K., Curio, G., Müller, K.-R., Wiegand, T., and Blankertz, B. (2015). EEG-based classification of video quality perception using steady state visual evoked potentials (SSVEPs). *Journal of Neural Engineering*, 12(2):026012.

[Albrecht et al., 2018] Albrecht, M., Bartnik, C., Bosse, S., Brandenburg, J., Bross, B., Erfurt, J., George, V., Haase, P., Helle, P., Helmrich, C., Henkel, A., Hinz, T., de Luxan Hernandes, S., Kaltenstadler, S., Keydel, P., Kirchhoffer, H., Lehmann, C., Lim, W.-Q., Ma, J., Maniry, D., Marpe, D., Merkle, P., Nguyen, T., Pfaff, J., Rasch, J., Rischke, R., Rudat, C., Schaefer, M., Schierl, T., Schwarz, H., Siekmann, M., Skupin, R., Stallenberger, B., Stegemann, J., Suehring, K., Tech, G., Venugopal, G., Walter, S., Wieckowiski, A., Wiegand, T., and Winken, M. (2018). Description of SDR, HDR and 360° video coding technology proposal by Fraunhofer HHI. In *Joint Video Exploration Team (JVET) of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, JVET-J0014*, San Diego, CA, USA.

[Ales et al., 2012] Ales, J. M., Farzin, F., Rossion, B., and Norcia, A. M. (2012). An objective method for measuring face detection thresholds using the sweep steady-state visual evoked response. *Journal of Vision*, 12(10):1–18.

[Ales and Norcia, 2009] Ales, J. M. and Norcia, A. M. (2009). Assessing direction-specific adaptation using the steady-state visual evoked potential: Results from EEG source imaging. *Journal of Vision*, 9(7):8–8.

[Alonso-Prieto et al., 2013] Alonso-Prieto, E., Van Belle, G., Liu-Shuang, J., Norcia, A. M., and Rossion, B. (2013). The 6 Hz fundamental stimulation frequency rate for individual face discrimination in the right occipito-temporal cortex. *Neuropsychologia*, 51(13):2863–75.

[Antons et al., 2010] Antons, J.-N., Porbadnigk, A. K., Schleicher, R., Blankertz, B., Möller, S., and Curio, G. (2010). Subjective listening tests and neural correlates of speech degradation in case of signal-correlated noise. In *Audio Engineering Society Convention 129*, pages 1–4.

[Antons et al., 2012a] Antons, J.-N., Schleicher, R., Arndt, S., Möller, S., and Curio, G. (2012a). Too tired for calling? A physiological measure of fatigue caused by bandwidth limitations. *4th International Workshop on Quality of Multimedia Experience (QoMEX)*, pages 63–67.

[Antons et al., 2012b] Antons, J.-N., Schleicher, R., Arndt, S., Möller, S., Porbadnigk, A. K., and Curio, G. (2012b). Analyzing speech quality perception using electroencephalography. *IEEE Journal of Selected Topics in Signal Processing*, 6(6):721–731.

[Arndt et al., 2014a] Arndt, S., Antons, J.-N., Schleicher, R., Möller, S., and Curio, G. (2014a). Using electroencephalography to measure perceived video quality. *IEEE Journal of Selected Topics in Signal Processing*, 8(3):366–376.

[Arndt et al., 2014b] Arndt, S., Radun, J., Antons, J. N., and Möller, S. (2014b). Using eye tracking and correlates of brain activity to predict quality scores. In *Sixth International Workshop on Quality of Multimedia Experience*, pages 281–285.

[Avarvand et al., 2017a] Avarvand, F. S., Bosse, S., Müller, K.-R., Schäfer, R., Nolte, G., Wiegand, T., Curio, G., and Samek, W. (2017a). Objective quality assessment of stereoscopic images with vertical disparity using EEG. *Journal of Neural Engineering*, 14(4):046009.

[Avarvand et al., 2017b] Avarvand, F. S., Bosse, S., Nolte, G., Wiegand, T., and Samek, W. (2017b). Measuring the quality of 3D visualizations using EEG: A time-frequency approach. In *Proceedings of the 7th Graz Brain-Computer Interface Conference*, pages 441–446. Verlag der TU Graz.

[Bach and Meigen, 1999] Bach, M. and Meigen, T. (1999). Do's and dont's in Fourier analysis of steady-state potentials. *Documenta ophthalmologica. Advances in ophthalmology*, 99(1):69–82.

[Bach et al., 2015] Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE*, 10(7):e0130140.

[Bang et al., 2014] Bang, J. W., Heo, H., Choi, J. S., and Park, K. R. (2014). Assessment of eye fatigue caused by 3D displays based on multimodal measurements. *Sensors*, 14(9):16467–16485.

[Bell and Sejnowski, 1997] Bell, A. J. and Sejnowski, T. J. (1997). The 'independent components' of natural scenes are edge filters. *Vision Research*, 37(23):3327–3338.

[Berger, 1933] Berger, H. (1933). Über das Elektrencephalogramm des Menschen. *European Archives of Psychiatry and Clinical Neuroscience*, 98(1):231–254.

[Blankertz et al., 2011] Blankertz, B., Lemm, S., Treder, M., Haufe, S., and Müller, K.-R. (2011). Single-trial analysis and classification of ERP components - A tutorial. *NeuroImage*, 56(2):814–825.

[Blankertz et al., 2010] Blankertz, B., Sannelli, C., Halder, S., Hammer, E. M., Kübler, A., Müller, K.-R., Curio, G., and Dickhaus, T. (2010). Neurophysiological predictor of SMR-based BCI performance. *NeuroImage*, 51(4):1303–1309.

[Blankertz et al., 2008] Blankertz, B., Tomioka, R., Lemm, S., Kawanabe, M., and Müller, K.-R. (2008). Optimizing spatial filters for robust EEG single-trial analysis. *IEEE Signal Processing Magazine*, 25(1):41–56.

[Blum and Rutkove, 2007] Blum, A. S. and Rutkove, S. B. (2007). *The clinical neurophysiology primer*. Humana Press, Totowa, NJ, USA.

[Bosse et al., 2014] Bosse, S., Acqualagna, L., Porbadnigk, A., Blankertz, B., Curio, G., Müller, K.-R., and Wiegand, T. (2014). Neurally informed assessment of perceived natural texture image quality. In *Proceedings of IEEE International Conference on Image Processing (ICIP)*, pages 1987–1991.

[Bosse et al., 2015] Bosse, S., Acqualagna, L., Porbadnigk, A. K., Curio, G., Müller, K.-R., Blankertz, B., and Wiegand, T. (2015). Neurophysiological assessment of perceived image quality using steady-state visual evoked potentials. In *Applications of Digital Image Processing XXXVIII*, volume 9599, pages 959914–959914.

[Bosse et al., 2017a] Bosse, S., Acqualagna, L., Samek, W., Porbadnigk, A. K., Curio, G., Blankertz, B., Müller, K.-R., and Wiegand, T. (2017a). Assessing perceived image quality using steady-state visual evoked potentials and spatio-spectral decomposition. *IEEE Transactions on Circuits and Systems for Video Technology*, 8215(c):1–1.

[Bosse et al., 2018a] Bosse, S., Arndt, S., Engelke, U., Martini, M., Ramzan, N., and Brunnström, K. (2018a). The VQEG testplan for psychophysiological video quality assessment. *Quality and User Experience, submitted.*

[Bosse et al., 2018b] Bosse, S., Bagdasarian, M., Samek, W., Curio, G., and Wiegand, T. (2018b). On the stimulation frequency in SSVEP-based image quality assessment. In *10th International Conference on Quality of Multimedia Experience (QoMEX), accepted for publication.*

[Bosse et al., 2018c] Bosse, S., Becker, S., Fisches, Z., Samek, W., and Wiegand, T. (2018c). Neural network-based estimation of distortion sensitivity for image quality prediction. In *Proceeding of the IEEE International Conference on Image Processing (ICIP), accepted for publication.*

[Bosse et al., 2018d] Bosse, S., Becker, S., Müller, K.-R., Samek, W., and Wiegand, T. (2018d). Estimation of distortion sensitivity for visual quality prediction using a convolutional neural network. *Digital Signal Processing, submitted.*

[Bosse et al., 2016a] Bosse, S., Chen, Q., Siekmann, M., Samek, W., and Wiegand, T. (2016a). Shearlet-based reduced reference image quality assessment. In *Proceeding of the IEEE International Conference on Image Processing (ICIP)*, pages 2052–2056.

[Bosse et al., 2017b] Bosse, S., Helmrich, C., Schwarz, H., Marpe, D., and Wiegand, T. (2017b). Perceptually optimized QP adaptation and associated distortion measure. In *Joint Video Exploration Team (JVET) of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, JVET-H0047*, Macao, China.

[Bosse et al., 2018e] Bosse, S., Maniry, D., Müller, K.-R., Wiegand, T., and Samek, W. (2018e). Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on Image Processing*, 27(1):206–219.

[Bosse et al., 2016b] Bosse, S., Maniry, D., Müller, K.-R. R., Wiegand, T., and Samek, W. (2016b). Neural network-based full-reference image quality assessment. In *Proceedings of the Picture Coding Symposium (PCS)*, pages 1–5.

[Bosse et al., 2016c] Bosse, S., Maniry, D., Wiegand, T., and Samek, W. (2016c). A deep neural network for image quality assessment. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pages 3773–3777.

[Bosse et al., 2016d] Bosse, S., Müller, K.-R., Wiegand, T., and Samek, W. (2016d). Brain-computer interfacing for multimedia quality assessment. In *Proc. of the IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2834–2839.

[Bosse et al., 2012] Bosse, S., Schwarz, H., Hinz, T., and Wiegand, T. (2012). Encoder control for renderable regions in high efficiency multiview video plus depth coding. In *Picture Coding Symposium (PCS)*, pages 129–132.

[Bosse et al., 2016e] Bosse, S., Siekmann, M., Rasch, J., Wiegand, T., and Samek, W. (2016e). Quality assessment of image patches distorted by image Compression using crowdsourcing. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6.

[Bosse et al., 2017c] Bosse, S., Siekmann, M., Samek, W., and Wiegand, T. (2017c). A perceptually relevant shearlet-based adaptation of the PSNR. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pages 315–319.

[Bossen, 2013] Bossen, F. (2013). Common test conditions and software reference configurations Output. In *Joint Collaborative Team on Video Coding (JCT-VC) of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG1, JCTVC-L110*, San José CA, USA.

[Boto et al., 2018] Boto, E., Holmes, N., Leggett, J., Roberts, G., Shah, V., Meyer, S. S., Muñoz, L. D., Mullinger, K. J., Tierney, T. M., Bestmann, S., Barnes, G. R., Bowtell, R., and Brookes, M. J. (2018). Moving magnetoencephalography towards real-world applications with a wearable system. *Nature*, 555(7698):657–661.

[Brandenburg et al., 2013] Brandenburg, K., Faller, C., Herre, J., Johnston, J. D., and Kleijn, W. B. (2013). Perceptual coding of high-quality digital audio. *Proceedings of the IEEE*, 101(9):1905–1919.

[Bromley et al., 1993] Bromley, J., Bentz, J. W., Bottou, L., Guyon, I., LeCun, Y., Moore, C., Säckinger, E., and Shah, R. (1993). Signature verification using a "Siamese" time delay neural network. *International Journal of Pattern Recognition and Artificial Intelligence*, 7(04):669–688.

[Cacioppo et al., 2016] Cacioppo, J. T., Tassinary, L. G., and Berntson, G. G. (2016). *Handbook of Psychophysiology*. Cambridge University Press.

[Callet et al., 2012] Callet, P. L., Möller, S., and Perkis, A. (2012). Qualinet white paper on definitions of quality of experience. *European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003)*.

[Chen et al., 2013] Chen, C., Li, K., Wu, Q., Wang, H., Qian, Z., and Sudlow, G. (2013). EEG-based detection and evaluation of fatigue caused by watching 3DTV. *Displays*, 34(2):81–88.

[Chen et al., 2014] Chen, C., Wang, J., Li, K., Wu, Q., Wang, H., Qian, Z., and Gu, N. (2014). Assessment visual fatigue of watching 3DTV using EEG power spectral parameters. *Displays*, 35(5):266–272.

[Cho et al., 2012] Cho, H., Kang, M.-K., Yoon, K.-J., and Jun, S. C. (2012). Feasibility study for visual discomfort assessment on stereo images using EEG. *2012 International Conference on 3D Imaging (IC3D)*, pages 1–6.

[Chopra et al., 2005] Chopra, S., Hadsell, R., and Y., L. (2005). Learning a similiarty metric discriminatively, with application to face verification. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 349–356.

[Cisco, 2017] Cisco (2017). Cisco visual networking index: Forecast and methodology, 2016–2021. Technical report, Cisco.

[Daly, 1992] Daly, S. (1992). Visible differences predictor: an algorithm for the assessment of image fidelity. *Proc. SPIE 1666, Human Vision, Visual Processing, and Digital Display III*, 1666(1992):2–16.

[Darcy et al., 2016] Darcy, D., Gitterman, E., Brandmeyer, A., Daly, S., Crum, P., Laboratories, D., and Francisco, S. (2016). Physiological capture of augmented viewing states : objective measures of high-dynamic-range and wide-color-gamut viewing experiences. *Electronic Imaging*, 2016(16):1–9.

[Deng et al., 2009] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255.

[Dias et al., 2015a] Dias, A., Schwarz, S., Siekmann, M., Bosse, S., Schwarz, H., Marpe, D., Zubrzycki, J., and Mrak, M. (2015a). Perceptually Optimised Video Compression. In *IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 1–4.

[Dias et al., 2015b] Dias, A. S., Siekmann, M., Bosse, S., Schwarz, H., Marpe, D., and Mrak, M. (2015b). Rate-distortion optimised quantisation for HEVC using spatial just noticeable distortion. In *Proceedings of the European Signal Processing Conference (EUSIPCO)*, pages 110–114.

[Dmochowski and Norcia, 2015] Dmochowski, J. P. and Norcia, A. M. (2015). Cortical components of reaction-time during perceptual decisions in humans. *PLoS ONE*, 10(11):e0143339.

[Donley et al., 2015] Donley, J., Ritz, C., and Shujau, M. (2015). Analysing the Quality of Experience of Multisensory Media from the Measurements of Physiological Responses. In *Proc. of the IEEE International Conference on Wireless Communications and Signal Processing*, pages 1–6.

[Dosselmann and Yang, 2009] Dosselmann, R. and Yang, X. D. (2009). A comprehensive assessment of the structural similarity index. *Signal, Image and Video Processing*, 5(1):81–91.

[Engelke et al., 2017] Engelke, U., Darcy, D., Mulliken, G., Bosse, S., Martini, M., Arndt, S., Antons, J.-N., Chan, K., Ramzan, N., and Brunnström, K. (2017). Psychophysiology-Based QoE Assessment: A Survey. *IEEE Journal of Selected Topics in Signal Processing*, 11(1):6–21.

[Everett, 1963] Everett, H. (1963). Generalized Lagrange Multiplier Method for Solving Problems of Optimum Allocation of Resources. *Operations Research*, 11(3):399–417.

[Farzin et al., 2012] Farzin, F., Hou, C., and Norcia, A. M. (2012). Piecing it together: Infants' neural responses to face and object structure. *Journal of Vision*, 12(13):6–6.

[Fechner, 1907] Fechner, G. T. (1907). *Elemente der Psychophysik. 2.* Breitkopf & Härtel.

[Fisher et al., 2005] Fisher, R. S., Harding, G., Erba, G., Barkley, G. L., and Wilkins, A. (2005). Photic- and pattern-induced seizures: A review for the epilepsy foundation of america working group. *Epilepsia*, 46(9):1426–1441.

[Frey et al., 2015] Frey, J., Appriou, A., Lotte, F., and Hachet, M. (2015). Classifying EEG Signals during Stereoscopic Visualization to Estimate Visual Comfort. *Computational Intelligence and Neuroscience*, 2016:7.

[Gao et al., 2017] Gao, F., Wang, Y., Li, P., Tan, M., Yu, J., and Zhu, Y. (2017). DeepSim: Deep similarity for image quality assessment. *Neurocomputing*, 257:104–114.

[Garcia et al., 2014] Garcia, M.-N., Argyropoulos, S., Staelens, N., Naccari, M., Rios-Quintero, M., Raake, A., and Möller, S. (2014). *Quality of experience*. Springer.

[Garcia Freitas et al., 2015] Garcia Freitas, P., Redi, J. A., Farias, M. C. Q., and Silva, A. F. (2015). Video quality ruler: A new experimental methodology for assessing video quality. In *Seventh International Workshop on Quality of Multimedia Experience (QoMEX)*, pages 1–6.

[Ghadiyaram and Bovik, 2015] Ghadiyaram, D. and Bovik, A. C. (2015). LIVE In the Wild Image Quality Challenge Database. http://live.ece.utexas.edu/research/ChallengeDB/index.html.

[Ghadiyaram and Bovik, 2016] Ghadiyaram, D. and Bovik, A. C. (2016). Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing*, 25(1):372–387.

[Ghadiyaram and Bovik, 2017] Ghadiyaram, D. and Bovik, A. C. (2017). Perceptual Quality Prediction on Authentically Distorted Images Using a Bag of Features Approach. *Journal of Vision*, 17(1):32.

[Girod, 1993] Girod, B. (1993). What's Wrong with Mean-squared Error? In *Digital Images and Human Vision*, pages 207–220. MIT Press.

[Girshick, 2015] Girshick, R. (2015). Fast R-CNN. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448.

[Hammer et al., 2012] Hammer, E. M., Halder, S., Blankertz, B., Sannelli, C., Dickhaus, T., Kleih, S., Müller, K.-R., and Köbler, A. (2012). Psychological predictors of SMR-BCI performance. *Biological Psychology*, 89(1):80–86.

[Hands, 2004] Hands, D. S. (2004). A basic multimedia quality model. *IEEE Transactions on Multimedia*, 6(6):806–816.

[Handy, 2005] Handy, T. C. (2005). *Event-related potentials: A methods handbook*. MIT press.

[Hanhart and Ebrahimi, 2014] Hanhart, P. and Ebrahimi, T. (2014). Calculation of average coding efficiency based on subjective quality scores. *Journal of Visual Communication and Image Representation*, 25(3):555–564.

[Harmeling et al., 2006] Harmeling, S., Dornhege, G., Tax, D., Meinecke, F., and Müller, K. R. (2006). From outliers to prototypes: Ordering data. *Neurocomputing*, 69(13-15):1608–1618.

[Haufe et al., 2014a] Haufe, S., Dähne, S., and Nikulin, V. V. (2014a). Dimensionality reduction for the analysis of brain oscillations. *NeuroImage*, 101:583–597.

[Haufe et al., 2014b] Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J. D., Blankertz, B., and Bießmann, F. (2014b). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, 87:96–110.

[He et al., 2015] He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034.

[He et al., 2016] He, Y., Tu, Y., Wang, L., Jia, L., Guo, J., and Chen, Y. (2016). A study on the relationship between event-related potentials and image quality variation. *Journal of Display Technology*, 12(6):532–541.

[Herculano-Houzel, 2009] Herculano-Houzel, S. (2009). The human brain in numbers: a linearly scaled-up primate brain. *Frontiers in Human Neuroscience*, 3(November):1–11.

[Hess and Polt, 1964] Hess, E. H. and Polt, J. M. (1964). Pupil Size in Relation to Mental Activity during Simple Problem-Solving. *Science*, 143(3611):1190–2.

[Hoßfeld et al., 2011] Hoßfeld, T., Schatz, R., and Egger, S. (2011). SOS : The MOS is not enough. In *Third International Workshop on Quality of Multimedia Experience (QoMEX)*, pages 131–136.

[Hotelling, 1933] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417.

[Howell, 2013] Howell, D. C. (2013). *Statistical Methods for Psychology*. Cengage Learning.

[Hu et al., 2015] Hu, S., Jin, L., Wang, H., Zhang, Y., Kwong, S., and Kuo, C.-C. J. (2015). Compressed image quality metric based on perceptually weighted distortion. *IEEE Transactions on Image Processing*, 24(12):5594–5608.

[Huynh-Thu et al., 2011] Huynh-Thu, Q., Garcia, M. N., Speranza, F., Corriveau, P., and Raake, A. (2011). Study of rating scales for subjective quality assessment of high-definition video. *IEEE Transactions on Broadcasting*, 57(1):1–14.

[ITU-R Rec. BT.500-13, 2012] ITU-R Rec. BT.500-13 (2012). Methodology for the subjective assessment of the quality of television pictures.

[ITU-T Rec. P.910, 2008] ITU-T Rec. P.910 (2008). Subjective video quality assessment methods for multimedia applications.

[ITU-T Rec. P.911, 1998] ITU-T Rec. P.911 (1998). Subjective audiovisual quality assessment methods for multimedia applications.

[ITU-T Rec. P.912, 2016] ITU-T Rec. P.912 (2016). Subjective video quality assessment methods for recognition tasks.

[Janowski and Pinson, 2015] Janowski, L. and Pinson, M. (2015). The accuracy of subjects in a quality experiment: A theoretical subject model. *IEEE Transactions on Multimedia*, 17(12):2210–2224.

[Jansen et al., 1995] Jansen, A. S. P., Van Nguyen, X., Karpitskiy, V., Mettenleiter, T. C., and Loewy, A. D. (1995). Central command neurons of the sympathetic nervous system: basis of the fight-or-flight response. *Science*, 270(5236):644.

[JCT-VC, 2014] JCT-VC (2014). Subversion repository for the HEVC test model reference software.

[Kafkas, 2012] Kafkas, A. (2012). Familiarity and recollection produce distinct eye movement, pupil and medial temporal lobe responses when memory strength is matched. *Neuropsychologia*, 50(13):3080–3093.

[Kang et al., 2014] Kang, L., Ye, P., Li, Y., and Doermann, D. (2014). Convolutional neural networks for no-reference image quality assessment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1733–1740.

[Kim and Lee, 2017] Kim, J. and Lee, S. (2017). Fully deep blind image quality predictor. *IEEE Journal on Selected Topics in Signal Processing*, 11(1):206–220.

[Kingma and Ba, 2014] Kingma, D. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[Knoche et al., 1999] Knoche, H., Meer, H. D., and Kirsh, D. (1999). Utility curves: Mean opinion scores considered biased. In *Seventh International Workshop on Quality of Service*, pages 12–14.

[Kohler et al., 2018] Kohler, P. J., Cottereau, B. R., and Norcia, A. M. (2018). Dynamics of perceptual decisions about symmetry in visual cortex. *NeuroImage*, 167:316–330.

[Krauledat et al., 2008] Krauledat, M., Tangermann, M., Blankertz, B., and Müller, K. R. (2008). Towards zero training for brain-computer interfacing. *PLoS ONE*, 3(8):1435–1439.

[Kroupi et al., 2014] Kroupi, E., Hanhart, P., Lee, J.-S., Rerabek, M., and Ebrahimi, T. (2014). EEG correlates during video quality perception. In *Proceedings of the 22nd European Signal Processing Conference (EUSIPCO)*, pages 1–4. IEEE.

[Kwak et al., 2015] Kwak, N.-S., Müller, K.-R., and Lee, S.-W. (2015). A lower limb exoskeleton control system based on steady state visual evoked potentials. *Journal of Neural Engineering*, 12(5):056009.

[Kylberg, 2011] Kylberg, G. (2011). The Kylberg Texture Dataset v. 1.0. External report (Blue series) 35, Centre for Image Analysis, Swedish University of Agricultural Sciences and Uppsala University, Uppsala, Sweden.

[Laparra et al., 2016] Laparra, V., Ballé, J., Berardino, A., and Simoncelli, E. P. (2016). Perceptual image quality assessment using a normalized Laplacian pyramid. *Electronic Imaging*, 2016(16):1–6.

[Larson and Chandler, 2010] Larson, E. C. and Chandler, D. M. (2010). Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of Electronic Imaging*, 19(1):011006.

[Lecun et al., 2015] Lecun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

[LeCun et al., 1998] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2323.

[LeCun et al., 2012] LeCun, Y., Bottou, L., Orr, G. B., and Müller, K.-R. (2012). Efficient backprop. In *Neural networks: Tricks of the trade*, pages 9–50. Springer.

[Lee et al., 2015] Lee, S.-W., Bülthoff, H. H., and Müller, K.-R. (2015). *Recent progress in brain and cognitive engineering*. Springer.

[Li et al., 2008] Li, H. O., Seo, J., Kham, K., and Lee, S. (2008). Measurement of 3D visual fatigue using event-related potential (ERP): 3D oddball paradigm. In *2008 3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video*, pages 213–216. IEEE.

[Li and Bampis, 2017] Li, Z. and Bampis, C. G. (2017). Recover subjective quality scores from noisy measurements. *Data Compression Conference (DCC)*, Part F1277:52–61.

[Likert, 1932] Likert, R. (1932). A technique for the measurement of attitudes. *Archives of psychology*, (140):1–55.

[Lin et al., 2014] Lin, J. Y., Liu, T. J., Wu, E. C. H., and Kuo, C. C. (2014). A fusion-based video quality assessment (FVQA) index. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pages 1–5.

[Lin and Kuo, 2011] Lin, W. and Kuo, C.-C. J. (2011). Perceptual visual quality metrics: A survey. *Journal of Visual Communication and Image Representation*, 22(4):297–312.

[Lindemann and Magnor, 2011] Lindemann, L. and Magnor, M. (2011). Assessing the quality of compressed images using EEG. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pages 3109–3112.

[Lindemann et al., 2011] Lindemann, L., Wenger, S., and Magnor, M. (2011). Evaluation of video artifact perception using event-related potentials. In *Proceedings of the ACM SIGGRAPH Symposium on Applied Perception in Graphics and Visualization*, pages 53–58.

[Little, 2007] Little, D. (2007). The issue: The common European framework of reference for languages: Perspectives on the making of supranational language education policy. *Modern Language Journal*, 91(4):645–655.

[Liu et al., 2012] Liu, A., Lin, W., and Narwaria, M. (2012). Image Quality Assessment Based on Gradient Similarity. *IEEE Transaction on Image Processing*, 21(4):1500–1512.

[Liu-Shuang et al., 2015] Liu-Shuang, J., Ales, J. J. M., Rossion, B., and Norcia, A. M. (2015). The effect of contrast polarity reversal on face detection: Evidence of perceptual asymmetry from sweep VEP. *Vision research*, 108:8–19.

[Long et al., 2015] Long, J., Shelhamer, E., and Darrell, T. (2015). Fully Convolutional Networks for Semantic Segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440.

[Lopes da Silva, 2013] Lopes da Silva, F. (2013). EEG and MEG: Relevance to neuroscience. *Neuron*, 80(5):1112–1128.

[Lubin, 1997] Lubin, J. (1997). A human vision system model for objective picture quality measurements. *International Broadcasting Convention*, pages 498–503.

[Luck, 2014] Luck, S. J. (2014). *An introduction to the event-related potential technique*. MIT Press.

[Luck and Kappenman, 2011] Luck, S. J. and Kappenman, E. S. (2011). *The Oxford handbook of event-related potential components*. Oxford University Press.

[Lukin et al., 2015] Lukin, V. V., Ponomarenko, N. N., Ieremeiev, O. I. O., Egiazarian, K. O., and Astola, J. (2015). Combining full-reference image visual quality metrics by neural network. In *SPIE/IS&T Electronic Imaging*, volume 9394, pages 93940K—-93940K.

[Marpe et al., 2010a] Marpe, D., Schwarz, H., Bosse, S., Bross, B., Helle, P., Hinz, T., Kirchhoffer, H., Lakshman, H., Nguyen, T., Oudin, S., Siekmann, M., Sühring, K., Winken, M., and Wiegand, T. (2010a). Highly efficient video compression using quadtree structures and improved techniques for motion representation and entropy coding. In *Picture Coding Symposium (PCS)*, pages 206–209.

[Marpe et al., 2010b] Marpe, D., Schwarz, H., Bosse, S., Bross, B., Helle, P., Hinz, T., Kirchhoffer, H., Lakshman, H., Nguyen, T., Oudin, S., Siekmann, M., Suhring, K., Winken, M., and Wiegand, T. (2010b). Video compression using nested quadtree structures, leaf merging, and improved techniques for motion representation and entropy coding. *IEEE Transactions on Circuits and Systems for Video Technology*, 20(12):1676–1687.

[Marpe et al., 2011] Marpe, D., Schwarz, H., Wiegand, T., Bosse, S., Bross, B., Helle, P., Hinz, T., Kirchhoffer, H., Lakshman, H., Nguyen, T., Oudin, S., Siekmann, M., Sühring, K., and Winken, M. (2011). Improved video compression technology and the emerging high efficiency video coding standard. In *Proc. of the IEEE International Conference on Consumer Electronics (ICCE)*, pages 52–56.

[Meigen and Bach, 1999] Meigen, T. and Bach, M. (1999). On the statistical significance of electrophysiological steady-state responses. *Documenta Ophthalmologica*, pages 207–232.

[Merke et al., 2011] Merke, P., Jäger, F., Bosse, S., and Müller, K. (2011). HEVC Anchors and Target Bit Rates for 3DV CfP. In *MPEG Meeting - ISO/IEC JTC1/SC29/WG11, Doc. MPEG11/M21217*, Turin, Italy.

[Mittal et al., 2012] Mittal, A., Moorthy, A. K., and Bovik, A. C. (2012). No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708.

[Mittal et al., 2013] Mittal, A., Soundararajan, R., and Bovik, A. C. (2013). Making a 'completely blind' image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212.

[Montavon et al., 2017] Montavon, G., Lapuschkin, S., Binder, A., Samek, W., and Müller, K. R. (2017). Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognition*, 65(May 2016):211–222.

[Montavon et al., 2018] Montavon, G., Samek, W., and Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15.

[Moon and Lee, 2015] Moon, S.-E. E. and Lee, J.-S. S. (2015). Perceptual experience analysis for tone-mapped HDR videos based on EEG and peripheral physiological signals. *IEEE Transactions on Autonomous Mental Development*, 7(3):236–247.

[Moorthy and Bovik, 2011] Moorthy, A. K. and Bovik, A. C. (2011). Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE Transactions on Image Processing*, 20(12):3350–3364.

[Müller et al., 2013] Müller, K., Schwarz, H., Marpe, D., Bartnik, C., Bosse, S., Brust, H., Hinz, T., Lakshman, H., Merkle, P., Rhee, F. H., Tech, G., Winken, M., and Wiegand, T. (2013). 3D high-efficiency video coding for multi-view video and depth data. *IEEE Transactions on Image Processing*, 22(9):3366–3378.

[Müller et al., 2003] Müller, K.-R., Anderson, C. W., and Birch, G. E. (2003). Linear and nonlinear methods for brain-computer interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 11(2):165–169.

[Müller et al., 2008] Müller, K.-R., Tangermann, M., Dornhege, G., Krauledat, M., Curio, G., and Blankertz, B. (2008). Machine learning for real-time single-trial EEG-analysis: From brain-computer interfacing to mental state monitoring. *Journal of Neuroscience Methods*, 167(1):82–90.

[Müller-Putz et al., 2006] Müller-Putz, G. R., Scherer, R., Neuper, C., and Pfurtscheller, G. (2006). Steady-state somatosensory evoked potentials: Suitable brain signals for brain-computer interfaces? *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 14(1):30–37.

[Mustafa et al., 2012a] Mustafa, M., Guthe, S., and Magnor, M. (2012a). Single-trial EEG classification of artifacts in videos. *ACM Transactions on Applied Perception*, 9(3):1–15.

[Mustafa et al., 2012b] Mustafa, M., Lindemann, L., and Magnor, M. (2012b). EEG analysis of implicit human visual perception. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 513–516. ACM.

[Nair and Hinton, 2010] Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning*, pages 807–814.

[Nikulin et al., 2011] Nikulin, V. V., Nolte, G., and Curio, G. (2011). A novel method for reliable and fast extraction of neuronal EEG/MEG oscillations on the basis of spatio-spectral decomposition. *NeuroImage*, 55(4):1528–1535.

[Norcia et al., 2014] Norcia, A., Ales, J., Cooper, E., and Wiegand, T. (2014). Measuring perceptual differences between compressed and uncompressed video sequences using the swept-parameter visual evoked potential. *Journal of Vision*, 14(10):649–649.

[Norcia et al., 2015] Norcia, A. M., Appelbaum, L. G., Ales, J. M., Cottereau, B. R., and Rossion, B. (2015). The steady-state visual evoked potential in vision research: A review. *Journal of vision*, 15(6):1–46.

[Nunez and Srinivasan, 2006] Nunez, P. L. and Srinivasan, R. (2006). *Electric fields of the brain: the neurophysics of EEG*. Oxford University Press, USA.

[Ojala et al., 2002] Ojala, T., Viertola, J., Huovinen, S., Vision, M., and Unit, M. P. (2002). Outex - New framework for empirical evaluation of texture analysis algorithms. *Pattern Recognition*, 1:701–706.

[Ortiz-Jaramillo et al., 2015] Ortiz-Jaramillo, B., Niño-Castañeda, J., Platiša, L., and Philips, W. (2015). Content-aware video quality assessment: predicting human perception of quality using peak signal to noise ratio and spatial/temporal activity. In *Image Processing: Algorithms and Systems XIII*, volume 9399, page 939917.

[Palmer, 1999] Palmer, S. E. (1999). *Vision science: Photons to phenomenology*. MIT press.

[Parra et al., 2005] Parra, L. C., Spence, C. D., Gerson, A. D., and Sajda, P. (2005). Recipes for the linear analysis of EEG. *NeuroImage*, 28(2):326–341.

[Pei and Chen, 2015] Pei, S.-C. and Chen, L.-H. (2015). Image quality assessment using human visual DOG model fused with random forest. *IEEE Transactions on Image Processing,*, 24(11):3282–3292.

[Pinson et al., 2012] Pinson, M. H., Janowski, L., Pepion, R., Huynh-Thu, Q., Schmidmer, C., Corriveau, P., Younkin, A., Le Callet, P., Barkowsky, M., and Ingram, W. (2012). The influence of subjects and environment on audiovisual subjective tests: An international study. *IEEE Journal on Selected Topics in Signal Processing*, 6(6):640–651.

[Polich, 2009] Polich, J. (2009). Updating P300:An integrative theory of P3a and P3b. *Clin Neurophysiol*, 118(10):2128–2148.

[Polich, 2012] Polich, J. (2012). Neuropsychology of P300. *Oxford handbook of event-related potential components*, pages 159–188.

[Ponomarenko et al., 2013] Ponomarenko, N., Ieremeiev, O., Lukin, V., Egiazarian, K., Jin, L., Astola, J., Vozel, B., Chehdi, K., Carli, M., Battisti, F., and Kuo, C.-C. J. (2013). Color Image Database TID2013: Peculiarities and preliminary results. *4th European Workshop on Visual Information Processing (EUVIP)*, pages 106–111.

[Ponomarenko et al., 2009] Ponomarenko, N., Lukin, V., Zelensky, A., Egiazarian, K., Astola, J., Carli, M., and Battisti, F. (2009). TID2008-A database for evaluation of full-reference visual quality assessment metrics. *Advances of Modern Radioelectronics*, 10(January 2016):30–45.

[Porbadnigk et al., 2010] Porbadnigk, A. K., Antons, J.-N., Blankertz, B., Treder, M. S., Schleicher, R., Möller, S., and Curio, G. (2010). Using ERPs for assessing the (sub) conscious

perception of noise. In *Annual International Conference of the IEEE on Engineering in Medicine and Biology Society (EMBC)*, volume 2010, pages 2690–2693.

[Porbadnigk et al., 2015] Porbadnigk, A. K., Görnitz, N., Sannelli, C., Binder, A., Braun, M., Kloft, M., and Müller, K.-R. (2015). Extracting latent brain states - Towards true labels in cognitive neuroscience experiments. *NeuroImage*, 120:225–253.

[Porbadnigk et al., 2011] Porbadnigk, A. K., Scholler, S., Blankertz, B., Ritz, A., Born, M., Scholl, R., Müller, K.-R., Curio, G., and Treder, M. S. (2011). Revealing the neural response to imperceptible peripheral flicker with machine learning. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, volume 2011, pages 3692–3695.

[Porbadnigk et al., 2013] Porbadnigk, A. K., Treder, M. S., Blankertz, B., Antons, J. N., Schleicher, R., Möller, S., Curio, G., and Müller, K.-R. (2013). Single-trial analysis of the neural correlates of speech quality perception. *Journal of Neural Engineering*, 10(5):056003.

[Prechelt, 2012] Prechelt, L. (2012). Early stopping – but when? In *Neural Networks: Tricks of the Trade*, pages 53–67. Springer.

[Regan, 1966] Regan, D. (1966). Some characteristics of average steady-state and transient responses evoked by modulated light. *Clinical Neurophysiology*, 20(3):238–248.

[Regan, 1977] Regan, D. (1977). Steady-state evoked potentials. *Journal of the Optical Society of America*, 67(11):1475–1489.

[Regan, 1989] Regan, D. (1989). *Human brain electrophysiology: evoked potentials and evoked magnetic fields in science and medicine*. Elsevier.

[Regan et al., 1995] Regan, M. P., He, P., and Regan, D. (1995). An audio-visual convergence area in the human brain. *Experimental brain research. Experimentelle Hirnforschung. Experimentation cerebrale*, 106(3):485–487.

[Regan and Regan, 1988] Regan, M. P. and Regan, D. (1988). A frequency domain technique for characterizing nonlinearities in biological systems. *Journal of Theoretical Biology*, 133(3):293–317.

[Reisenhofer et al., 2018] Reisenhofer, R., Bosse, S., Kutyniok, G., and Wiegand, T. (2018). A Haar wavelet-based perceptual similarity index for image quality assessment. *Signal Processing: Image Communication*, 61:33–43.

[Riskey, 1986] Riskey, D. R. (1986). Use and Abuses of Category Scales in Sensory Measurement. *Journal of Sensory Studies*, 1(3-4):217–236.

[Ruderman, 1994] Ruderman, D. (1994). The statistics of natural images. *Network: Computation in Neural Systems*, 5:517–548.

[Russakovsky et al., 2015] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A., and Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252.

[Saad et al., 2012] Saad, M. A., Bovik, A. C., and Charrier, C. (2012). Blind image quality assessment: A natural scene statistics approach in the DCT domain. *IEEE Transactions on Image Processing*, 21(8):3339–3352.

[Samek et al., 2014] Samek, W., Kawanabe, M., and Müller, K.-R. (2014). Divergence-based framework for common spatial patterns algorithms. *IEEE Reviews in Biomedical Engineering*, 7:50–72.

[Scholler et al., 2012] Scholler, S., Bosse, S., Treder, M. S., Blankertz, B., Curio, G., Müller, K.-R., and Wiegand, T. (2012). Toward a direct measure of video quality perception using EEG. *IEEE Transactions on Image Processing*, 21(5):2619–29.

[Schwarz et al., 2011a] Schwarz, H., Bartnik, C., Bosse, S., Brust, H., Hinz, T., Lakshman, H., Marpe, D., Merke, P., Müller, K., Rhee, H., Tech, G., Winken, M., and Wiegand, T. (2011a). Description of 3D video coding technology proposal by Fraunhofer HHI (HEVC compatible configuration B). In *MPEG Meeting ISO/IEC JTC1/SC29/WG11, Doc. MPEG11/M22571*, Geneva, Switzerland.

[Schwarz et al., 2011b] Schwarz, H., Bartnik, C., Bosse, S., Brust, H., Hinz, T., Lakshman, H., Marpe, D., Merke, P., Müller, K., Rhee, H., Tech, G., Winken, M., and Wiegand, T. (2011b). Description of 3D video coding technology proposal by Fraunhofer HHI (HEVC compatible configuration A),. In *MPEG Meeting ISO/IEC JTC1/SC29/WG11, Doc. MPEG11/M22570*, Geneva, Switzerland.

[Schwarz et al., 2011c] Schwarz, H., Bartnik, C., Bosse, S., Brust, H., Hinz, T., Lakshman, H., Marpe, D., Merke, P., Müller, K., Rhee, H., Tech, G., Winken, M., and Wiegand, T. (2011c). Description of 3D video coding technology proposal by Fraunhofer HHI (MVC compatible). In *MPEG Meeting ISO/IEC JTC1/SC29/WG11, Doc. MPEG11/M22569*, Geneva, Switzerland.

[Schwarz et al., 2012a] Schwarz, H., Bartnik, C., Bosse, S., Brust, H., Hinz, T., Lakshman, H., Marpe, D., Merkle, P., Müller, K., Rhee, H., Tech, G., Winken, M., and Wiegand, T. (2012a). 3D video coding using advanced prediction, depth modeling, and encoder control methods. In *Picture Coding Symposium (PCS)*, pages 1–4.

[Schwarz et al., 2012b] Schwarz, H., Bartnik, C., Bosse, S., Brust, H., Hinz, T., Lakshman, H., Merkle, P., Müller, K., Rhee, H., Tech, G., Winken, M., Marpe, D., and Wiegand, T. (2012b). Extension of High Efficiency Video Coding (HEVC) for multiview video and depth data. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pages 205–208.

[Shahbazi et al., 2017] Shahbazi, F., Bosse, S., Nolte, G., Wiegand, T., and Samek, W. (2017). Quality assessment of 3D visualizations with vertical disparity: An ERP approach. In

Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, pages 4391–94.

[Sharbrough et al., 1991] Sharbrough, F., Chatrian, G. E., Lesser, R., Luders, H., Nuwer, M., and Picton, T. (1991). American Electroencephalographic Society guidelines for standard electrode position nomenclature. *Clinical Neurophysiology*, 8:200–202.

[Sheikh et al., 2006] Sheikh, H. R., Sabir, M. F., and Bovik, A. C. (2006). A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on image processing*, 15(11):3440–51.

[Siekmann et al., 2010] Siekmann, M., Bosse, S., Schwarz, H., and Wiegand, T. (2010). Separable Wiener filter based adaptive in-loop filter for video coding. In *Picture Coding Symposium (PCS)*, pages 70–73.

[Simonyan and Zisserman, 2014] Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, pages 1–10.

[Slater, 2004] Slater, M. (2004). How colorful was your day? Why questionnaires cannot assess presence in virtual environments. *Presence: Teleoperators and Virtual Environments*, 13(4):484–493.

[Soundararajan and Bovik, 2012] Soundararajan, R. and Bovik, A. C. (2012). RRED indices: Reduced reference entropic differencing for image quality assessment. *IEEE Transactions on Image Processing*, 21(2):517–526.

[Srivastava et al., 2014] Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout : A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research (JMLR)*, 15:1929–1958.

[Statistica, 2017] Statistica (2017). https://www.statista.com/chart/10913/number-of-photos-taken-worldwide/.

[Stefanoski et al., 2011] Stefanoski, N., Espinosa, P., Wang, O., Lang, M., Smolic, A., Bosse, S., Farre, M., Müller, K., Schwarz, H., Winken, M., and Wiegand, T. (2011). Description of 3D video coding technology proposal by Disney research Zurich and Fraunhofer HHI,. In *MPEG Meeting - ISO/IEC JTC1/SC29/WG11, Doc. MPEG11/M22668*, Geneva, Switzerland.

[Stevens, 1946] Stevens, S. (1946). *On the theory of scales of measurement*, volume 103. Bobbs-Merrill, College Division.

[Strasburger et al., 1988] Strasburger, H., Scheidler, W., and Rentschler, I. (1988). Amplitude and phase characteristics of the steady-state visual evoked potential. *Applied Optics*, 27(6):1069–1088.

[Streijl et al., 2016] Streijl, R. C., Winkler, S., and Hands, D. S. (2016). Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives. *Multimedia Systems*, 22(2):213–227.

[Suk et al., 2014] Suk, H. I., Fazli, S., Mehnert, J., Müller, K.-R., and Lee, S. W. (2014). Predicting BCI subject performance using probabilistic spatio-temporal filters. *PLoS ONE*, 9(2):e87056.

[Sullivan et al., 2012] Sullivan, G. J., Ohm, J. R., Han, W.-J. J., and Wiegand, T. (2012). Overview of the High Efficiency Video Coding (HEVC) Standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 22(12):1649–1668.

[Tukey, 1977] Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley.

[Urvoy et al., 2013] Urvoy, M., Barkowsky, M., and Le Callet, P. (2013). How visual fatigue and discomfort impact 3D-TV quality of experience: A comprehensive review of technological, psychophysical, and psychological factors. *Annales des Telecommunications/Annals of Telecommunications*, 68(11-12):641–655.

[VQEG, 2004] VQEG (2004). Objective perceptual assessment of video quality: full reference television.

[Wandell, 1995] Wandell, B. A. (1995). *Foundations of Vision*. Sinauer Associates.

[Wang, 2011] Wang, Z. (2011). Applications of objective image quality assessment methods. *IEEE Signal Processing Magazine*, 28(6):137–142.

[Wang and Bovik, 2009] Wang, Z. and Bovik, A. C. (2009). Mean squared error: Love it or leave it? A new look at signal fidelity measures. *IEEE Signal Processing Magazine*, 26(1):98–117.

[Wang et al., 2004] Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612.

[Wang and Li, 2011] Wang, Z. and Li, Q. (2011). Information content weighting for perceptual image quality assessment. *IEEE Transactions on Image Processing*, 20(5):1185–1198.

[Wang and Shang, 2006] Wang, Z. and Shang, X. (2006). Spatial pooling strategies for perceptual image quality assessment. In *Proc. of the IEEE International Conference on Image Processing (ICIP)*, pages 2945–2948.

[Wang et al., 2003] Wang, Z., Simoncelli, E. P., and Bovik, A. C. (2003). Multiscale structural similarity for image quality assessment. In *IEEE Asilomar Conference on Signals, Systems and Computers*, pages 1398–1402.

[Ward, 2003] Ward, L. M. (2003). Synchronous neural oscillations and cognitive processes. *Trends in Cognitive Sciences*, 7(12):553–559.

[Watson et al., 1997] Watson, A., Borthwick, R., and Taylor, M. (1997). Image quality and entropy masking. In *SPIE Proceedings*, volume 3016, pages 1–11.

[Wenzel et al., 2016] Wenzel, M. A., Schultze-Kraft, R., Meinecke, F. C., Cardinaux, F., Kemp, T., Müller, K.-R., Curio, G., and Blankertz, B. (2016). EEG-based usability assessment of 3D shutter glasses. *Journal of Neural Engineering*, 13(1):016003.

[Wiegand and Schwarz, 2016] Wiegand, T. and Schwarz, H. (2016). Video Coding: Part II of Fundamentals of Source and Video Coding. *Foundations and Trends® in Signal Processing*, 10(1–3):1–346.

[Wiegand et al., 2003] Wiegand, T., Sullivan, G. J., Bjontegaard, G., and Luthra, A. (2003). Overview of the H. 264/AVC video coding standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 13(7):560–576.

[Winken et al., 2010] Winken, M., Bosse, S., Benjamin, B., Helle, P., Hinz, T., Kirchoffer, H., Lakshman, H., Marpe, D., Oudin, S., Preiss, M., Schwarz, H., Siekmann, M., Sühring, K., and Wiegand, T. (2010). Description of video coding technology proposal by Fraunhofer HHI. In *Joint Collaborative Team on Video Coding, JCTVC-A116*, Dresden, Germany.

[Winken et al., 2011] Winken, M., Marpe, D., Schwarz, H., Wiegand, T., Boße, S., Bross, B., Helle, P., Hinz, T., Kirchhoffer, H., Lakshman, H., Nguyen, T., Oudin, S., Siekmann, M., and Sühring, K. (2011). Highly efficient video coding based on quadtree structures, improved motion compensation, and probability interval partitioning entropy coding. In *Proc. of the ITG Conference on Electronic Media Technology, CEMT*.

[Winkler, 2014] Winkler, S. (2014). Does inter-subject variability depend on test material? In *Sixth International Workshop on Quality of Multimedia Experience (QoMEX)*, pages 37–38.

[WMA General Assembly, 2013] WMA General Assembly (2013). Declaration of Helsinki. Ethical principles for medical research involving human subjects.

[Won et al., 2015] Won, D.-O., Hwang, H.-J., Dähne, S., Müller, K.-R., and Lee, S.-W. (2015). Effect of higher frequency on the classification of steady-state visual evoked potentials. *Journal of Neural Engineering*, 13(1):016014.

[Xue et al., 2014] Xue, W., Zhang, L., Mou, X., and Bovik, A. C. (2014). Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. *IEEE Transactions on Image Processing*, 23(2):668–695.

[Yarbus, 1967] Yarbus, A. L. (1967). Eye movements during perception of complex objects. In *Eye movements and vision*, pages 171–211. Springer.

[Ye et al., 2012] Ye, P., Kumar, J., Kang, L., and Doermann, D. (2012). Unsupervised feature learning framework for no-reference image quality assessment. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1098–1105.

[You et al., 2010] You, J., Reiter, U., Hannuksela, M. M., Gabbouj, M., and Perkis, A. (2010). Perceptual-based quality assessment for audiovisual services: A survey. *Signal Processing: Image Communication*, 25(7):482–501.

[Zhang et al., 2014a] Zhang, L., Gu, Z., Liu, X., Li, H., and Lu, J. (2014a). Training quality-aware filters for no-reference image quality assessment. *IEEE MultiMedia*, 21(4):67–75.

[Zhang et al., 2014b] Zhang, L., Shen, Y., and Li, H. (2014b). VSI: A visual saliency-induced index for perceptual image quality assessment. *IEEE Transactions on Image Processing*, 23(10):4270–4281.

[Zhang et al., 2011] Zhang, L., Zhang, L., Mou, X., Zhang, D., and Mou, X. (2011). FSIM: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, 20(8):2378–2386.

[Zhang et al., 2015] Zhang, P., Zhou, W., Wu, L., and Li, H. (2015). SOM: Semantic obviousness metric for image quality assessment. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2394–2402.

[Zhang et al., 2016] Zhang, W., Borji, A., Wang, Z., Le Callet, P., and Liu, H. (2016). The application of visual saliency models in objective image quality assessment: A statistical evaluation. *IEEE Transactions on Neural Networks and Learning Systems*, 27(6):1266–1278.