Chenwei Deng · Lin Ma
Weisi Lin · King Ngi Ngan *Editors*

Visual Signal Quality Assessment

Quality of Experience (QoE)

Springer

Visual Signal Quality Assessment

Chenwei Deng • Lin Ma • Weisi Lin
King Ngi Ngan
Editors

Visual Signal Quality Assessment

Quality of Experience (QoE)

Springer

Editors

Chenwei Deng
School of Information and Electronics
Beijing Institute of Technology
Beijing, China

Weisi Lin
School of Computer Engineering
Nanyang Technological University
Singapore

Lin Ma
Huawei Noah's Ark Lab
Shatin, Hong Kong SAR

King Ngi Ngan
Department of Electronic Engineering
The Chinese University of Hong Kong
Shatin, Hong Kong SAR

ISBN 978-3-319-10367-9 ISBN 978-3-319-10368-6 (eBook)
DOI 10.1007/978-3-319-10368-6
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014953787

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

With rapidly advancing computer and network technologies, various visual signals (including image, video, graphics, animation, etc.) are produced, and visual quality of experience (QoE) plays an important role in multimedia applications and services. Visual QoE evaluation is essential not only on its own for testing, optimizing, and inspecting related algorithms, systems and services, but also for shaping and decision-making for virtually all multimedia signal processing and transmission algorithms. It is not an exaggeration to say that how visual signal quality is evaluated shapes the making of almost all multimedia processing algorithms and systems, since the ultimate goal of processing is to achieve the highest possible perceived quality.

During the past two decades, the research field of visual quality assessment has experienced significant growth and great progress. With the rapid development of the sensing and imaging devices, newly emerged visual signals are presented to human viewers, such as stereoscopic/3D image/video, high dynamic range (HDR) image/video, retargeted image/video, graphics, medical image, and so on. Meanwhile, recent psychophysical and neurological findings enable us to more clearly understand the human visual system. There is a considerable need for books like this one, which attempts to provide a comprehensive reviewing of recent progresses of visual signal quality assessment and shape the future research directions.

The objective of this book is to firstly present the latest achievements in quality assessment of visual signals. It reviews the current status, new trends, and challenges of quality assessment for traditional visual signals. More attentions are devoted to the newly emerged visual signals for better QoE. With the systematic and up-to-date review of the quality assessment of emerged visual signals, new trends of developing quality assessment methods for the specific visual signals are discussed and believed to be helpful to the researchers and readers of this book.

This book provides readers a comprehensive coverage of the latest trends/advances in visual signal quality assessment. The principal audience of this book will be mainly researchers and engineers as well as graduate students working on various disciplines related to visual signals, such as imaging, displaying, processing, storage, transmission, etc. The discussed contents in this book are expected to not only inspire newly research trends and directions for QoE but also benefit the development of multimedia products, applications, and services.

Chapter 1 introduces the current status, challenges, and new trends of visual quality assessment. Entropy and rate-distortion-based quality assessment methods are first reviewed, and then perception-oriented approaches are discussed, including pixel, feature, and model-based ones, etc. Perception distortion measures for rate-perceptual-distortion optimization (RpDO) in visual quality regulated services are finally presented.

Chapter 2 presents a detailed review of subjective image quality assessment. The previous research of subjective assessment targeted at measuring visual impairments induced by limited spatial, temporal resolutions in displays, bandwidth and storage constraints, etc. However, elements such as visual semantics, user personality, preferences and intent, social and environmental context of media fruition also have great impact on the perceived experience. In order to adapt the traditional visual quality gauging metrics to QoE, a few models have been proposed throughout the last decade, and a significant improvement has been achieved.

Chapter 3 provides a survey and tutorial on objective quality assessment. The public image databases are first introduced, including those popular ones (e.g., TID2008, LIVE) and new quality databases (e.g., TID2013, LIVEMD, CID2013). Most of the proposed objective metrics are tested on those image databases. As for objective assessment, numerous approaches have been developed including full-reference, reduced-reference, and no-reference ones. With the development of multimedia technology, some emerging directions in quality assessment are also in demand for specific applications, such as multiply distorted quality assessment, mobile quality assessment, etc.

In particular, Chapter 4 addresses the issue of understanding and modeling the perceptual mechanism of QoE of mobile videos. For quality perception of mobile videos, compression and transmission artifacts, and the video scalability, specifically the spatial, temporal, and quality scalability as the most effective factors are discussed. Several quality metrics for mobile videos are introduced by considering the general purposes, video scalability, and standardization. Finally, the characteristics of the public mobile video quality databases are introduced and summarized.

Chapter 5 focuses on the quality evaluation of HDR images. HDR is opposite to low dynamic range (LDR). The major difference between them is that HDR has much more bits than LDR to represent the dynamic range of visual signals in the real world. On the other hand, the existing display devices are not available for HDR contents. Tone mapping operators (TMOs) can efficiently solve the above-mentioned issues, but also lead to the loss of visual details affecting the perceived quality of HDR contents. A detailed discussion on the relationship between tone mapping and image quality has been presented from perceptual visual quality, visual attention, and naturalness aspects.

Chapter 6 reviews recent progresses of quality assessment for medical imaging systems and medical images. The concepts and definitions of traditional medical image quality metrics are introduced. A mutual information-based quality metric for medical imaging systems is introduced. Two clinical applications related to quality assessment for medical images are addressed. One application deals with the improvement of image quality in mammography, while the second one addresses the effect of radiation dose reduction on image quality in digital radiography.

Chapter 7 discusses visual quality assessment on stereoscopic images/videos, including the challenges and difficulties, such as visual discomfort, binocular vision, and extra dimensionality. Recently, quality metrics for stereoscopic images/videos not only utilize the information from images/frames but also consider the obtained depth or computed depth/disparity information. Evaluation of the 2D quality metrics and the 3D quality metrics considering depth/disparity information on the publicly available stereoscopic databases confirms the necessity of utilizing and incorporating accessible 3D information.

Chapter 8 addresses the quality assessment of retargeted images from both subjective and objective perspectives. Subjective evaluation process of retargeted images is firstly introduced. Specifically, two publicly available image retargeting quality databases are introduced. The representative objective quality assessment algorithms are evaluated and compared on the two built subjective databases, which provide helpful insights for image retargeting quality assessment. Future trends are discussed to handle the challenges during the objective quality metric construction for retargeted images.

Chapter 9 presents some introductions of computer graphics quality assessment from subjective and objective aspects. Image rendering evaluation is one of the important applications for graphics, and a few models have been proposed including visual model-based ones, data-driven ones, etc. Apart from image rendering, numerous assessment metrics have been developed for evaluating the quality of 3D models.

Chapter 10 gives the conclusions and perspectives of each chapter. It outlines the main contents from multiple aspects of quality assessment and reviews the main contributions of those existing works. Then perspectives related to QoE are presented to readers for further investigations.

This book collects related but distinctive works from 10 active research groups in different part of the world, in addressing QoE challenges from different perspectives. We hope that this gives a comprehensive overview in the area and inspires new thinking and ideas as the next steps of research and development.

Beijing, China	Chenwei Deng
Shatin, Hong Kong SAR	Lin Ma
Singapore	Weisi Lin
Shatin, Hong Kong SAR	King Ngi Ngan

Editor Bios

Chenwei Deng received his Ph.D. degree in signal and information processing from Beijing Institute of Technology, Beijing, China, in 2009. He is currently an Associate Professor at the School of Information and Electronics, Beijing Institute of Technology. Prior to this, he was a postdoctoral Research Fellow in the School of Computer Engineering, Nanyang Technological University, Singapore. He has authored or co-authored over 50 technical papers in refereed international journals and conferences. He was awarded the titles of Beijing Excellent Talent and Excellent Young Scholar of Beijing Institute of Technology in 2013. His current research interests include multimedia coding, quality assessment, perceptual modeling, and pattern recognition.

Weisi Lin received his Ph.D. degree from King's College, London University, London, UK. He is currently an Associate Professor and Associate Chair (Graduate Studies) with the School of Computer Engineering, Nanyang Technological University, Singapore. He has published over 270 refereed papers at international journals and conferences. His current research interests include image processing, visual quality evaluation, and perception-inspired signal modeling. He is a senior member of IEEE and a fellow of IET.

Lin Ma received his Ph.D. degree in Department of Electronic Engineering from the Chinese University of Hong Kong in 2013. He received his B.E. and M.E. degrees from Harbin Institute of Technology, Harbin, China, in 2006 and 2008, respectively, both in computer science. He is currently a Researcher in Huawei Noah's Ark Lab, Hong Kong. He got the best paper award in Pacific-Rim Conference on Multimedia 2008. He was awarded the Microsoft Research Asia fellowship in 2011. He was a finalist to HKIS young scientist award in engineering science in 2012. His research interests include visual quality assessment, super-resolution, restoration, and compression.

King Ngi Ngan received his Ph.D. degree in electrical engineering from Loughborough University, Loughborough, UK. He is currently a Chair Professor with the Department of Electronic Engineering, Chinese University of Hong Kong, Shatin, Hong Kong. He holds honorary and visiting professorships with numerous universities in China, Australia, and South East Asia. He has published extensively, including three authored books, six edited volumes, over 300 refereed technical papers, and has edited nine special issues in journals. He holds ten patents in image or video coding and communications. He is a fellow of IEEE, IET, and IEAust.

Contents

Chapter 1
Introduction: State of the Play and Challenges of Visual Quality Assessment

Hong Ren Wu

Visual communications, broadcasting and entertainment set out to overcome time, distance or other barriers between people and/or places, which hinder face-to-face contact with one another or between human subject(s) and the environment [25]. Quality of visual signals or pictures[1] perceived by human observers as compared with what they would experience should they be able to be present at the natural scene has always been a critical issue, so has been the measurement of the signal quality throughout a process chain of acquisition/reproduction, encoding, transmission or storage, decoding, and visualisation/display associated with a designated application or service regardless whether visual signals are in analogue or digital form [29, 31, 100, 129]. Taking advantage of digital communications, the first digital coding system for television (TV) signals using pulse code modulation (PCM) [82] as reported in 1949 revealed, among other issues, that the digital TV signal yielded approximately 1,250 times the bitrate of telephone signals,[2] and using 5 (instead of 8) bits per sample achieved visual picture quality which was comparable to that of the original analogue TV signal, achieving a 1.6 compression gain on the grounds of perceptual picture quality [24]. The sheer volume of digital TV signal necessitated research and development in signal compression theories [18, 27, 38, 63, 64, 81, 83, 118] and technology as well as in high speed and bandwidth communication and storage technologies [26]. Digital

[1] Visual signals or pictures refer to images, video, image sequences or motion pictures [118].

[2] In [24], 10 MHz sampling rate was used for PCM coding of a 5 MHz analogue TV signal with 8 bits per sample, compared with 8 kHz sampling rate and 8 bits per sample for voice using telephone at the time.

H.R. Wu (✉)
School of Electrical and Computer Engineering, RMIT University,
Swanston Street, Melbourne, Victoria 3000, Australia
e-mail: henry.wu@rmit.edu.au

C. Deng et al. (eds.), *Visual Signal Quality Assessment*,
DOI 10.1007/978-3-319-10368-6_1

visual signals compressed using various coding techniques [14, 39, 95] exhibited coding distortions which differed from those known to be associated with analogue visual signals and, therefore, required provision of both subjective and objective distortion or quality measures which quantitatively assess and evaluate the visual picture quality for the purposes of system or service evaluation and optimization [4, 5, 34, 41, 46, 49, 55, 85, 118].

Digital visual signal compression theory, technology, and standardizations have come of age. Visual communications, broadcasting, entertainment and recreation video and photography have been completely transformed in the past 20 years from analogue based devices, products, systems, and services to ever increasingly diverse forms of digital counterparts, exemplified most noticeably via popular products or events such as film based cameras replaced by digital cameras, tape based analogue video camcorders by digital storage[3] based digital video cameras, and free-to-air analogue television (TV) broadcasting having been switched to digital-only TV services in many countries.[4] Heralding a transition from technology driven services to user-centric (visual or perceived) quality assured services [118] comes an increasing emphasis on visual quality of picture (QoP) assessments and measures [7, 10, 44, 48, 59, 108, 119], and quality of experience (QoE)[5] [35, 53] as compared with quality of service (QoS)[6] [36] in the aforementioned applications and services. To understand the importance, imperative, and relevance of this transition, a number of fundamental issues deserve clarification in order to put the current discussions and activities into perspective and context, including relationship between picture quality assessment and coding designs, how to measure effectiveness of visual signal compression performance, different scales used for visual quality assessment and their intended applications, picture distortion or quality ratings for rate-perceptual-distortion (R$\mathfrak{p}$D) optimization [38].

1.1 Quality Assessments Based on Entropy and Rate-Distortion Theories

Entropy and rate-distortion theories have been design principles for visual signal coding and forged the inseparable nexus between visual signal coding design and

[3] Digital storage media commonly used by digital video cameras currently include memory stick, memory card, and flash memory.

[4] For example, Australia switched to digital-only TV broadcasting on 10 December 2013 as per Australian Government announcement via “Australia’s Ready for Digital TV.”

[5] QoE as defined by International Telecommunication Union Study Group (ITU SG) 12 is application or service specific and influenced by user expectations and context [35], and therefore necessitates assessments of perceived service quality and/or utility (or usefulness) of the service [76].

[6] QoS as defined by ITU SG 12 is the totality of characteristics of a telecommunications service that bear on its ability to satisfy stated and implied needs of the user of the service [36], e.g., error rates, bandwidth, throughput, transmission delay, availability, jitter, and so on [91].

quality assessment. To further advance the theory and practice in visual signal coding and transmission, three issues are examined in this section based on entropy and rate-distortion theories to clarify the "effectiveness" criterion for picture coder design, to raise questions regarding quality scales which are currently used in subjective assessment to collect the "ground truth" quality or distortion data as perceived by human viewers, and to place a focus on perception based principles for visual signal coding and performance assessment.

1.1.1 Quality Assessment for Bitrate- and Quality-Driven Coding Designs

It is widely known and acknowledged that visual signal coding/compression and transmission have been developed based on three fundamental theories [118], i.e., Nyquist–Shannon sampling theory which governs the required sample rate for a faithful digital representation of an underpinning analogue counterpart or another digital waveform at a reduced rate [82], Shannon's entropy theory which defines the lower bound for information lossless compression [81], and Shannon's rate-distortion (R-D) theory for information lossy compression design optimization [3, 81, 83]. With regard to entropy and R-D theories, two observations have been made known since the very beginning of visual signal coding and compression research and development [117]. First, when taking account of visual picture quality as perceived by human visual system (HVS), significantly higher compression ratio than what is achievable by information lossless (or entropy) coding is possible where the compressed images are visually comparable to [24] or indistinguishable from their originals [116, 118]. Second, constant bitrate and constant distortion (or quality) coder designs can be formulated when coding distortion is either inevitable or impractical due to various constraints [3,39,81,83]. These observations underscore that both perceptually lossless and perceptually lossy coding and transmission designs and evaluation have an inseparable nexus with human perception-based quality assessment and measures, including both psychophysical/subjective [15,33,37,75,120] and quantitative [4,5,7,10,34,41,44,46,48,49,55,59,85,108,119] methods or approaches. Acknowledging this nexus will allow better understanding of how to determine whether a coding system is effective or otherwise and how coding performance evaluations ought to be conducted. Based on the R-D theory, for a constant bitrate coder to lay a claim to its effectiveness in compression performance, it has to hold a fixed bitrate and then to maximize picture quality, whereas for a constant quality coder to be effective, it must, first and foremost, be able to hold a given picture quality and, then, to do so at the lowest possible bitrate.

The research efforts and developments in quality performance assessment to date have been mainly focused on and reasonably successfully addressed two significant issues, i.e., why the time-honored distortion (or quality) measure such as mean squared error (MSE) (or peak signal-to-noise ratio, PSNR for short) does not always qualify or suitable for picture coding visual quality performance assessments [5, 23, 55, 100, 119], and formulation and development of human visual perception-based quantitative distortion or quality measures [4, 5, 7, 10, 29, 34, 41, 44, 48, 49, 55, 59, 85, 108, 119, 123] which are able to grade picture quality consistently with respect to subjective test data collected following the current standard practices [15, 33, 37]. These achievements notwithstanding, coding performance evaluations including those where perception-based measures are considered have been confined, so it appears, by a mindset based on bitrate-driven design approach. The winner was usually declared if for the same given bitrate, it demonstrated a superior picture quality measured by either the time honored measures such as the PSNR or the MSE [63] or quantitative perceptual distortion/quality measures [4, 5, 7, 10, 34, 41, 44, 48, 49, 55, 59, 85, 108, 119] or subjective assessments [15, 33, 37]. When the visual quality was considered, the focus was still on which coder achieved significant bitrate savings at a comparable visual quality, instead of whether it was able to deliver designated picture quality levels discernible to human viewers at lower rates. In other words, performance evaluations based on the current mindset are able to assess effectiveness of bitrate-driven coder design, nonetheless do not address the key issue regarding effectiveness of (visual) quality-driven coder design, i.e., the question whether the distortion (or quality) measure is able to predict discernible levels by human visual perception in terms of, e.g., just-not-noticeable-difference (JNND), just-noticeable-difference level 1 (JND_1), JND level 2 (JND_2), etc., consistently [118].

Take the performance evaluation of H.265/HEVC (high efficiency video coding) recently reported in [63] for example. The superior effectiveness of H.265/HEVC over all its predecessors as a bitrate-driven coder and coding standard has been amply demonstrated in terms of delivering a designated bitrate at a significantly better quality using either the PSNR or visual quality. If, however, visual quality assured service at a designated quality level is set as the performance criterion, according to the performance shown in Fig. 1.1 [63], there is an up to 3 dB difference by the HEVC MP (main profile) coder for two different test sequences (i.e., *Kimono1* and *Park Scene*) for a given bitrate of 2 Mbps, even if the PSNR is accepted as an appropriate fidelity measure. Alternatively, given a PSNR value, say in this example at 37 dB, it is most likely representing different levels of visual picture quality for these two videos which are coded at about 2 Mbps (megabit per second) and 512 kbps (kilobit per second), respectively. In any event, effectiveness of all coders under this comparative R-D performance study as a quality-driven coder has not been considered, demonstrated, or established. (It is noted that in Fig. 1.1, YUV-PSNR is defined as $PSNR_{YUV} = (6 \cdot PSNR_Y + PSNR_U + PSNR_V)/8$, where $PSNR_Y$, $PSNR_U$, and $PSNR_V$ are each computed as $PSNR = 10 \cdot log_{10}((2^8 - 1)^2/MSE)$ in dB [63].)

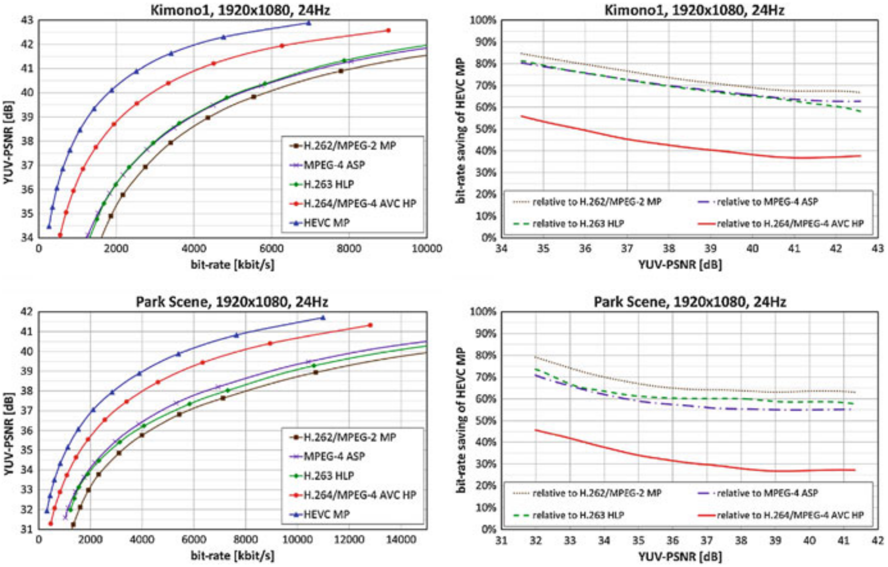


Fig. 1.1 Selected rate-distortion curves and bit-rate savings plots in coding performance comparison for entertainment applications (©IEEE, 2012) [63]

In contrast, effectiveness of a perceptually lossless[7] picture coder as a constant visual quality coder has to be evaluated in terms of whether it is able to maintain a designated visual picture quality at lower bitrate than its competitors. In [116], a JPEG 2000 bit-stream compliant perceptually lossless image coder was compared with a JPEG-LS (information) lossless coder and a JPEG-LS near-lossless coder (with d = 2, i.e., the maximum pixel difference between the compressed and the original images less than or equal to two) in double blind subjective evaluations. While the perceptually lossless coder demonstrated its effectiveness to hold visual distortions at or below JNND level, it achieved a compression gain of 1.48 times on average compared with the JPEG-LS lossless coder for medical images. The JPEG-LS near-lossless (d = 2) coder achieved a comparable bitrate compared with the perceptually lossless coder, but failed to deliver perceptually lossless coding results in the same double blind subjective evaluations [116]. An example is given in [118] to show the effectiveness of perceptually lossless coding which compresses *Shannon* image at 1.370 bpp (bit per pixel) using the aforementioned JPEG 2000 bit-stream compliant perceptually lossless image coder, compared with the abovementioned JPEG-LS lossless coder at 3.179 bpp, a JPEG 2000 lossless coder at 3.196 bpp, and the abovementioned JPEG-LS near-lossless coder at 1.424 bpp for the same image. In this sense, the perception-based approaches are

[7] Perceptually lossless coded visual signals incur no discernable visual difference compared with their originals while they may have undergone irreversible information loss [106, 118].

most likely to provide the means to delivery of effective, efficient compression and transmission strategies for visual signals in terms of *perceptual entropy*[8] or RpD criterion [38].

1.1.2 Scales for Subjective Perceptual Distortion and Quality Measurement

Constant perceptual quality coder design and visual signal transmission services rely on perceptual distortion or quality measures, which are designed to assess levels of quality discernible to human viewers, for RpDO to uphold an agreed or a designated visual quality acceptable to the users at the minimum bitrate. Goodness of these perceptual measures is appraised and validated using subjective test data as the ground truth [7, 9, 10, 34, 43, 45, 49, 84, 85, 125]. Absolute category rating (ACR) has been widely used in subjective picture quality evaluations [15, 33, 37] whose data have been often assumed as the ground truth and used to evaluate or validate perceptual distortion or quality metrics [9, 10, 34, 43, 45, 49, 84, 85, 125]. However, it is not entirely clear whether this ground truth so acquired and claimed using the existing ACR or similar schemes is sufficiently adequate, accurate, or suitable for design of constant quality or quality regulated picture coders and performance evaluations. For example, there is no guarantee that a score of "excellent" in a five-level [33] (or eleven-level [37, 63]) scale for rating overall picture quality, when the actual scores marked by viewers do not (and they rarely do [63]) achieve the full mark, corresponds to the JNND level which can be used to guide perceptually lossless picture coding. Nor a score out of 100 necessarily commits itself to a discernible level of quality or distortion comparable to that perceived by the HVS, which is able to uphold a constant visual quality in perceptual picture coding based on the RpDO.

Furthermore, human perception and judgment in a psychophysical measurement task perform usually better in comparison tasks than casting an absolute rating [120]. To address the issue regarding unreliability and fluctuations associated with subjective test data using absolute rating schemes due to contextual effects [16] and varying experience and expectations of observers,[9] a distortion detection strategy is considered compared with the ACR to ascertain JND levels [118] or VDUs (visual distortion units) [75] or distinguishable utility levels pertaining to a designated application [76] in correspondence to constant picture coding approach and design

[8] Perceptual entropy defines the theoretical lower bound of perceptually lossless visual signal coding in a similar way that entropy does the lower bound of information lossless coding [81].

[9] Prior knowledge plays an important part in subjective rating exercise using ACR which forms a benchmark experience or a point of reference in what constitutes the "best" or "excellent" picture quality as they have seen or experienced, and is also exemplified by the entropy masking effect which is imposed solely by an observer's unfamiliarity with the masker [110].

philosophy. As shown in [75], the relative threshold elevation for a VDU varies from one VDU to the next as a function of spatial frequency and orientation which does not appear to be linear.

1.1.3 QoE in Perceptual-Based Visual Signal Coding

QoP and QoE assessments are not just for their own sakes and they are linked closely to visual signal compression and transmission where R-D theory is applied for product, system, and service quality optimization [5, 55, 68]. The nexus between QoP/QoE and coder design is clearly borne out in the previous section. From R-D optimization perspective [3, 39, 83], it is widely understood that use of raw mathematical distortion measures, such as the MSE, does not guarantee visual superiority since the HVS does not compute the MSE [100, 118]. In RpDO [38] where perceptual distortion or utility measure matters, the setting of rate constraint, R_c, in Fig. 1.2 is redundant from the perceptual distortion controlled coding point of view. For bitrate controlled coder design, the perceptual bitrate constraint, R_{pc}, makes more sense which delivers a picture quality comparable to JND_1. In comparison, R_c is neither sufficient to guarantee a distortion level at $JNND$ nor necessary to achieve, e.g., JND_1 in Fig. 1.2. By the same token, for constant visual quality design, a constant distortion setting at D_c is not effective in holding a designated visual quality appreciable to the HVS since it cannot guarantee JND_2 nor is it necessary to deliver JND_3. As the entropy defines the lower bound of the bitrate required for information lossless picture coding [14, 81], the perceptual entropy [38] sets the minimum bitrate required for perceptually lossless picture coding [62, 89, 116]. Similarly, in UoP (perceptual utility of picture) regulated picture coding in terms of a utility measure [76], utility entropy can be defined as the minimum bitrate to reconstruct a picture as required to achieve complete feature recognition equivalent to the perceptually lossless picture including the original as illustrated in Fig. 1.2.

1.2 Perception-Based Approaches to Picture Quality Assessment

There exist different approaches to visual distortion and quality metric designs based on different HVS or visual weighting models whose parameters are optimized using subjective test data which are collected based on standard ACR schemes [61, 85] in terms of the Spearman rank-order correlation or Kandell rank correlation for prediction monotonicity, the Pearson correlation and the average absolute error or the root mean square error for prediction accuracy, and outlier ratio for prediction consistency [45, 104, 125, 130]. Visual modeling or weighting is

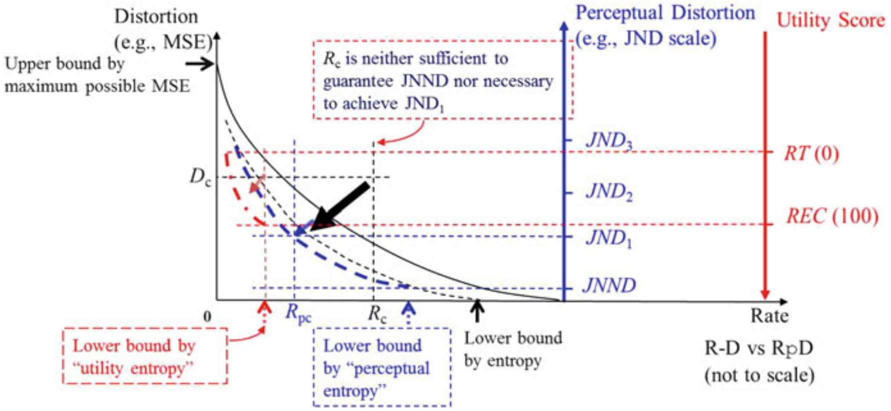


Fig. 1.2 Rate-distortion optimization considering a perceptual distortion measure (PDM) [118] or a utility score [76] for QoE regulated services compared with the MSE. In [76], RT (recognition threshold) is defined as a perceived utility score threshold with a value of zero (0) below which an image deems to be useless and REC (recognition equivalence class) defines a class of images whose perceived utility score with a value of 100 [i.e., REC (100)] is statistically equivalent to that of a perceptually lossless image with respective to and including the reference. *Black solid line* corresponds to R-D curve which can be optimized towards *black dash line*. *Thick blue dash line* represents RpD curve where a PDM is applied, while *thick red dash-dotted line* corresponds to RpD curve where utility rating is used as the measure

usually devised for distortion or quality measures at local level (via windowing or transform/decomposition or feature extraction/segmentation) and/or for overall measure at global level (via pooling) in either pixel, transform or feature domains. Performance advantages and limitations of these metrics may be better appreciated by looking into the very models on which they are constructed in terms of their designated applications, prediction accuracy, and computational complexity. Readers may refer to [29] for further discussions on application scenarios of quality metrics or estimators, and [6,45,85] for image quality metrics, and [10,49] for video quality metrics (VQMs) performance comparisons.

1.2.1 MSE/PSNR with Visual Weighting

Applying visual weighting to the time-honored distortion measure, MSE, was explored in early days of digital picture coding for perception-based quantizer design [46] to address the discrepancy between the raw pixel value differences and what transpired on a video monitor and was perceived by human observers [9]. Contrast masking was considered for contrast weighted computation of the MSE in the DCT domain for alternative PSNR computations [21, 73]. More sophisticated visually weighted distortion or quality measures have been reported recently, e.g., deploying a spatiotemporal JND model in a PSPNR (peak signal-to-perceptual-noise ratio) formulation for video quality assessment [11, 122], and considering both contrast threshold detection model and global-precedence-preserving contrast model in construction of a VSNR (visual signal-to-perceptual-noise ratio) based on the discrete wavelet decomposition for image quality evaluations [9]. A method of *information content weighting* (ICW) was reported in [103], in comparison with distortion, saliency, and contrast weighting techniques, where various perceptual weighting methods were applied to the PSNR showing noticeable improvement in quality prediction performance. The mathematical construct of the MSE is straightforward and simple by today's standard, and so is the PSNR. The MSE based R-D optimization is mathematically tractable which is extremely attractive to picture coder designers and implementers [3]. MSE with visual weighting leads to solutions in familiar distortion/quality assessment or R-DO framework [46]. Computation complexity of visually weighted MSE or PSNR increases with increase in the complexity of the vision model used to devise the visual weighting strategies [9, 11, 46, 103, 122]. The VSNR reported in [9] using a wavelet based visual model of masking and summation claims to have low computation complexity and low memory requirements.

1.2.2 Visual Feature-Driven Quality Metrics

Feature extraction based approach to picture quality metric design formulates a linear or nonlinear cost function of various distortion measures using features extracted from given reference and processed pictures, considering aspects of HVS (e.g., contrast sensitivity function or CSF for short, luminance adaption and spatiotemporal masking effects), and optimizes coefficients of the cost function to maximize the correlation of picture quality or distortion estimate with the MOS (mean opinion score) from subjective test data.

PQS (objective picture quality scale) was first introduced in [56] and further refined in [58]. The design philosophy of PQS is summarized in [57] which leads to a metric construct consisting of generation of visually adjusted and/or weighted distortion and distortion feature maps (i.e., images), computation and normalization of distortion indicators (i.e., measures), decorrelated principal perceptual indicators by the principal decomposition analysis (PDA), and pooling principal distortion indicators with weights determined by multiple regression analysis to fit subjective test data (e.g., MOS) to form the quality estimator (i.e., PQS in this case). The PQS considers a number of visual features, including luminance coding error weighted by contrast sensitivity and brightness sensitivity described by the Weber–Fechner's law [99], perceptible difference normalized as per Weber's law, perceptible blocking artifacts, perceptible correlated errors, and localized errors of high contrast/intensity transitions by visual masking. Use of the PDA in the PQS was intended to decorrelate any overlapping between the distortion indicators generated from the feature maps which are more or less extracted empirically, and omitted in many later distortion metric implementations only to be compensated by the regression (or optimization) process in terms of the least mean square error, linear correlation, or some other measures [125].

An example which followed and adopted this approach was an early representative video quality assessment metric, $\hat{s}$ (s-hat),[10] by ITS[11] [111], leading to the standardized VQM in the ANSI[12] and the ITU-T[13] objective perceptual video quality measurement standards [34, 72]. Six impairment indicators/measures and one picture quality improvement indicator/measure[14] are formulated based on spatial and temporal features of luminance and chrominance channels in a linear combination to form the VQM general model with parameters/coefficients of the model optimized using the iterative nested least squares algorithm to fit against a set of subjective training data. The impairment measures of the VQM were designed to measure perceptual impact of blurring, block distortion, jerky and unnatural motion, luminance and chrominance channel noise, and error blocks due to transmission error or packet loss. The quality improvement measure of the VQM was designed to assess the visual improvement resulted from edge sharpening or enhancement. The VQM general model was reported in [72] to have performed statistically better than or at least equivalent to others recommended in [34] in either the 525-line or 625-line video test.

Various picture distortion or quality metrics designed using this approach rely on extraction of spatial and/or temporal features, notably edge features [32, 58, 72], which deem to be visually significant to perception of picture distortion/quality, and a pooling strategy for formulation of an overall distortion/quality measure with parameters optimized by a regression process to fit subjective test data. Computation complexity of visual feature based perceptual measures varies from low to medium high depending on the number of features required and algorithms used for the feature extraction.

[10] $\hat{s}$ consists of three distortion measures, including blur-ringing and false edges, localized jerky motion due to frame repetition, and temporal distortion due to periodic noise, uncorrected block errors due to transmission errors or packet loss and maximum jerky motion of the time history [111].

[11] Institute for Telecommunication Sciences, National Telecommunications & Information Administration (NTIA), USA.

[12] American National Standards Institute.

[13] International Telecommunication Union, Telecommunication Standardization Sector.

[14] In [34, 72], these seven indicators/measures or sub-metrics were referred to as parameters. Weighting constants for the seven measures are referred to parameters or coefficients here which are determined or optimized for the outputs of VQM to fit the subjective test data.

1.2.3 Natural Scene Statistics Based Perceptual Metrics

Natural scene statistics (NSS) model-based approach to picture quality measurement assumes that the HVS and natural scene modelings are dual problems and, therefore, visual quality or distortion can be captured by NSS [4, 69]. Representatives in this category include the structure similarity index (SSIM) [101] and its variants [79, 103], visual information fidelity measure (VIF) [84], and texture similarity measure [67, 130]. The SSIM and the VIF have been frequently referenced and used in QoP performance benchmarking in recent years, and thanks to its low computation complexity, the SSIM has been applied to perceptual picture coding design using RpD optimization [118]. A number of weighting techniques have been reported for pooling of local distortion or quality measures to form an overall distortion or quality metric, including saliency, distortion, contrast and ICW. It is noted that an ICW technique may lead to better estimation outcomes for existing distortion and quality measures, including the MSE and the PSNR, as reported in [103].

1.2.3.1 Structure Similarity

Formulation of the SSIM is based on the assumption that structure information perception plays an important role in perceived QoP by the HVS and structural distortions due to additive noise, low-pass filtering induced blurring and other coding artifacts affect perceived picture quality more than non-structural distortions such as a change in brightness and contrast, or spatial shift or rotation and Gamma correction or change [100]. The SSIM replaces pixel-wise distortion or quality measurements by patch-wise metrics using region statistics [79, 101]. The SSIM adopted a two-step approach allowing human perception based formulations of localized quality or distortion measurements [101] and visual or ICW in pooling of the measurements [103]. To address the issues with non-stationary nature of spatial (and temporal) picture and distortion signals as well as visual attention of the HVS, the SSIM is applied locally, e.g., to a defined window, leading to windowed SSIM. The localized SSIM quality metric is formulated as a product of luminance, contrast, and structure similarity measures, which use, respectively, the means of the reference and the processed images for the luminance similarity measure (or comparison function) accommodating the luminance masking effects, the standard deviations for contrast similarity measure considering the contrast masking effects, and the cross correlation of the mean removed and normalized images from the original and the processed image for structure similarity measure (equivalent to and represented by the cross correlation of the original and the processed image). The local SSIM is computed pixel-by-pixel over the entire image with a moving window, which generates an SSIM map. To avoid the blocking artifacts caused by an initial 8×8 window, an 11×11 window with weights defined by circular-symmetric Gaussian function (of standard deviation being 1.5 samples) normalized to unity sum was used for computations of the mean, the standard deviation and the cross correlation in [101]. The overall SSIM is then computed as the average of relevant local SSIMs.

The SSIM first developed for monochrome images has been extended to multi-scale representation of images, color images with channel weights being 0.8 for Y, 0.1 for C_B and 0.1 for C_R, respectively, [28, 102] and video [102, 104]. To address a major issue of the SSIM which is its high sensitivity to picture translation, rotation and scaling, the complex wavelet SSIM (CW-SSIM) was devised whose picture similarity estimation performance was shown to be more robust to small rotations and translations [79]. To further improve SSIM's overall picture estimation performance, various visual weighting schemes have been investigated, including distortion, saliency, and contrast weighting, in pooling of localized SSIMs. ICW for pooling of multi-scale SSIM (MS-SSIM) was reported to demonstrate consistent picture quality estimation performance in terms of key prediction performance indicators [103].

1.2.3.2 Visual Information Fidelity

Using source (natural scene picture statistics) model, distortion model and HVS "visual distortion"[15] model, VIF formulation takes an information theoretic approach to QoP assessment where the picture quality measure is defined as the ratio between the mutual information representing the perceivable information of the processed picture and that representing the perceivable information of the reference. As shown in Fig. 1.3, a Gaussian scale mixture (GSM) model, $\mathscr{C}$, in the wavelet decomposition domain[16] is used to represent the reference picture. A random field (RF), $\mathscr{D}$, models the attenuation such as blur and contrast changes, and additive noise of the channel and/or coding which represent equal perceptual-annoyance by the distortion instead of modeling specific image artifacts. All HVS effects are considered as uncertainty and treated as visual distortion which is modeled as a stationary, zero-mean, additive white Gaussian noise model, $\mathscr{N}$, corresponding to the reference (or $\mathscr{N}_d$ for the processed), in the wavelet domain.

To describe detailed mathematical formulation of the VIF in correspondence to Fig. 1.3 for monochrome images, mathematical representations of a monochrome image and its transform are prescribed as follows. A monochrome image, $\mathbf{x}[\mathbf{n}]$, with a height of N_1 pixels and a width of N_2 pixels where $\mathbf{n} = [n_1, n_2]$ for $0 \leq n_1 \leq N_1 - 1$ and $0 \leq n_2 \leq N_2 - 1$, has a transform or decomposition, $\mathbf{X}[\mathbf{b}, \mathbf{k}]$, where $\mathbf{k} = [k_1, k_2]$ defines the position row and column indices of a coefficient in a block of a frequency band $\mathbf{b}$ in the

[15] In [84], it is referred to as "HVS distortion visual noise."

[16] The GSM model in wavelet domain is an RF expressed as a product of two independent RFs and is used to approximate key statistical features of natural pictures [98].

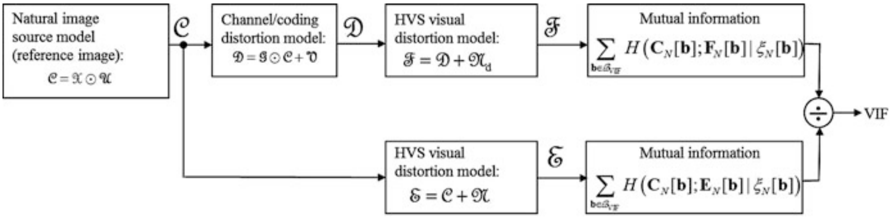

Natural image source model (reference image): $\mathfrak{C} = \mathfrak{X} \odot \mathfrak{U}$
$\mathfrak{C}$
Channel/coding distortion model: $\mathfrak{D} = \mathfrak{G} \odot \mathfrak{C} + \mathfrak{V}$
$\mathfrak{D}$
HVS visual distortion model: $\mathfrak{F} = \mathfrak{D} + \mathfrak{N}_d$
$\mathfrak{F}$
Mutual information
$\sum_{b \in \mathcal{B}_{VIF}} H(\mathbf{C}_N[\mathbf{b}]; \mathbf{F}_N[\mathbf{b}] \mid \xi_N[\mathbf{b}])$
$\div$
VIF
HVS visual distortion model: $\mathfrak{E} = \mathfrak{C} + \mathfrak{N}$
$\mathfrak{E}$
Mutual information
$\sum_{b \in \mathcal{B}_{VIF}} H(\mathbf{C}_N[\mathbf{b}]; \mathbf{E}_N[\mathbf{b}] \mid \xi_N[\mathbf{b}])$

Fig. 1.3 An information theoretic framework used by VIF measurement (after [84])

decomposition domain. For example, the subband $\mathbf{b} = [s, \theta]$ for a three level DWT decomposition, where $s = 3$ and $\theta \in \Theta = \{\theta_0|\text{LL band}, \theta_1|\text{LH band}, \theta_2|\text{HL band}, \theta_3|\text{HH band}\}$.

For a selected subband $\mathbf{b}$ where $\mathbf{b} = [s, \theta]$ with level s and orientation θ, in the wavelet transform domain, the VIF measure is defined as [84]

$$VIF = \frac{\sum_{\mathbf{b} \in \mathscr{B}_{VIF}} H(\mathbf{C}_N[\mathbf{b}]; \mathbf{F}_N[\mathbf{b}], \boldsymbol{\xi}_N[\mathbf{b}])}{\sum_{\mathbf{b} \in \mathscr{B}_{VIF}} H(\mathbf{C}_N[\mathbf{b}]; \mathbf{E}_N[\mathbf{b}], \boldsymbol{\xi}_N[\mathbf{b}])} \tag{1.1}$$

where the mutual information between the reference image and the perceived reference image in the same subband $\mathbf{b}$ is defined as $H(\mathbf{C}_N[\mathbf{b}]; \mathbf{E}_N[\mathbf{b}], \boldsymbol{\xi}_N[\mathbf{b}])$ with $\boldsymbol{\xi}_N[\mathbf{b}]$ being a realization of N elements in $\mathscr{X}$ for a given reference image, and that between the processed image and the perceived processed image by the HVS is $H(\mathbf{C}_N[\mathbf{b}]; \mathbf{F}_N[\mathbf{b}], \boldsymbol{\xi}_N[\mathbf{b}])$, and $\mathbf{C}_N[\mathbf{b}] = [\mathbf{C}_1, \mathbf{C}_2, \ldots, \mathbf{C}_N] \in \mathscr{C}$, $\mathscr{C} = \{\mathbf{C}_k | k \in \mathscr{K}\} = \mathscr{X} \odot \mathscr{U} = \{\xi_k \mathbf{U}_k | k \in \mathscr{K}\}$ is the GSM, a random field (RF), as the NSS model in the wavelet domain, approximating the reference image, $\mathbf{C}_k$ and $\mathbf{U}_k$ are M-dimensional vectors consisting of non-overlapping blocks of M coefficients in a given subband, $\mathscr{U} = \{\mathbf{U}_k | k \in \mathscr{K} \text{and } \mathbf{U}_{k_1} \text{is independent of } \mathbf{U}_{k_2}, \forall k_1 \neq k_2, \text{and } k_1, k_2 \in \mathscr{K}\}$ a Gaussian vector RF with zero-mean and covariance $\mathbf{C}_{\mathscr{U}}$, $\boldsymbol{\xi}_N[\mathbf{b}] = [\xi_1, \xi_2, \ldots, \xi_N] \in \mathscr{X}$, $\mathscr{X} = \{\xi_k | k \in \mathscr{K}\}$ an RF of positive scalars, symbol "$\odot$" defines element-by-element product of two RFs [84], and $\mathscr{K}$ is the set of location indices in the wavelet decomposition domain, $\mathbf{D}_N[\mathbf{b}] = [\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_N] \in \mathscr{D}$, $\mathscr{D} = \{\mathbf{D}_k | k \in \mathscr{K}\} = \mathscr{G} \odot \mathscr{C} + \mathscr{V} = \{g_k \cdot \mathbf{C}_k + \mathbf{V}_k | k \in \mathscr{K}\}$, the RF representing the distorted image in the same subband, $\mathscr{G} = \{g_k | k \in \mathscr{K}\}$ a deterministic scalar field which is slow varying, $\mathscr{V} = \{\mathbf{V}_k | k \in \mathscr{K}\}$ a stationary additive zero-mean Gaussian noise RF with variance $\mathbf{C}_{\mathscr{V}} = \sigma^2_{\mathscr{V}} \mathbf{I}$ which is white and independent of $\mathscr{X}$ with identity matrix, $\mathbf{I}$, and $\mathscr{U}$, $\mathbf{E}_N[\mathbf{b}] = [\mathbf{E}_1, \mathbf{E}_2, \ldots, \mathbf{E}_N] \in \mathscr{E}$, $\mathbf{F}_N[\mathbf{b}] = [\mathbf{F}_1, \mathbf{F}_2, \ldots, \mathbf{F}_N] \in \mathscr{F}$, $\mathscr{E} = \{\mathbf{E}_k | k \in \mathscr{K}\} = \mathscr{C} + \mathscr{N}$ and $\mathscr{F} = \{\mathbf{F}_k | k \in \mathscr{K}\} = \mathscr{D} + \mathscr{N}_d$ modeling HVS visual distortions to the reference $\mathscr{C}$ and channel/coding distortion $\mathscr{D}$, respectively, with RFs $\mathscr{N}$ and $\mathscr{N}_d$ being zero-mean uncorrelated multivariate Gaussian of M-dimensions with their covariance $\mathbf{C}_{\mathscr{N}} = \mathbf{C}_{\mathscr{N}_d} = \sigma^2_{\mathscr{N}} \mathbf{I}$ and $\sigma^2_{\mathscr{N}}$ variance of the visual noise, and $\mathscr{B}_{VIF}$ the select subband critical to VIF computation.

If $M \times N$ elements of $\mathbf{C}_N[\mathbf{b}]$ contain all DWT coefficients of subband $\mathbf{b}$ for all the subbands (i.e., $\forall\mathbf{b}$), the VIF calculated by (1.1) predicts the information fidelity of the entire image perceivable by the HVS. If, however, $\mathbf{C}_N[\mathbf{b}]$ only contains DWT coefficients of a spatially localized region of subband $\mathbf{b}$, a VIF map can be generated using a sliding window to show picture quality variations of the image as measured by the VIF.

When there is no distortion, VIF equals unity. When VIF is greater than unity, the processed picture is perceptually superior to the reference picture as may be the case in a visually enhanced picture.

Computation complexity of the VIF is much higher than other NSS based metrics, e.g., SSIM and MS-SSIM, due to the wavelet decomposition and the distortion model parameterization process [84]. Detailed parameterization algorithms for the source, the distortion and the HVS models can be found in [84, 87].

1.2.3.3 Texture Similarity

Structure texture similarity metric (STSIM) measures perceived texture similarity between a reference picture and a processed counterpart to address an issue with the SSIM which tends to give low similarity values to textures which are perceptually similar [130]. The framework used by the STSIM consists of subband decomposition, e.g., using steerable filterbanks [86], computation of a set of statistics including the mean, the variance, horizontal and vertical autocorrelations and crossband correlation, statistic comparisons and pooling scores across statistics, subbands and window positions. More detailed discussions and reviews of various texture similarity metrics can be found in [67, 130].

A major issue with using STSIM and the like in perceptual video coding is that regions covering the same spatial location in consecutive frames rated by the STSIM as having the same structural texture or quality may result in frame differences in the very regions greater than the JND leading to perceptible temporal fluctuation artifacts or flickering [126] and may or may not achieve RpD optimization performance gain in video coding applications [60].

1.2.4 HVS Model-Based Perceptual Distortion Metrics

HVS model-based approach to picture distortion or quality estimation employs human visual perception model in metric design, which characterizes low-level vision in terms of spatiotemporal response, color vision, and foveation [118]. Three types of HVS models have emerged, including JND models, multichannel contrast gain control (CGC) models and suprathreshold models, which have been successfully applied to picture quality assessment and perceptual picture coding design using RpD optimization. The multichannel structure of the HVS decomposes visual signal into several spatial, temporal, and orientation bands where masking parameters will be determined based on human visual experiments [34, 125].

1.2.4.1 JND Models

The HVS cannot perceive all changes in an image or a video sequence, nor does it respond to varying changes in a uniform manner [80, 99]. In picture coding, JND threshold detection based HVS models are extensively reported [10, 41, 44, 47–49, 108] and used in QoP assessment, perceptual quantization for picture coding and perceptual distortion measures (PDMs) in RpD performance optimization for visual signal processing and transmission services [118].

JND models reported currently in the literature consider (1) spatial/temporal CSF which describes the sensitivity of the HVS to each frequency component, as determined by psychophysical experiments; (2) background luminance adaptation (LA) which refers to how the contrast sensitivity of the HVS changes as a function of the background luminance; and (3) contrast masking (CM) which refers to the masking effect of the HVS in the presence of two or more simultaneous frequency components where all orientations for a given scale and adjacent scales for a given orientation are usually considered in addition to components in the same subband [19, 22, 52, 92, 93, 109, 113, 125, 130]. The JND model can be represented in either spatiotemporal domain, or transform/decomposition domain, or both. Examples of JND models are found with CSF, CM and LA modeling in the DCT domain [1, 70, 77, 112], and CSF and CM modeling using sub-band decomposition [12,30,50,78]; or in the pixel domain [43,122], where the key issue is to differentiate edge from textured regions [51, 121].

The local luminance JND model in sub-band decomposition domain is generally formulated by the base visibility threshold at the given location in the given sub-band of the given frame determined by spatiotemporal CSF modulated by different elevation factors due to intra-band masking, inter-band masking, temporal masking, and luminance adaptation [49, 118]. Global response of the HVS has been accommodated by modulating the local JND model using a VA (visual attention) or foveation model centered at the point of VA [118].

Two approaches have been reported to modeling of just noticeable color difference (JNCD), i.e., to model each color component channel independently in a similar way by which the luminance JND model is formulated, and, alternatively, to model JNCD by a base visibility threshold of distortion for all colors, JNCD$_{00}$ at a given spatiotemporal pixel location which is modulated by masking effect of non-uniform neighborhood (measured by the variance) and a scale function, modeling the masking effect induced primarily by local changes of luminance (measured by the gradient of the luminance component) [118]. It is noted that the base color difference visibility threshold is determined using the point-by-point color difference, ΔE_{00}, as measured by a perceptually more uniform color metric calculated in polar coordinates of the CIELAB space with luminance, chroma, and hue components [13, 54]:

$$\Delta E_{00} = \sqrt{(\frac{\Delta L'}{\alpha_L \cdot S_L})^2 + (\frac{\Delta C'_{ab}}{\alpha_C \cdot S_C})^2 + (\frac{\Delta H'_{ab}}{\alpha_H \cdot S_H})^2 + R_T(\frac{\Delta C'_{ab}}{\alpha_C \cdot S_C}) \cdot (\frac{\Delta H'_{ab}}{\alpha_H \cdot S_H})} \tag{1.2}$$

where $\Delta L'$, $\Delta C'_{ab}$, and $\Delta H'_{ab}$ are the luminance, chroma, and hue components of the color difference, respectively, α_L, α_C, and α_H are parameters, S_L, S_C, and S_H are weighting functions, and R_T is a parameter to adjust the orientation of the discrimination ellipsoids in the blue region. Subscript "00" indicates the point-by-point color difference formula defined in CIEDE00 in 2000 [54].

Perceptually lossless visual signal coding requires a PDM which controls the distortion at no greater than JNND level or below JND level [89].

1.2.4.2 Multi-Channel Vision Model

Contrast gain control or CGC [22, 109] has been successfully used in various implementations for JND detection [19, 52, 92], QoP assessment [93, 113, 125] and perceptual picture coding in either standard alone [68] or embedded forms [88, 90, 116, 119]. The general CGC model consists of color space transform, visual decomposition, spatiotemporal contrast sensitivity modeling, luminance adaptation, contrast and texture masking, visible difference detection, and pooling of perceptual differences over all channels (frequency and orientation bands). Usually, visually uniform color space, e.g., CIELab or opponent color space [99, 127], is preferred as result of the color space transformation. Various transforms have been used for visual decomposition including over-complete transforms such as steerable pyramid transform [86, 92] and the cortex transform [105], and complete transforms such as the DCT [107] and the DWT [90]. The over-complete transforms are shift-invariant and free from aliasing which is an inherent problem for complete transforms [86,88]. It is noted that the DWT as commonly adopted for visual signal decomposition [2] is not able to represent directional features in diagonal and anti-diagonal directions separately, which exacerbates visual impact of pattern aliasing effects as shown in Fig. 1.4. Spatiotemporal contrast sensitivity is represented by contrast weights which are reciprocally proportional to the base visibility threshold determined by the CSF [88]. Excitatory and inhibitory nonlinearities are formulated by power-law, which are then used as the inputs to a divisive gain control [92, 109]. The divisive gain control implements the visual masking effects by normalizing the excitatory channel by weighted sum of responses of adjacent frequencies and orientations [88, 92].

Based on the original Sarnoff's visual discrimination model-JNDmetrix$^{\text{TM}}$ [96, 97], PQR (picture quality rating) is devised and extensively documented in ITU-T J.144 recommendation and frequently used as a benchmark [34]. This model is also used in a standalone implementation for RpD optimization for MPEG-2 video encoding [68].

Another example of the multichannel CGC model in visual decomposition domain is briefly described in [88] for embedding a PDM in RpD optimization of a standard compliant coder, which consists of a frequency transform (with a 9/7 filter), CSF weighting, intra-band and inter-orientation contrast masking including texture masking, detection, and pooling.

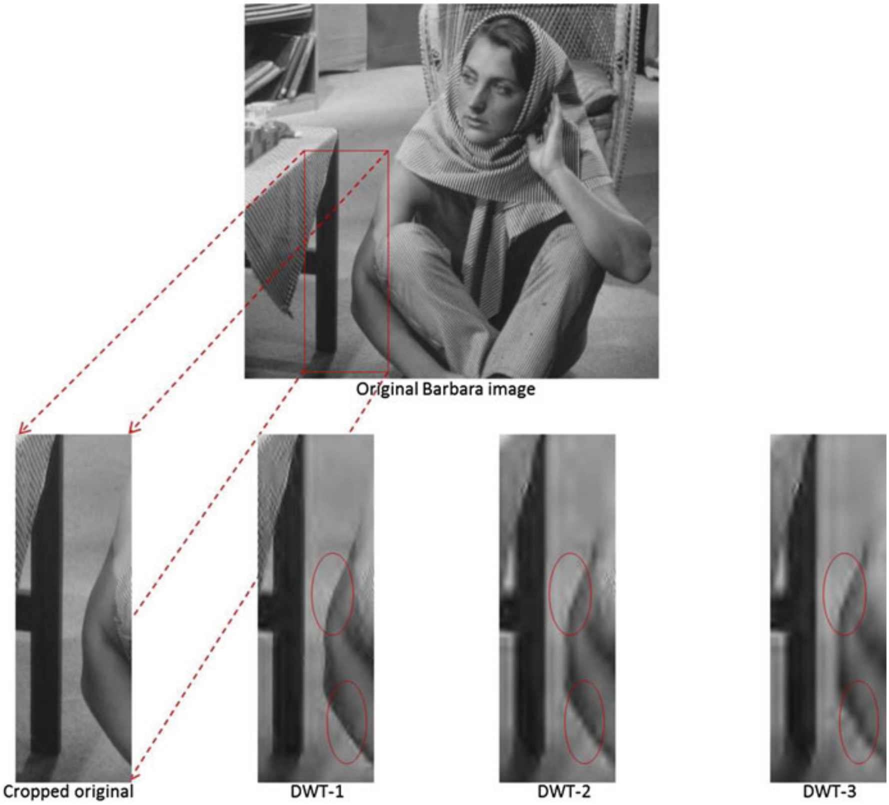


Fig. 1.4 An example of pattern aliasing in images coded by a DWT based coder at three different bitrates. The original *Barbara* image is shown on the *top*. Images on the *bottom row* from left to right are the cropped section from the original *Barbara* image, the same section from coded images by the DWT coder at reduced bitrates. Pattern aliasing distortions are highlighted by *red elongated circles* (Courtesy of Dr. D.M. Tan)

1.2.4.3 Suprathreshold Vision Models

To be cost-effective, various visual signal processing and compression applications operate in the so-called suprathreshold domain, where processing distortions or compression artifacts are visible to human observers [99, 118]. It has been questioned whether extension of threshold vision models discussed in previous sections by linear scaling or weighting for quality assessment of processed or coded visual signals with suprathreshold distortions is theoretically plausible and practically effective [75].

A suprathreshold wavelet coefficient quantization experiment has reported that the first three visible differences (relative to the original image) are well predicted by an exponential function of sub-band standard deviation, and regression lines corresponding to JND_2 and JND_3 are parallel to that of JND_1 [75]. Therefore, it was suggested that if a perceptual quantization strategy was formulated for coding an image at JND_1, it could be scaled to encode the image at JND_2 and JND_3, etc. The quantization design strategy based on this suprathreshold model was reported to contradict those which used visual weighting or scaling of threshold models for perceptually lossy visual signal compression achieving improved picture quality as perceivable to the HVS.

A composite model approach has been reported which integrates the threshold detection model and the suprathreshold model in perception based visual signal coding [8] and quality or distortion measurement design [9]. A more recent example of this approach is the MAD (most apparent distortion) which measures suprathreshold distortion using a detection model and appearance model in the form of [45]

$$MAD = (D_{\text{detection}})^{\alpha} (D_{\text{appearance}})^{\alpha-1} \tag{1.3}$$

where $D_{\text{detection}}$ is the perceived distortion due to visual detection which is formulated in a similar way to JND models and $D_{\text{appearance}}$ a visual appearance based distortion measure dependent on changes in log-Gabor statistics such as the standard deviation, skewness and kurtosis of sub-band coefficients, and is weight adapted to severity of the distortion as measured by $D_{\text{detection}}$ as follows:

$$\alpha = \frac{1}{1 + \beta_1 (D_{\text{detection}})^{\beta_2}}, \tag{1.4}$$

where $\beta_1 = 0.467$ and $\beta_2 = 0.130$.

1.2.5 Light-Weight Bit-Stream-Based Models [123, 124]

In real-time visual communications, broadcasting and entertainment services, QoE assessment and monitoring tasks face various constraints such as availability of full or partial information on reference pictures, computation power, and real-time or on-line assessment. While no-reference picture quality metrics provide feasible solutions [29, 124], it prompted investigations into light-weight QoE methods and associated standardization activities. There are at least three identifiable models, including parametric model, packet layer model, and bit-stream layer model. With very limited information acquired or extracted from the transmission payload, stringent transmission delay constraint, and limited computation resources, these models share a common technique, i.e., optimization of perceptual quality or distortion predictors via, e.g., regression or algorithms of similar trade using ground truth subjective test data (e.g., the MOS or the DMOS) and optimization criteria such as Pearson linear correlation, Spearman rank-order correlation, outlier ratio, and the RMSE (root mean square error) [29, 125].

Relying on KPI (key performance indicators) collected by network equipment via statistical analysis, a crude prediction of perceived picture quality or distortion is made by a *parametric model* using bitrate (R) and packet loss rate (PLR) along with side information, e.g., codec type and video resolution, to assist with adaptation of model parameters to differently coded visual signals. Since the bitrate does not correlate well with the MOS data for pictures of varying contents and packet losses which occur at different locations in a bit-stream may have significantly different impacts on perceived picture quality [118], the quality estimation accuracy based on this model is limited while computation required is usually trivial.

With more information available via packet header analysis to the *packet layer model*, distortions at picture frame level can be better estimated with information on coding parameters such as frame type and bitrate per frame, frame rate and position of lost packets as well as the PLR. Temporal complexity of the video contents is estimated using ratios between bitrates of different types of frames. The packet layer model incorporates temporal pooling for better quality or distortion prediction with moderate computational costs [123].

By accessing the media payload as well as packet layer information, the *bit-stream layer model* allows picture quality estimation either with or without pixel information [123, 124]. It usually incurs the highest computational cost amongst the light-weight picture quality or distortion models.

1.2.6 Perceptual Quality and Distortion Assessment of Audiovisual Signals

Investigations on QoE assessment, which integrates audio and visual components beyond the preliminary based on human perception and integrated human audiovisual system modeling, have been very limited [17, 71, 115].

1.2.7 Perceptual Quality/Distortion Assessment of 3-D/Multiview Visual Signals

Technological advances in visual communications, broadcasting and entertainment continue to captivate the general public, offering a new height of viewing quality and experience with three-dimensional (3-D) full HD digital video [42, 65, 94]. Many theoretical and practical challenges remain in 3-D video acquisition, display, coding and compression, and quality assessment and metrics [74, 114]. While various visual distortions associated with 3-D visual signal coding have been identified and investigated [20], perceptual quality/distortion measures based on the HVS have yet been further developed as well as subjective quality assessment methods [114].

1.3 PDMs for RpDO in Visual Quality Regulated Services

A fundamental principle for quality-driven picture coding design is RpD optimization, where PDM plays a key role in aligning steps/units of distortions, hopefully, consistently with discernible levels by human visual perception in terms of, e.g., JNND, JND_1, JND_2, and so on [116, 118]. Using RpDO based approach to constant quality picture coding design, the designated visual quality level is achieved by controlling perceptual distortions estimated by the PDM at a corresponding (constant) level as perceived and desired by the HVS while minimizing the required coding bitrate. It then begs the question whether the existing PDMs can consistently predict discernible levels by the HVS, while they have been reported to grade distortions reasonably successfully in correspondence with HVS perception [10, 34, 49, 85].

A preliminary experimental investigation has been conducted to ascertain if various perceptual image metrics under evaluation are able to consistently grade different images at various JND levels. Images were generated using an open source JPEG 2000 coder [66] at various (increasingly higher) compression ratios for a total of eighty-one (81) variations for each of forty-one (41) well-known test images. This provides a range of test pictures that capture the transition points between JND levels. An image at JND_n is determined relative to the image at $\text{JND}_{(n-1)}$, except for JND_1 which was relative to the reference such that JND_2 is relative to JND_1 and JND_3 to JND_2, etc. Perceptual distortion or quality measures were computed for sets of images at JND_1, JND_2, JND_3, JND_4, and JND_5, respectively.

Small data samples with normal distribution was assumed as in the present case (41 test images per a JND level) and, therefore, the 95 % confidence interval (CI) was used to identify the upper and lower bounds relative to the mean and standard deviation of the data in which 95 % of the responses resided [40]. If the variation is such that most of the responses from a metric (i.e., >50 %) do sit outside the 95 % CI range, then one may be inclined to conclude that the behavior of that metric is inconsistent, i.e., the metric is ineffective.

In Table 1.1, preliminary data were collected for a well-known early perceptual distortion metric based on the DCT decomposition, using DCTune 2.0 in error calculation mode [107]. Images in JNND test set were encoded using a perceptually lossless coder [89]. Two observations deserve immediate attention. First, for images at the same JND level, the metric produced a range of values. Second, acceptance rates were lower than 50 %, indicating that the metric is ineffective as a measure to predict discernible visual quality levels in terms of JNDs. Similar results were obtained for a number of other perceptual distortion or quality measures including the SSIM [101], the VIF [84], the PSNR-HVS [21], the PSNR-HVS-M [73], the VSNR [9], the MAD [45], and the FSIM (feature similarity index) [128].

It is noted that when non-overlapping boundaries were applied to JNND and JND levels, the aforementioned metrics tested were able to achieve an acceptance rates greater than 50 % for JNND, while falling below 50 % for the rest of JND levels. In other words, reliable prediction of JND_n levels is still a challenge when $n \geq 1$.

Table 1.1 Perceptual distortion metric outputs when evaluating images at first five JND levels compared with JNND

	DCTune metric [107]											
	Metric values						CI acceptance[a]					
Image	JNND	JND_1	JND_2	JND_3	JND_4	JND_5	JNND	JND_1	JND_2	JND_3	JND_4	JND_5
Beachbum	1.774	1.771	2.342	2.949	4.010	4.439	0	0	0	0	0	0
Bikes	3.069	6.795	7.208	10.056	10.529	11.805	0	0	0	0	0	0
Buildings	3.290	3.266	4.901	6.338	8.373	10.668	0	1	1	1	0	0
Caps	2.116	1.758	2.438	1.938	2.562	3.193	0	0	0	0	0	0
Flowers	2.124	1.919	2.897	4.089	5.272	6.098	0	0	0	0	0	0
Frontbuilding	2.946	2.691	4.968	5.661	6.480	8.276	0	0	1	1	1	1
Girl	3.068	2.505	3.728	4.569	5.500	7.154	0	0	0	0	0	0
House	2.340	2.737	4.016	5.502	6.320	7.497	0	0	0	1	0	1
Lighthouse	2.405	1.706	2.510	3.186	4.464	4.902	0	0	0	0	0	0
Lighthouse2	2.417	2.061	2.789	3.728	4.523	5.675	0	0	0	0	0	0
Ocean	2.354	1.679	2.546	3.017	4.366	5.610	0	0	0	0	0	0
Paintedhouse	2.779	2.970	5.436	5.990	7.776	8.221	1	0	0	1	1	1
Parrots	1.919	1.357	1.642	2.335	3.001	3.948	0	0	0	0	0	0
Plane	2.274	2.342	3.185	4.671	5.692	6.022	0	0	0	0	0	0
Rapids	2.780	4.155	6.802	8.452	10.255	11.400	1	0	0	0	0	0
Reddoor	2.329	2.244	3.410	4.419	5.425	6.501	0	0	0	0	0	0
Sailing1	2.645	2.871	5.257	5.748	8.212	9.191	1	0	1	1	1	1
Sailing2	1.847	1.985	2.653	3.452	3.651	4.409	0	0	0	0	0	0
Sailing3	2.104	1.909	2.543	3.324	4.061	4.587	0	0	0	0	0	0
Sailing4	2.849	3.901	4.152	5.870	6.983	7.510	1	1	1	1	1	1

(continued)

Table 1.1 (continued)

	DCTune metric [107]											
	Metric values						CI acceptance[a]					
Image	JNND	JND_1	JND_2	JND_3	JND_4	JND_5	JNND	JND_1	JND_2	JND_3	JND_4	JND_5
Statue	3.023	4.831	4.953	6.472	7.626	8.549	0	0	1	1	1	1
Stream	3.042	5.552	6.844	10.169	11.316	11.891	0	0	0	0	0	0
Womanhat	2.284	2.883	3.520	4.110	5.685	7.378	0	0	0	0	0	1
Woman	3.075	1.972	3.249	4.062	5.361	5.851	0	0	0	0	0	0
Merry0080	5.235	8.375	10.339	12.423	14.910	16.122	0	0	0	0	0	0
Merry_0006_Lasalle	2.764	4.122	5.336	5.934	6.083	7.398	1	0	1	1	0	1
Merry_mtl07_041	2.999	5.184	7.481	9.378	10.511	11.293	0	0	0	0	0	0
Pippin_jtalon0002	2.319	3.395	3.767	5.006	5.807	6.397	0	1	0	0	0	0
Pippin_jtalon0022	2.088	2.645	3.881	4.876	5.484	6.563	0	0	0	0	0	0
Pippin_Mex07_014	2.941	3.978	5.004	6.531	7.182	8.626	0	1	1	1	1	1
Pippin_park0037	2.732	4.004	5.276	6.210	9.390	10.482	1	0	1	1	1	1
Merry_0064_Lasalle	3.049	5.279	5.396	6.733	8.239	8.328	0	0	0	0	0	0
Merry_florida0034	3.857	7.013	9.817	14.043	16.468	17.180	0	0	0	0	0	0
Merry_italy0009	2.346	3.138	3.659	4.484	5.486	6.027	1	1	1	1	1	1
Merry_italy0044	3.041	3.971	4.485	6.230	7.125	7.876	0	0	0	0	0	0
Merry_italy0179	3.031	5.021	5.654	6.827	8.448	9.863	0	0	0	0	0	0
Merry_mexico0061	2.998	3.960	5.479	7.441	10.948	12.345	0	0	0	0	0	0
Merry_mexico0118	2.932	6.847	8.478	12.363	13.869	15.862	0	0	0	0	0	0
Merry_mexico0123	2.487	3.876	6.674	7.136	7.630	9.818	0	0	0	0	0	0
Merry_mexico0211	2.436	3.906	5.754	6.259	7.186	9.196	0	0	0	0	0	0
Merry_mexico0223	2.877	4.389	4.805	6.869	7.452	7.972	0	0	0	0	0	0

Mean (μ)	2.707	3.584	4.763	6.069	7.309	8.344						
s.d.[b] (σ)	0.600	1.667	2.007	2.742	3.102	3.257						
L.[c] bound 95 % CI[e]	2.517	3.058	4.129	5.204	6.330	7.316						
U.[d] bound 95 % CI[e]	2.896	4.111	5.396	6.935	8.288	9.372						
% within CI							17.07 %	21.20 %	21.95 %	26.86 %	19.51 %	26.83 %

t-distribution analysis with 40 degrees of freedom at $\alpha = 0.025$, 95 % CI (confidence interval)

[a] Acceptance (1)/Rejection (0)

[b] s.d. stands for standard deviation

[c] L. (lower) bound

[d] U. (upper) bound

[e] 95 %($\alpha = 0.025$) confidence interval with 40 degrees of freedom = 2.021

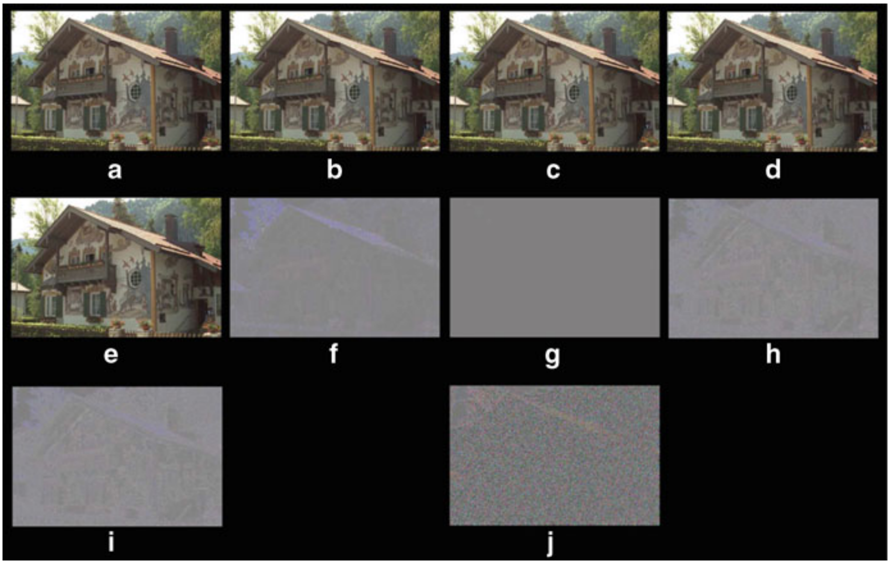


Fig. 1.5 Sample images representing the first three JND levels as well as a perceptually lossless coding result. (**a**) original Paintedhouse image; (**b**) perceptually lossless coded Paintedhouse image [104]; (**c**) processed Paintedhouse with JND_1; (**d**) processed Paintedhouse with JND_2; (**e**) processed Paintedhouse with JND_3; (**f**) the difference between (**a**) and (**b**) with an offset of 128; (**g**) the difference between (**a**) and (**c**) with an offset of 128; (**h**) the difference between (**a**) and (**d**) with an offset of 128; (**i**) the difference between (**a**) and (**e**) with an offset of 128; (**j**) the difference between (**a**) and (**c**) with an offset of 128 and contrast enhancement which assist in visualization using PDF format or photo quality printing

Figure 1.5 shows a representative test image, *Paintedhouse*, with its coded versions at JNND, JND_1, JND_2 and JND_3 levels, respectively. The difference images with an offset of 128 between the reference and coded images at the first three JND levels are also shown in Fig. 1.5 to appreciate where noticeable distortions between each distortion levels are observed on a reference monitor (e.g., a Sony BVM-L231 23-inch trimaster LCD reference monitor was used in this case).

1.4 Summary and Remarks

The on-going research and development efforts on quality and distortion assessment of visual signal coding and transmission are viewed from visual information entropy and rate-distortion theoretic perspective to appreciate the inevitability of the current transition from technology driven services governed by QoS to user-centric (perceived) quality assured services measured and regulated by QoP and QoE in visual communications, broadcasting, entertainment, and consumer electronics applications and services. Availability of ever increasing transmission bandwidth and audiovisual systems of ever increasing spatiotemporal resolutions and dimensions have relaxed the bitrate constraint while raised users' expectations of service quality and experience at a justifiable cost.

Making every bit count – A. Pica, 1999

or making every bit accountable is founded on the perceptual entropy and RpD theories and spells out a philosophy for sustainable future development of visual communications, broadcasting, entertainment and consumer electronics industries, applications and services.

Visual signal quality and distortion assessment methods are reviewed based on ways in which they incorporate the HVS' characteristics and/or factors into their quality or distortion metric designs and their theoretical or practical grounding. Two obvious areas of research in perceptual quality/distortion assessment and measurement for audiovisual signal coding and transmission/storage are highlighted, including QoE measures based human audiovisual perception and integrated human audiovisual system modeling, and QoE measures for 3-D and multiview visual signals.

As obvious to some of readers and contentious to others as it may sound, quality driven visual communication applications and services require perceptual quality/distortion measures to estimate or predict consistently quality/distortion levels discernible by human viewers for a wide range of picture contents. Using a small sample pool of 41 well-known test images, a number of existing perceptual image quality/distortion measures (some of which are more well-known than others) were taken to the task of predicting quality/distortion of images coded at the first five JND levels as well as by a perceptually lossless coder, where 95 % confidence interval was used to analyze statistical reliability and acceptance rate of the measures as an image quality predictor for constant quality (or quality driven) image coder designs. The initial findings seem to vindicate that a change of the mindset in quality performance evaluation including subjective and objective assessments may not be entirely unreasonable or unfounded.

Acknowledgements H.R. Wu thanks Dr D.M. Tan of HD2 Technologies for his inputs and valuable discussions during compilation of Chapter 1 and all his past and present colleagues who have contributed to the subject matters covered by Chapter 1.

References

1. Ahumada AJ (1992) Luminance-model-based DCT quantization for color image compression. In: Proc SPIE 1666:365–374
2. Antonini M, Barlaud M, Mathieu P et al (1992) Image coding using wavelet transform. IEEE Trans Image Process 1(2):205–220
3. Berger T, Gibson JD (1998) Lossy source coding. IEEE Trans Inf Theory 44(10):2693–2723
4. Bovik AC (2013) Automatic prediction of perceptual image and video quality. Proc IEEE 101(9):2008–2024

5. Budrikis ZL (1972) Visual fidelity criterion and modeling. Proc IEEE 60(7):771–779
6. Carnec M, Le Callet P, Barba D (2008) Objective quality assessment of color images based on a generic perceptual reduced reference. Signal Process Image Commun 23(4):239–256
7. Chandler DM (2013) Seven challenges in image quality assessment: Past, present, and future research. ISRN Signal Processing Article ID 905685
8. Chandler DM, Hemami SS (2005) Dynamic contrast-based quantization for lossy wavelet image compression. IEEE Trans Image Process 14(4):397–410
9. Chandler DM, Hemami SS (2007) VSNR: A wavelet-based visual signal-to-noise ratio for natural images. IEEE Trans Image Process 16(9):2284–2298
10. Chikkerur S, Sundaram V, Reisslein M et al (2011) Objective Video Quality Assessment Methods: A Classification, Review, and Performance Comparison. IEEE Trans Broadcasting 57(2):165–182
11. Chou C-H, Chen C-W (1996) A perceptually optimized 3-D subband image codec for video communication over wireless channels. IEEE Trans Circuits Syst Video Technol 6(4):143–156
12. Chou C-H, Li Y-C (1995) A perceptually tuned subband image coder based on the measure of just-noticeable distortion profile. IEEE Trans Circuits Syst Video Technol 5(6):467–476
13. Chou C, Liu K (2010) A perceptually tuned watermarking scheme for color images. IEEE Trans Image Process 19(11):2966–2982
14. Clarke RJ (1985) Transform Coding of Images. Academic Presss, New York, NY
15. Corriveau P (2006) Video Quality Testing. In: Wu HR, Rao KR (eds) Digital video image quality and perceptual coding. CRC Press, Boca Raton, FL. 125–153
16. Corriveau P, Gojmerac C, Hughes B. et al (1999) All subjective scales are not created equal: The effect of context on different scales. Signal Processing 77(1):1–9
17. Coverdale P, Möller S, Raake A et al (2011) Multimedia quality assessment standards in ITU-T SG12. IEEE Signal Process Mag 28(6):91–97
18. Cutler CC (1952) Differential quantization of communication signals. U.S. Patent 2 605 361
19. Daly S (1993) The visible differences predictor: An algorithm for the assessment of image fidelity. In: Watson AB (ed) Digital Images and Human Vision. MIT Press, Cambridge, MA. 179–206.
20. Daly SJ, Held RT, Hoffman DM (2011) Perceptual issues in stereoscopic signal processing. IEEE Trans Broadcast 57(2):347–361
21. Egiazarian K, Astola J, Ponomarenko N, et al (2006) New full-reference quality metrics based on HVS. In: Proc VPQM-06. Paper 9
22. Foley JM (1994) Human luminance pattern-vision mechanisms: Masking experiments require a new model. J Opt Soc Amer A 11(6):1710–1719
23. Girod B (1993) What's wrong with mean-squared error. In: Watson AB (ed) Digital images and human vision. MIT Press, Cambridge, MA. 207–220.
24. Goodall WM (1951) Television by pulse code modulation. Bell Syst Tech J 28:33–49
25. Gonzalez RC, Woods RE (2008) Digital image processing, 3rd edn. Prentice Hall, Upper Saddle River, NJ
26. Green PE Jr (1993) Fiber optic networks. Prentice-Hall, Englewood Cliffs, NJ
27. Harrison CW (1952) Experiments with Linear Prediction in Television. Bell Sys Techn J 31(4):764–783
28. Hassan M, Bhagvati C (2012) Structural similarity measure for color images. Int J Comput Appl (0975 – 8887) 43(14):7–12
29. Hemami SS, Reibman AR (2010) No-reference image and video quality estimation: Applications and human-motivated design. Signal Process Image Commun 25:469–481
30. Höntsch I, Karam LJ (2000) Locally adaptive perceptual image coding. IEEE Trans Image Process 9(9):1285–1483
31. Inglis AF, Luther AC (1993) Video engineering, 2nd edn. McGraw-Hill, New York
32. ITU-R (2004) Objective perceptual video quality measurement techniques for standard definition digital broadcast television in the presence of a full reference. Rec. BT.1683

33. ITU-R (2012) Methodology for the subjective assessment of the quality of television pictures. Rec. BT.500-13
34. ITU-T (2004) Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference, Rec. J.144
35. ITU-T (2008) Vocabulary for performance and quality of service, Amendment 2: New definitions for inclusion in Recommendation ITU-T P.10/G.100. Rec. P.10/G.100
36. ITU-T (2011) Vocabulary for performance and quality of service, Amendment 3: New definitions for inclusion in Recommendation ITU-T P.10/G.100. Rec. P.10/G.100
37. ITU-T (2008) Subjective video quality assessment methods for multimedia applications. Rec. P.910
38. Jayant NS, Johnston J, Safranek and R (1993) Signal compression based on models of human perception. Proc IEEE 81(10):1385–1422
39. Jayant NS, Noll P (1984) Digital coding of waveforms: Principles and applications to speech and video. Prentice-Hall, Englewood Cliffs, NJ
40. Johnson RA, Bhattacharyya GK (1992) Statistics: Principles and methods, 2nd edn. John Wiley & Sons, New York
41. Karam LJ, Ebrahimi T, Hemami S et al (eds) (2009) Special issue on visual media quality assessment. IEEE J Sel Topics Signal Process 3(2)
42. Kubota A, Smolic A, Magnor M et al (2007) Multiview imaging and 3DTV-Special issue overview and introduction. IEEE Signal Process Mag 24(6):10–21
43. Jia Y, Lin W, Kassim AA (2006) Estimating just-noticeable distortion for video. IEEE Trans Circuits Syst Video Technol 16(7):820–829
44. Van den Branden Lambrecht CJ (ed) (1998) Special issue on image and video quality metrics. Signal Process 70(3)
45. Larson EC, Chandler DM (2010) Most apparent distortion: Full-reference image quality assessment and the role of strategy. J Electron Imaging 19(1):ID 011006.
46. Limb JO (1967) Source-receiver encoding of television signals. Proc. IEEE 55(3): 364–379
47. Lin W. (2006) Computational models for just-noticeable difference. In: Wu HR, Rao KR (eds) Digital video image quality and perceptual coding, CRC Press, Boca Raton, FL 281–303
48. Lin W, Ebrahimi T, Loizou PC et al (eds) (2012) Special issue on new subjective and objective methodologies for audio and visual signal processing. IEEE J Sel Top Signal Process 6
49. Lin W, Kuo C-CJ (2011) Perceptual visual quality metrics: A survey. J Vis Commun Image R 22:297–312
50. Liu Z, Karam LJ, Watson AB (2006) JPEG2000 encoding with perceptual distortion control. IEEE Trans Image Process 15(7):1763–1778
51. Liu A, Lin W, Paul M et al (2010) Just noticeable difference for image with decomposition model for separating edge and textured regions. IEEE Trans Circuits Syst Video Technol 20(11):1648–1652
52. Lubin J (1993) The use of psychophysical data and models in the analysis of display system performance. In: Watson AB (ed) Digital Images and Human Vision. MIT Press, Cambridge, MA. 163–178.
53. Luigi A, Chen CW, Tasos D (eds) (2012) QoE Management in Emerging Multimedia Services. IEEE Communications Magazine 50(4):18–19
54. Luo MR, Cui G, Rigg B (2001) The development of the CIE 2000 colour-difference formula: CIEDE2000. Col Res App 26(5):340–350
55. Mannos JL, Sakrison DJ (1974) The effects of a visual fidelity criterion on the encoding of images. IEEE Trans Inf Theory IT-20(4):525–536
56. Miyahara M. (1988) Quality assessments for visual service. IEEE Commun Mag 26:51–60
57. Miyahara M, Kawada R (2006) Philosophy of picture quality scale. In: Wu HR, Rao KR (eds) Digital video image quality and perceptual coding. CRC Press, Boca Raton, FL. 181–223
58. Miyahara M, Kotani K, Algazi VR (1998) Objective picture quality scale (PQS) for image coding. IEEE Trans Commun 46(9):1215–1226
59. Muntean G-M, Ghinea G, Frossard P et al (eds) (2008) Special Issue: Quality Issues on Multimedia Broadcasting. IEEE Trans Broadcast 54(3), Pt.II

60. Naser K, Ricordel V, Le Callet P (2014) Experimenting texture similarity metric STSIM for intra prediction mode selection and block partitioning in HEVC. In: Proc DSP2014: 882–887
61. National Institute of Standards and Technology (2000) Final report from the video quality experts group on the validation of objective models of video quality assessment. Available [Online] via: ftp.its.bldrdoc.gov
62. Oh H, Bilgin A, Marcellin MW (2013) Visually lossless encoding for JPEG2000. IEEE Trans Image Process 22(1):189–201
63. Ohm J-R, Sullivan GJ, Schwarz H et al (2012) Comparison of the coding efficiency of video coding standards-including high efficiency video coding (HEVC). IEEE Trans Circuits Syst Video Technol 22(12):1669–1684
64. O'Neal JB Jr (1966) Predictive quantizing aystems (differential pulse code modulation) for the transmission of television signals. Bell Sys Techn J 45(5):689–721
65. Onural L (2007) Television in 3-D: What are the prospects? Proc IEEE 95(6):1143–1145
66. OpenJPEG (2014) Windows Binaries of OpenJPEG library and codecs Labels: OpSys-Windows Type-Executable, (Version 2.0). Available: http://www.openjpeg.org/
67. Pappas TN, Neuhoff DL, de Ridder H et al (2013) Image analysis: Focus on texture similarity," Proc IEEE 101(9):2044–2057
68. Pica A, Isnardi M, Lubin J (2006) HVS based perceptual video encoders. In: Wu HR, Rao KR (eds) Digital video image quality and perceptual coding. CRC Press, Boca Raton, FL. 337–360
69. Párraga CA, Troscianko T, Tolhurst DJ (2005) The effects of amplitude-spectrum statistics on foveal and peripheral discrimination of changes in natural images, and a multi-resolution model. Vis Res 45(25–26):3145–3168
70. Peterson HA, Ahumada AJ, Watson AB (1993) Improved detection model for DCT coefficient quantization. In: Proc SPIE 1913:191–201
71. Pinson MH, Ingram W, Webster A (2011) Audiovisual quality components. IEEE Signal Process Mag 28(6):60–67
72. Pinson MH, Wolf S (2004) A new standardized method for objectively measuring video quality. IEEE Trans Broadcast 50(3):312–322
73. Ponomarenko N, Silvestri F, Egiazarian K et al (2007) On Between-Coefficient Contrast Masking of DCT Basis Functions. In: Proc. VPQM-07. Paper 11
74. Quan H-T, Le Callet P (2010) Video quality assessment: From 2D to 3D-challenges and future trends. In: Proc IEEE ICIP2010 4025–4028
75. Ramos MG, Hemami SS (2001) Suprathreshold wavelet coefficient quantization in complex stimuli: psychophysical evaluation and analysis. J Opt Soc Am A 18(10):2385–2397
76. Rouse DM, Hemami SS, Pépion R et al (2011) Estimating the usefulness of distorted natural images using an image contour degradation measure. J Opt Soc Am A 28(2):157–188
77. Safranek RJ (1994) A JPEG compliant encoder utilizing perceptually based quantization. In: Proc SPIE 2179:117–126
78. Safranek RJ, Johnston JD (1989) A perceptually tuned subband image coder with image dependent quantization and post-quantization. In: Proc IEEE ICASSP 1945–1948
79. Sampat MP, Wang Z, Gupta S et al (2009) Complex wavelet structural similarity: A new image similarity index. IEEE Trans Image Process 18(11):2385–2401
80. Sekuler R, Blake R (1994) Perception, 3rd ed. McGraw-Hill, New York, NY
81. Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27:379–423 and 623–656
82. Shannon CE (1949) Communication in the presence of noise. Proc IRE 37(1):10–21
83. Shannon CE (1959) Coding theorems for a discrete source with a fidelity criteria. In: IRE Nat. Conv. Record. 7:142–163.
84. Sheikh HR, Bovik AC (2006) Image information and visual quality. IEEE Trans Image Process 15(2):430–444
85. Sheikh HR, Sabir MF, Bovik AC (2006) A statistical evaluation of recent full reference image quality assessment algorithms. IEEE Trans Image Process 15(11):3440–3451

86. Simoncelli EP, Freeman WT, Adelson EH et al (1992) Shiftable multiscale transforms. IEEE Trans Inform Theory 38(2):587–607
87. Strela V, Portilla J, Simoncelli E (2000) Image denoising using a local Gaussian scale mixture model in the wavelet domain. Proc SPIE 4119:363–371
88. Tan DM, Tan C-S, Wu HR (2010) Perceptual colour image coder with JPEG2000. IEEE Trans Image Process 19(2):374–383
89. Tan DM, Wu D (2013) Perceptually lossless and perceptually enhanced image compression system & method, Patent Appl No:WO2013/063638 A2. WIPO, Geneva, Switzerland
90. Tan DM, Wu HR, Yu Z (2004) Perceptual coding of digital monochrome images. IEEE Signal Process Lett 11(2):239–242
91. Tanenbaum AS (2003) Computer networks, 4th ed. Pearson Education Inc., Upper Saddle River, NJ
92. Teo PT, Heeger DJ (1994) Perceptual image distortion. In: Proc IEEE ICIP 1994 2:982–986
93. van den Branden Lambrecht CJ (1996) Perceptual models and architectures for video coding applications. Ph.D. dissertation, Swiss Federal Inst Technol, Zurich, Switzerland
94. Vetro A, Tourapis AM, Müller K et al (2011) 3D-TV content storage and transmission. IEEE Trans Broadcast 57(2):384–394
95. Vetterli M, Kovačević J (1995) Wavelets and subband coding. Prentice-Hall, Englewood Cliffs, NJ
96. Visual Information Systems Research Group (1995) A methodology for imaging system design and evaluation. Sarnoff Corporation, Princeton, NJ
97. Visual Information Systems Research Group (1997) Sarnoff JND vision model algorithm description and testing. Sarnoff Corporation, Princeton, NJ
98. Wainwright MJ, Simoncelli EP, Wilsky and AS (2001) Random cascades on wavelet trees and their use in analyzing and modeling natural images. Appl Comput Harmon Anal 11:89–123
99. Wandell BA (1995) Foundations of vision. Sinauer, Sunderland, MA
100. Wang Z, Bovik AC (2009) Mean squared error: Love it or leave it. IEEE Signal Process Mag 26(1):98–117
101. Wang Z, Bovik AC, Sheikh HR et al (2004) Image quality assessment: From error visibility to structural similarity. IEEE Trans Image Process 13(4):600–612
102. Wang Z, Li Q (2007) Video quality assessment using a statistical model of human visual speed perception. J Opt Soc Amer A 24(12):B61–B69
103. Wang Z, Li Q (2011) Information content weighting for perceptual image quality assessment. IEEE Trans Image Process 20(5):1185–1198
104. Wang Z, Lu L, Bovik AC (2004) Video quality assessment based on structural distortion measurement. Signal Process Image Commun 19(2):121–132
105. Watson AB (1987) The cortex transform: Rapid computation of simulated neural images. Comput Vision Graphics Image Process 39:311–327
106. Watson AB (1989) Receptive fields and visual representations. In: Proc SPIE 1077:190–197
107. Watson AB (1993) DCTune: A technique for visual optimization of DCT quantization matrices for individual images. In: Soc Inf Display Dig Tech Papers XXIV:946–949
108. Watson AB (ed) (1993) Digital images and human vision. MIT Press, Cambridge, MA
109. Watson AB, Solomon JA (1997) A model of visual contrast gain control and pattern masking. J Opt Soc Amer A 14(9):2379–2391
110. Watson AB, Taylor M, Borthwick R (1997) Image quality and entropy masking. In: Proc SPIE Int Soc Opt Eng 3016:2–12
111. Webster AA, Jones CT, Pinson MH et al (1993) An objective video quality assessment system based on human perception. In: Proc. SPIE-Human Vision, Visual Process Digital Display IV. 1913:15–26
112. Wei Z, Ngan KN (2009) Spatio-temporal just noticeable distortion profile for grey scale image/video in DCT domain. IEEE Trans Circuits Syst Video Technol 19(3):337–346
113. Winkler S (1999) A perceptual distortion metric for digital color video. In: Proc SPIE Human Vision and Electronic Imaging IV 3644:175–184

114. Winkler S, Min D (2013) Stereo/multiview picture quality: Overview and recent advances. Signal Process Image Commun 28:1358–1373
115. Winkler S, Mohandas P (2008) The evolution of video quality measurement: From PSNR to hybrid metrics. IEEE Trans Broadcast 54(3):660–668
116. Wu D, Tan DM, Baird M et al (2006) Perceptually lossless medical image coding. IEEE Trans Med Imaging 25(3):335–344
117. Wu HR, Lin W, Ngan KN (2014) Rate-perceptual-distortion optimisation (RpDO) based picture coding. In: Proc DSP 2014:777–782
118. Wu HR, Reibman AR, Lin W et al (2013) Perception-based visual signal compression and transmission. Proc IEEE 101(9):2025–2043
119. Wu HR, Rao KR (eds) (2006) Digital video image quality and perceptual coding. CRC Press, Boca Raton, FL
120. Wu HR, Yu Z, Qiu B (2002) Multiple reference impairment scale subjective assessment method for digital video. In: Proc DSP2002 1:185–189
121. Wu J, Lin W, Shi G (2014) Structural uncertainty based just noticeable difference estimation. In: Proc DSP2014
122. Yang XK, Lin WS, Lu ZK et al (2005) Just noticeable distortion model and its applications in video coding. Signal Process Image Commun 20(7):662–680
123. Yang F, Wan S (2012) Bitstream-based quality assessment for networked video: A review. IEEE Commun Mag 50(11):203–209
124. Yang F, Wan S, Xie Q et al (2010) No-reference quality assessment for networked video via primary analysis of bit stream. IEEE Trans Circuits Syst Video Technol 20(11):1544–1554
125. Yu Z, Wu HR, Winkler S et al (2002) Objective assessment of blocking artifacts for digital video with a vision model. Proc IEEE 90(1):154–169
126. Yuen M, Wu HR (1998) A survey of hybrid MC/DPCM/DCT video coding distortions. Signal Process 70:247–278
127. Zhang X, Wandell BA (1998) Color image fidelity metrics evaluated using image distortion maps. Signal Process 70(3):201–214
128. Zhang L, Zhang L, Mou X (2011) FSIM: A Feature Similarity Index for Image Quality Assessment. IEEE Trans Image Process 20(8):2378–2386
129. Zilly F, Kluger J, Fauff P (2011) Production rules for stereo acquisition. Proc IEEE 99(4):590–606
130. Zujovic J, Pappas TN, Neuhoff DL (2013) Structural texture similarity metrics for image analysis and retrieval. IEEE Trans Image Process 22(7):2545–2558

Chapter 2
How Passive Image Viewers Became Active Multimedia Users

New Trends and Recent Advances in Subjective Assessment of Quality of Experience

Judith A. Redi, Yi Zhu, Huib de Ridder, and Ingrid Heynderickx

2.1 Introduction

Billions of digital images and videos are produced, broadcasted, shared, and enjoyed by users every day. Especially with the advent of Internet-based image and video delivery, the amount of multimedia content consumed every day has dramatically increased [24], and will continue to grow in the foreseeable future. This enormous amount of information needs to be handled (i.e., captured, stored, transmitted, retrieved, and delivered) in a way that meets the end-users' expectations. However, technology still shows limitations, such as limited spatial, temporal, and bit rate resolution in displays, bandwidth and storage constraints introducing compression related artifacts, or error-prone transmission channels resulting in network related artifacts. As a result, multimedia material is often delivered affected by impairments which disrupt the overall appearance of the visual content. Impairments provoke a sense of dissatisfaction in the user [54, 129, 175], which, in turn, may decrease the willingness to pay for/use the multimedia application, service, or device [161].

J.A. Redi (✉) • Y. Zhu
Department of Intelligent Systems, Delft University of Technology, Delft, The Netherlands
e-mail: j.a.redi@tudelft.nl; Y.Zhu-1@tudelft.nl

H. de Ridder
Department of Industrial Design, Delft University of Technology, Delft, The Netherlands
e-mail: H.deRidder@tudelft.nl

I. Heynderickx
Department of Industrial Engineering and Innovation Sciences, Eindhoven University of Technology, Eindhoven, The Netherlands

Philips Research Laboratories, Eindhoven, The Netherlands
e-mail: I.E.J.Heynderickx@tue.nl

© Springer International Publishing Switzerland 2015
C. Deng et al. (eds.), *Visual Signal Quality Assessment*,
DOI 10.1007/978-3-319-10368-6_2

That's why, in the last three decades, a lot of effort has been devoted to the development of technologies that can either prevent the appearance of impairments, or repair for it when needed. Following initial attempts based on the quantification of signal errors [53], it became soon clear that a better understanding of how humans experience images and videos was necessary to properly optimize media delivery. As a result, multimedia delivery optimization was researched from its early days through collaboration between engineers and vision scientists. In fact, this community can be considered a pioneer in user-centered multimedia design and engineering (for an accurate historical overview, see [21]). Within this effort, dedicated psychometric techniques were developed [3, 40, 81] and standardized [73, 82, 137, 138] to support a reliable quantification of visual quality (i.e., the perceived overall degree of excellence of the image, [40]) from a subjective point of view. With these techniques a large body of psychophysical data was collected to unveil the perceptual functions of the human visual system (HVS) that regulate the sensitivity to impairments. The outcome of these experiments served as inspiration for designing objective visual quality assessment metrics [61, 105], whose output would steer then impairment concealment (i.e., image/video restoration) and technology tuning.

It is interesting to point out that the common, underlying assumption for those studies is that having an understanding (possibly a model) of the perceptual processes that regulate impairment sensitivity suffices to predict the impairments' annoyance. In practice, being able to measure impairment sensitivity is considered to be substantially equivalent to predicting the overall quality of the viewing experience. This impairment-centric definition of visual quality (also referred to as perceptual quality in the following) has yielded remarkable results [61, 105]. Still, large room for improvement exists [114, 144]. Furthermore, new imaging and media technologies are challenging this impairment-centric notion of visual quality. Visual media are nowadays consumed in more and more immersive contexts (e.g., 3DTV, virtual and augmented reality) or in social, interactive, and customizable contexts (e.g., social media, video on demand, mobile). The judge of the visual experience cannot be regarded as a mere passive observer anymore, but rather as an active user interacting with the systems on the basis of specific expectations from them. In such a scenario, impairment sensitivity cannot be expected to be the sole factor contributing to the final user satisfaction on viewing experience.

In fact, several models have been proposed during the last decade that attempted at expanding the concept of visual quality to a more encompassing idea of quality of the (viewing) experience (QoE) [51, 81, 99, 128, 133, 144, 149]. In general, QoE is defined as a multidimensional quantity, depending on a number of attributes or features (i.e., quantifiable properties of the viewing experience, such as blockiness, aesthetic appeal, subject uniqueness), which are not necessarily mutually independent. Of these features, only a subset addresses traditional impairment sensitivity issues; others take into account rather cognitive and affective aspects of the experience. Attributes of the experience can be in turn influenced by external factors (i.e., factors independent of the media visualization) such as context of usage, user background, personality or task. Indeed, it has been recently shown that elements such as context of fruition [79] or user affective state [180] have an impact on visual quality appreciation, actually compensating in some cases for visual impairments. For example, football fans were shown to be highly tolerant to low frame-rates, as long as they were watching a football video [126].

Unfortunately, despite a working framework for QoE seems to be established, neither agreement has been reached on a precise taxonomy of attributes and external factors, nor much knowledge has been developed on how these quantities are interrelated. As a result, more subjective studies are needed to unveil interdependencies of QoE attributes and external factors, towards defining a precise model of how these elements concur to the final QoE judgment. In addition, integration with qualitative and quantitative user study techniques developed for other fields (e.g., human computer interaction) and including existing results from image psychology [47] are needed to fully characterize visual experiences.

In the following, we review the steps that led, throughout the last few decades, to the evolution of the concept of visual quality (typically, impairment-centric) into that of quality of experience (QoE). We first summarize the research done to quantify visual quality and impairment acceptability in the fields of display, signal processing, and network optimization (Sect. 2.2). We then review in Sect. 2.3 the models that over the years have attempted at extending the impairment-centric conception of visual quality, finally converging into an operative definition of QoE, which takes into account also the influence of external factors on the final user satisfaction. The existing knowledge on these factors and their impact on QoE is summarized in Sect. 2.4. Finally, Sect. 2.5 outlines new trends in subjective assessment of QoE, discussing in more detail QoE of immersive imaging technologies (such as stereoscopic displays), unveiling the role of affective processes in QoE judgments, and pointing out a methodological shift from lab-based to real-world- and crowd-based subjective experiments.

2.2 Subjective Assessment of Visual Quality

Within the past decades, visual impairments produced by technological limitations (e.g., lossy compression, sub-optimal pixel size, unreliable network transmissions) have been for long considered the principal cause of user dissatisfaction with multimedia systems; as a result, subjective assessment studies have mainly focused on quantifying the annoyance of such impairments as a function of technology variables.

A framework that supported this research was Engeldrum's image quality circle (IQC), depicted in Fig. 2.1 [40]. Such framework aimed at providing an effective methodology for linking experienced visual quality to the setting of technological variables of a multimedia system. In the case of displays, technological variables of interest were, for example, pixel size, color filter thickness, driving voltages, etc; when dealing with processing algorithms (e.g., compression or sharpness enhancement), such variables could be identified as the relevant parameters in these algorithms; in the field of network optimization for video streaming, bandwidth

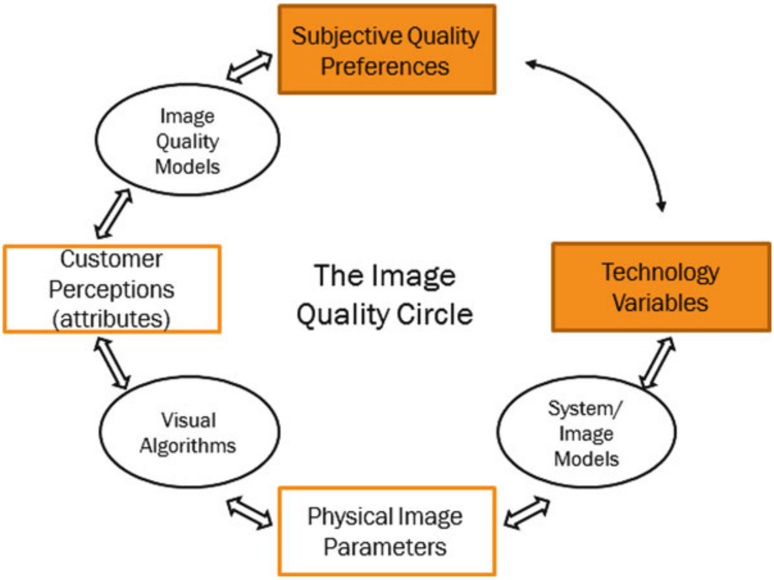


Fig. 2.1 Engeldrum's image quality circle

allocation was a main technological constraint. By varying the setting of technological variables, the eventual quality of the media delivery is affected. As a result, a holy grail for the multimedia community was (and still is) to infer relationships that, given a change in technological variables, accurately predict the experienced quality of the delivered media. In introducing the IQC, Engeldrum [40] argued that this problem was ill-posed: it aimed at linking a multi-dimensional description of a system (through its technological variables) to a one-dimensional overall quality preference. This relationship is not unique to begin with, and unveiling it requires an almost endless trial-and-error approach (i.e., a subjective test for every single change in a technological variable, which is costly and ineffective). Thus, rather than directly modeling the relationship between the technological variables and overall quality preferences, the IQC proposes a divide-and-conquer approach, involving three intermediate steps: (1) linking overall image quality to the (often unconsciously weighted) combination of underlying perceived attributes of the image, (2) linking each image attribute to the physical characteristics of the system output, and (3) linking the physical description of the system output to the system technological variables. By defining these three intermediate relationships, simulation and more accurate prediction of the effect that variations in technological variables have on the eventual perceived visual quality are allowed, limiting the need for subjective testing during system development.

Initially designed for display optimization, it is possible to adapt the IQC framework to the type of multimedia system under consideration (Fig. 2.2). Although rarely applied in practice, the IQC building blocks can be easily translated from

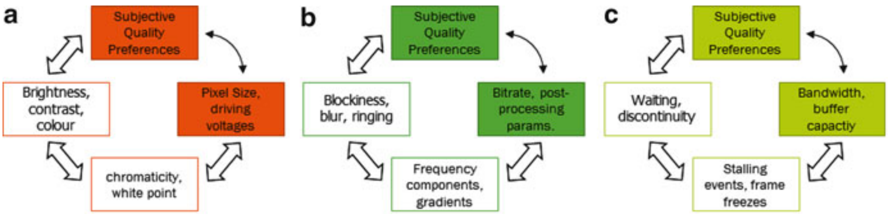

a
Subjective Quality Preferences
Brightness, contrast, colour
Pixel Size, driving voltages
chromaticity, white point
b
Subjective Quality Preferences
Blockiness, blur, ringing
Bitrate, post-processing params.
Frequency components, gradients
c
Subjective Quality Preferences
Waiting, discontinuity
Bandwidth, buffer capactiy
Stalling events, frame freezes

Fig. 2.2 Engeldrum's image quality circle adapted to different problem domains: display quality assessment (**a**), signal processing algorithm optimization (**b**) and network parameter optimization (**c**)

a display context to a signal or network context. Indeed, also technological signal variables can be related to physical characteristics of the light output of the multimedia system, resulting in partly different (e.g., compression artifacts such as blockiness and ringing) and partly similar (e.g., blur and noise) perceptual attributes, yielding an overall quality preference. Similarly, in a network context, physical media characteristics can be identified from technological quality of service (QoS) parameters (e.g., packet-loss ratio or number of stalling events), which result in attributes (such as jerkiness or perceived waiting time for the video to load/progress) that are weighted towards an overall quality preference. In the following, we review the efforts done within the display, signal processing, and networking communities to unveil relationships between technological variables, physical output characteristics, related features or attributes and the eventual user preference in terms of visual quality.

2.2.1 Visual Quality Preference in Displays

The impact of display related artifacts on the experience of viewers has commonly been associated with the concept of image quality, being a well-recognized concept to consumers [40], who considered it the key driver in selecting one device over the competitor's. The RaPID method first [9] and the IQC later constituted a solid framework to improve display quality. When applied to displays, a few issues needed to be addressed to make the IQC framework useful in practice: first of all, relevant image quality attributes, i.e., specific perceptual characteristics of the image, needed to be defined; then, a strategy to combine these attributes in a single overall quality measure needed to be determined.

With respect to the first issue, the RaPID method, more than the IQC model, contributed to determining unambiguous descriptions for the image perceptual attributes. These descriptions arose from a two-step procedure. First, expert (trained) viewers would discuss in team sessions which attributes characterized the image quality for a given set of images, the meaning (appearance) of these attributes and a meaningful way to quantify them. Naive viewers (a sample of standard customers) were then requested to score the image quality of the same set of stimuli. The relation with image quality was then established by means of a multivariate analysis and regression of the expert-defined attributes onto the overall quality scores. Applied to spatial resolution scaling on LCD monitors, for example, this method revealed that within this context image quality was a weighted sum of perceived blur, perceived pixellation (i.e., the fact that blocks of pixels were visible on diagonal lines), aliasing (geometrical deformations of spatially high frequency patterns), artifacts in letters (i.e., missing parts in letters/text), and increased ringing visibility [171]. An alternative approach to establish the attributes, focusing particularly on naive viewers, was described as the interpretation based quality (IBQ) method [134]. This approach combined qualitative (i.e., free image sorting and interviewing) and quantitative (i.e., magnitude estimation) methodologies, and as such allowed establishing the relationship between subjective preferences and the underlying features/attributes. The method has especially added value in detecting content and context dependency in high-quality images.

The second crucial issue is the combination of judgments of perceptual attributes into a single image quality value. Historically, several attempts have been made at creating a model of quality judgment based on assessment of image degradations and impairments. Allnatt's "Law of Subjective Addition" [3] was the first major achievement in this sense. This "law" stated that impairment annoyance was inversely proportional to image quality and that distinct impairments added up when they occurred simultaneously, to then linearly map into image quality. Minkowski metrics were later proved to be a better way of combining simultaneously occurring impairments into one overall impairment score [29,30]. This relationship was shown to hold also for extreme levels of impairment (i.e., such that the image content was almost unrecognizable) [193], as in the case of overlapping blur due to low-pass filtering using a 2D separable binomial filter [120] and noisiness due to normally distributed spatial noise. Finally, in a refined version of the Minkowski metrics accommodating an upper bound for impairment [30] it was shown that setting the exponent to approximately 2 yielded a good description of both quality and impairment judgments. In other words, it was concluded that the accumulation of perceptually distinguishable impairments could be described by a vector-sum model. As a result, the Minkowski metric provided a valuable tool for combining perceived display quality attributes into overall image quality [41, 121].

de Ridder et al. [149] pointed out later that, although successfully combining annoyance of different artifacts into an overall quality score, Minkowski metrics assumed image quality preference to be context and application independent. To verify whether this assumption was correct, they suggested to consider image quality as an indicator of the degree to which an observer could exploit images, i.e., to regard images "... not as signals but instead as carriers of information ..." [74]. This information-processing oriented definition assessed the degree of identifiability (i.e., the naturalness constraint) and discriminability (i.e., usefulness constraint) of the elements in the image. Fulfillment of both requirements would ensure high quality of the image; nevertheless, the degree to which constraints were expected to be fulfilled was highly content and task dependent [75]. In a series of experiments aimed at establishing color rendering preference in images, Janssen et al. manipulated the color characteristics of natural images, and then asked a pool of observers to judge their image quality [31, 33, 75]. Interestingly, image quality (Q) turned out to be a weighted sum of perceived naturalness (N) and colorfulness (C), or

$$Q = w \times N + (1 - w) \times C \tag{2.1}$$

where w is a weighting factor between zero and one. Moreover, it was observed that, when varying color contrast, participants showed a clear preference for more colorful, yet slightly unnatural images. Similar observations were made in a later study for the perceived quality of stereoscopic images where, under certain conditions, the quality of depth was found to be a weighted sum of naturalness of depth and perceived strength of depth [71, 72].

The implications of the findings above were that (1) image quality is typically judged based on a comparison of the experienced quality to an internal "reference" image, and that (2) this internal reference is not necessarily the most realistic one (i.e., a high fidelity reproduction of reality) but is rather influenced by other factors such as memory, content, and context of usage. These conclusions were later integrated in the so-called FUN model [39, 149]. In essence, this model assumes the existence of three major constraints determining image quality: Fidelity (i.e., degree of apparent match with an external reference, e.g., an original), Usefulness (i.e., degree of visibility of details), and Naturalness (i.e., degree of apparent match with an internal reference, e.g., memory colors). Overall image quality is then modeled as a weighted sum of the three constraints whereby the weighting depends on task, context, image content, etc. In fact, different people, different types of images, and different tasks may require different combinations of these weights, which implies that there is no single standard criterion for image quality, nor absolute perceptual preferences. It is interesting to note that this conclusion fits remarkably well with the ideas behind the so-called interface theory of human perception [63, 88], which states that perception is not about accurately reconstructing the physical world, but about constructing the properties and categories of an organism's perceptual world. Hoffman [63] argues that these perceptual structures are not intended to accurately match the physical world but, instead, are fast, intention-driven explorations of the meaningless physical world in preparation of "optically guided potential behavior," thus striving for utility and efficiency, not veridicality.

2.2.2 Visual Quality Preference of Processed Signals

The signal processing community evolved in many aspects rather distinct from the display community, and until recently its researchers largely focused on a signal fidelity approach. The goal of this branch of subjective studies was to understand the impact on perceived quality of a specific type of processing algorithm (e.g. compression, scaling, de-noising, and so on) towards identifying the optimized setting of the algorithm's parameters to produce the best visual quality.

In 1987, Watson [179] made an important distinction between perceptually lossless and perceptually lossy image coding, thus acknowledging the relevance of understanding and modeling the impact of coding artifacts on perceived image quality. Research into image integrity [145] or artifactual quality [81] referring to the relation between coding artifacts and quality preference became an important focus of the image and video processing community (for a thorough overview, see [21]). The approach consisted of incorporating models of low-level features of the HVS into image quality metrics. Subjective studies were therefore aimed at generating ground-truth data, and with that determining which HVS mechanisms were triggered by the appearance of impairments, leading to the identification and modeling of, e.g., contrast [8, 60, 160] and luminance masking mechanisms [106,127], spatial pooling strategies [178], and image structure perception [48,177]. Temporal and movement effects on artifact visibility have also been studied [122, 164]. Furthermore, initial attempts to understand the perceptual impact of overlapping video signal impairments (i.e., co-presence of e.g., blur, blockiness and noise) were carried out by Farias et al. [45], concluding Minkowski metrics were a powerful modeling tool in this context, as already proven for displays [44].

The large body of work done on artifact visibility estimation was inspired by the idea that the HVS remained constant over time [21], i.e., despite personal preferences, our visual processing strategies have barely evolved over the course of human history, thus it should be possible to model them in a meaningful and objective way (that is, independent of individual subject differences) such that HVS-based models could accurately predict and describe image quality. Interestingly, this soon turned out to be not true, even at threshold level. In an experiment on visibility of compression artifacts [182] it was shown that non-expert observers were less sensitive to compression-related artifacts (i.e., blockiness, blur, and quantization noise) than trained observers with detailed knowledge of the algorithm employed. Even more interestingly, it was shown that subjects who actually developed the algorithm were very sensitive to artifact visibility only in the images that were used within the algorithm design and test phases. Apparently, the well-informed experts knew exactly where to look for the impairments, a finding that cannot be accommodated by most of the low-level HVS-based models.

An initial attempt at studying the role of higher-level HVS features in signal impairment annoyance and related quality appreciation targeted visual attention mechanisms [42, 143]. When observing a scene, the human eye optimizes the information acquisition by focusing on specific, meaningful areas of the scene, and neglecting poorly informative areas [35]. As a consequence, it was hypothesized that signal impairments located in the visually attractive areas of an image were more likely to be noticed during the visual experience, resulting in a more negative judgment of visual quality (or higher annoyance). Evidence of this has been provided for, e.g., blocking artifacts in images [2]. As a consequence, the interplay between visual attention and visual quality assessment mechanisms was thoroughly studied. Research showed that the visual quality assessment task had a significant impact on visual attention deployment [123], which was also found to be the case for image aesthetic appeal assessment [145]. These findings stressed the importance of having control, task-free eye-tracking recordings to fairly evaluate the impact of signal impairment appearance on visual attention, and the impact of visual attention on the eventual quality judgment. Several eye-tracking studies reported information in this sense, yet without a clear consensus. In the work by Vuori and others [174] the quality of the judged image was shown to have an impact on the saccades' duration. In [35] the authors showed that saliency maps of pristine images obtained from free-looking eye-tracking data were poorly correlated to the maps derived from the image quality scoring of slightly impaired versions of the same images. This correlation was shown to increase with the amount of impairment visible in the images, and to be independent on the type of signal impairment. Vu et al. [173] identified instead an effect of the type of signal impairment (i.e., blur, compression, or noise) on the location of the fixations while scoring, though without quantifying it. As far as videos are concerned, Le Meur et al. [100] found that the quality evaluation task had a more limited impact in the video domain than in the image domain. Later, though, Mantel et al. [111] showed that the strength of signal impairments had an impact on the dispersion of the fixations (i.e., increasing with decreasing video quality) and was positively correlated with the duration of the fixations [111].

Despite the diversity of the abovementioned results, the study of visual attention in relation to signal impairment annoyance enabled the design of a wide range of image and video quality metrics, enhanced with either saliency or visual importance data (for a complete overview, see [42]). The added value of incorporating such information in quality metrics was clearly shown for images [107, 142], whereas it was found to be less relevant for video [42]. Furthermore, this activity produced an abundance of subjective data, most of which have been made publicly available for further research [184]. These data may be precious in further understanding the role of high-level HVS mechanisms in viewing experience appreciation.

2.2.3 Subjective Assessment of Network-Related Impairments

Nowadays quality of the (broadband) broadcasted or stored video content and of the displays used for their rendering is in most circumstances of such a high level that naive consumers hardly see improvements. The latter, however, is not yet true for multimedia content distributed over (mobile) IP networks. Bandwidth limitations, along with network unreliability (i.e., the possibility of losing parts of the streamed signal/packets) can cause impairments during visualization of the image/video, including frame freezes, deformations of the spatial and temporal structure of the content, and long stalling times.

Rather than based on subjective assessment of visual quality, network parameters have been for long optimized towards keeping an acceptable QoS, by taking into account parameters such as packet loss ratio, delay, jitter and available bandwidth [153], as well as video QoS parameters, such as buffering time and buffer ratio [5]. Lately, researchers have been aiming at correlating the QoS parameters to QoE measurements (typically, identified once again with visual quality subjective ratings [138]) by using fitting functions [49, 84, 158]. In general, low QoS performance leads to low QoE [70]. For example, it has been shown that the buffer ratio (i.e., the fraction of time spent in buffering over the total session time, including playing plus buffering) consistently had a high impact on user QoE [37]. Reduced buffering times resulted in higher user satisfaction. Similar conclusions were found for other QoS parameters, such as the join time in multicast video delivery, the buffering duration, the rate of buffering events, the average bit-rate, and the packet loss rate [70, 113].

In general, QoS metrics succeed in estimating QoE from a network efficiency point of view, but they do not necessarily reflect the overall viewing experience. In fact, QoS parameters fail in capturing all subjective aspects associated with the viewing experience [34, 131]. Note that typically QoS parameters are computed based on the encoded bit-stream, whereas no pixel information is analyzed; therefore, the impact of signal impairments such as blockiness and blur (see Sect. 2.2.2) is not taken into account in these approaches. In the case of packet loss, for example, it was found that the same packet loss ratio yielded more or less annoyance depending on the video encoding and video content [162]. The loss of bit-stream packets indeed can result in specific, spatiotemporal visual impairments, due to the poor concealment of the lost packet at the bit-stream decoder side. This type of impairments is more or less noticeable depending on the amount of movement in the video and can be annoying [140, 162] even more so when in combination with strong compression artifacts [80]. Furthermore, when studied in conjunction with visual attention, packet loss artifacts have been shown to be more annoying when located in visually important regions of the image [43], and to have a high potential for becoming salient, then altering the natural visual attention deployment [140]. Quite interestingly, the entity of this alteration has been shown to be negatively correlated with the perceived visual quality of the video [140].

2.3 From Visual Quality to Quality of (Viewing) Experience

As pointed out in Sect. 2.2, subjective studies from different communities (displays, signal processing, and networking) converged eventually towards a similar conclusion: quantifying impairment sensitivity, even by means of accurate HVS models, is necessary yet not sufficient to quantify the overall quality of the viewing experience. In fact, a few models have been proposed throughout the last decade to extend the impairment-centric notion of visual quality to a broader, more representative concept of quality of the viewing experience.

Keelan [81] defined visual quality as a multidimensional quantity evolving along a number of visual attributes, comparable to Engeldrum's IQC attributes. Keelan distinguished four different families of attributes: artifactual (e.g., blockiness and blurriness), preferential (e.g., brightness and contrast), aesthetic (e.g., symmetry or harmony [46]), and personal (e.g., user emotional connection and engagement with the visual content [90,180]). Of those, the first two were highly related to perceptual quality, whereas the latter two would contribute to the visual quality assessment by taking into account more implicit experiences of the viewer [110]. Because of this, aesthetic and personal attributes were considered "too subjective" and "unlikely to yield to an objective description" of an image, leading Keelan to the conclusion that their quantification would be "too cumbersome and expensive to use for routine image quality research" ([81], p. 6). As a result, Keelan privileged the investigation of artifactual and preferential attributes, leaving unexplained the contribution of personal and aesthetic attributes to the overall visual quality.

Ghinea and Thomas also attempted at reaching a more encompassing definition of visual quality by proposing the concept of quality of perception (QoP) [51]. Their reasoning started from the assumption that multimedia are primarily consumed for infotainment; therefore, viewing experience has a twofold purpose: that of transferring information to the user, and that of granting a sufficiently high level of satisfaction in terms of entertainment. To properly optimize viewing experience, then, both (1) the level of Information Assimilation (QoP-IA) and (2) the overall user satisfaction with respect to the media presentation (QoP-S) should be taken into account. QoP-IA represents the level of the user's understanding of the media content. It is typically measured as the performance (in terms of number of correct responses) on a questionnaire about the (semantic) content of the viewed media. QoP-S depends instead on two elements. The subjective level of quality (QoP-LoQ) measures the perceptual impact of losses on visual quality (e.g., due to the appearance of impairments), independent of the media content. The level of enjoyment (QoP-LoE) measures instead the overall enjoyability of the media presentation, taking into account also cognitive and affective aspects of the visual experience, such as watchability, ease of understanding, and level of interest in the subject matter. Throughout multiple studies [57, 58, 105] it was found that low QoP-LoQ did not impact the level of information assimilation (i.e., despite the presence of impairment, users could still fully understand the content of the media) and had limited impact on the level of enjoyment (QoP-LoE), indicating that other elements compensated for artifact appearance in viewing experience. Unfortunately, up to date, there is little known and studied on these elements that compensate for artifact visibility in overall viewing experience: it is not known yet which are these elements and how they contribute to the eventual experience appreciation.

From a different perspective, Pereira [128] proposed a three-level model for visual experience appreciation. In Pereira's model, visual experience is first evaluated at the sensorial level, which responds to the purely physical properties of the media (i.e., comparable to the IQC physical image characteristics and to some extent to Keelan's artifactual and preferential attributes). This level of evaluation contributes therefore to the first perceptual quality impression (a concept similar to QoP-LoQ [51]). Next, the viewing experience is evaluated at the "perceptual" level. Here, the media content and the potential for creating knowledge out of it are assessed (with similarities to QoP-IA in [51]). Note that the word "perceptual" here entails also cognitive processes such as content recognition and interpretation, and is not to be confused with the classic notion of perceptual quality, which in this framework is addressed at the sensorial level. Finally, the viewing experience is assessed at the emotional level, where the way in which the media impacts the user's affective state is evaluated. Pereira suggests that viewing experience appreciation results from a linear combination of a quantification of these three aspects; however, he did not provide an empirical validation of this hypothesis. Pereira's model was further extended in [133] to assess augmented reality experiences. The extended model accounted also for implicit experiences of the user (such as cultural background), and the context and goal of the viewing experience. As a result, on top of the three levels in the model of Pereira [128], the authors of [133] suggest to take into account both usability of the multimedia system and ethnographical assessment to obtain an accurate measure of the quality of the augmented reality experience.

The FUN model of de Ridder and Endrikhovski [149], already mentioned in Sect. 2.2, can also be considered a milestone in the road that took visual quality to evolve into QoE. The model is the first to introduce the concept of finality of usage of media (and user motivation for having the viewing experience), and to suggest that viewing experience cannot be quantified without taking this concept into account. The quality of a viewing experience, indeed, should depend on the degree to which the visual information can be successfully exploited by the user towards his/her goal. This in turn is quantified in terms of the fulfillment of the Fidelity, Usefulness, and Naturalness criteria already described in Sect. 2.2.2. Whereas the Fidelity criterion can be to a large extent equated to the assessment of impairment sensitivity, the Usefulness and Naturalness criteria introduce two rather new concepts in QoE evaluation. The Usefulness constraint indicates the maximum discriminability of perceived items in the image (or video); thus, the degree to which this criterion should be fulfilled is highly application and task-dependent, as, for example, the fulfillment threshold for Usefulness of a consumer display is different from that of a microscope. The Naturalness constraint refers instead to the fidelity of the media to what the authors call an "internal-reference," or an internal representation of how the media "should look like." Here, previous (quality) experiences and expectations come into play. Thus, the fulfillment threshold for Naturalness is intrinsically user-dependent. The paradigm shift in this model lays in the fact that constraints are not anymore assumed to be universal, but rather application and user dependent. Hence, factors external to the viewing experience (i.e., not directly related to vision) have an impact on its appreciation, and should be studied in relation to it.

This idea has been recently picked up by the Qualinet consortium, which has proposed a rather encompassing model for Quality of multimedia Experience [99]. Note that the experience here is not limited to vision, but is multisensory, thus it is not necessarily related to imaging systems only. In fact, the Qualinet model combines elements of all models described above: to begin with, QoE is described as a multidimensional quality, that can be decomposed in a set of perceptual attributes called features. QoE features are defined as "perceivable, recognized and namable characteristics of the individual's experience of a [multimedia] service which contributes to its quality" and can be classified into four categories: features at the level of perception, at the level of interaction, at the level of usage, and at the level of service. The features at the perceptual level entail experience characteristics that can be evaluated from immediate perception (e.g., blockiness, blurriness, brightness, and contrast). The features at the interaction level account for human–technology interaction aspects of the experience (e.g., responsiveness and communication efficiency between the user and the multimedia system). Features at the level of usage assess the accessibility and the stability of the service during usage. Finally, the long-term characteristics of the service beyond the single instance of usage, such as ergonomics, usability, and ease of use, are accounted for by the level of service, similarly to what was suggested in [133]. All these features are assumed not to be independent. Features appreciation is in turn mediated by a set of interrelated quantities called Influence Factors (IF). Influence Factors are defined as "characteristics of a user, system, service, application, or context whose actual state or setting may have influence on the QoE for the user." As such, they pre-exist the fruition of the media; nevertheless, they condition the final user satisfaction. IFs can be grouped into three categories, depending on whether they represent properties of the user, of the system (or application or service) or of the context of usage. User IFs entail characteristics of the user such as demographics, personality or emotional state, and can condition both the appreciation of technical quality (thus modulating the features of the level of perception) and that of the overall experience, also impacting on the interpretation and understanding of the media content. System IFs are those properties of the multimedia system/service that are responsible for the resulting technical quality: media encoding configuration, network parameters, display functions, etc. Finally, context IFs encompass all situational properties of the environment in which the experience takes place. Examples of context IFs are location and space, time of the day, task and social context. For a detailed overview of known influencing factors of QoE, see Sect. 2.4.

Similarly to the FUN model [149], the Qualinet model [99] also assumes the existence of an internal "reference" experience to which the real one is compared. All QoE features have an internal reference value that is modulated by IFs; the extent to which the features of the current experience match the reference ones builds the eventual user satisfaction. Although mapping the interactions of (reference and current) QoE features and influence factors is still beyond reach, a simplified representation of the model is attempted in Fig. 2.3. In this figure, the model is visualized as a network of computing units that modulate the assessment of the difference between the reference and current (i.e., "quality") feature values. Influencing factors not only determine the value of the "quality" (experienced) features, but also modulate the importance that the difference between their value and that of the internal reference has on the final quality judgment (fusion module). To draw a parallel with the models reviewed so far, we can consider the level of perception similar to the artifactual attributes in [81], the QoP-LoQ in [51], and the Fidelity dimension in [149]; they all consider the impact of system IFs on perceptual features (green unit in the figure). The emotional level of [128] could result from the impact of Human and Context IFs on the level of perception and the level of service

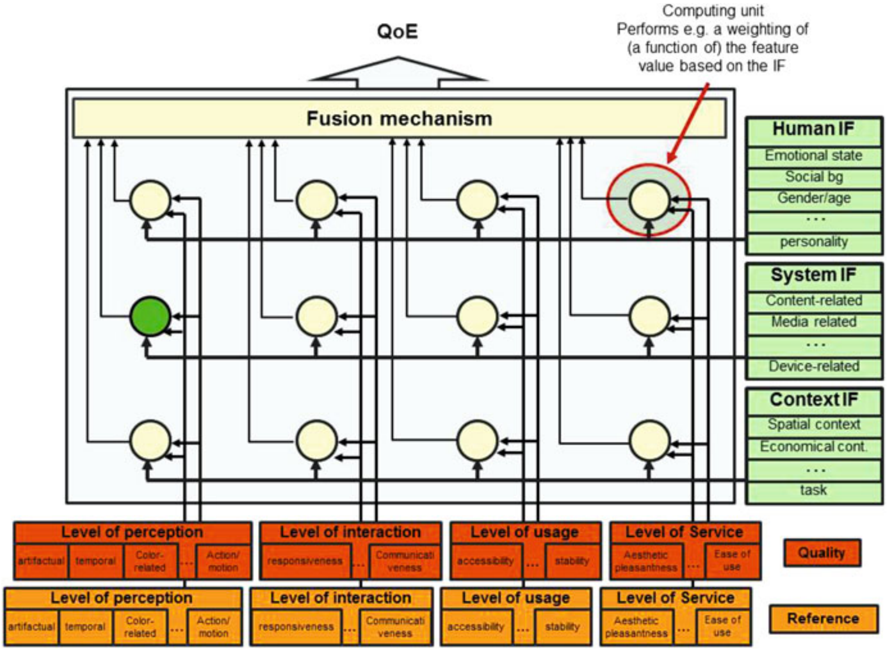


Fig. 2.3 Schematic representation of the Qualinet model for quality of experience [99, 144], fully connected into a network for the final QoE prediction. The *green* computing unit covers the influence of system factors (technology variables [81]) on the features of the level of perception; most research in the QoE domain has been focusing so far on this specific facet of QoE

features. Finally, the Usefulness dimension in [149] could be intended as the result of context IFs on the level of perception features. Some of these interactions are further explored in Sect. 2.4, but we first want to clarify an operative definition of QoE that we will adopt throughout the rest of this chapter.

2.3.1 Definition of QoE

The concept of QoE arose from the field of Telecommunication Engineering. In the past decades, the effectiveness of communication services was linked to the notion of QoS, which is defined as the “totality of characteristics of a telecommunication service that bears on its ability to satisfy stated and implied needs of the user of the service” [139]. QoS is mainly operationalized in terms of system and network performance-related measures (e.g., packet loss ratio, jitter, or delay). However, with the booming of online multimedia services, the notion of QoS has started showing its limitations, and was found to be poorly correlated to user satisfaction. As a result, the QoE concept emerged, and was initially defined by ITU [167] as “the overall acceptability of an application or service, as perceived subjectively by an end-user." This definition suggests that the scope of QoE has shifted from a rather narrow perspective of telecommunication systems to a broader perspective of multimedia services. Furthermore, this definition not only takes the complete end-to-end system in consideration to define QoE, but also includes the user's expectations and his/her context. Recently, the Qualinet White Paper [99] proposed a more explicit definition of QoE: "Quality of Experience (QoE) is the degree of delight or annoyance of the user of an application or service. It results from the fulfillment of his or her expectations with respect to the utility and/or enjoyment of the application or service in the light of the user's personality and current state." This second definition explicitly refers to the concepts of "personality," entailing long-term traits of the user such as feelings, thinking attitude, and behavior (as per [130]), as well as user "current state", i.e., the punctual set of feelings, thoughts and behavior contextual to the viewing experience [99]. Note that the current state is both an influencing factor of QoE and a consequence of the experience. Although both definitions describe a similar phenomenon, the definition given by Qualinet seems to be more complete than the one of ITU-T. In the ITU-T definition, QoE is related to acceptability in terms of the "characteristic of a service describing how readily a person will use the service." The Qualinet definition, instead, emphasizes that human factors, such as personality and current state, may significantly influence QoE. Given the evidence of the importance of such factors in properly estimating user satisfaction and QoE (which will be documented in Sect. 2.4 of this paper), we adopt the Qualinet definition as operational definition of QoE in the remainder of this chapter. Along with this operative definition, it is worth mentioning a few other concepts that closely relate to QoE. The term engagement, for example, often refers to positive aspects of user experience. Attfield [6] gives a definition of engagement as "a quality of the user experience that emphasizes the positive aspects of interaction—in particular the fact of being captivated by a resource." In prior research, engagement was described as the experience of a user who highly focused on the video and was affectively involved with it [153]. Studies showed that engagement played a crucial role in determining user satisfaction [37]. As a result, in Sect. 2.4 we also refer to literature on engagement with multimedia content to complement the knowledge existing on factors influencing QoE. Finally, it is interesting to relate QoE to the concept of endurability. The term endurability has been used to describe the consequence of a satisfactory experience, and specifically the likelihood of remembering it and the willingness to repeat it or recommend it [124]. Read et al. [136] studied endurability in children's satisfaction with tourist attractions. They organized a group of 45 children going on a school trip to a themed tourist attraction, where they could engage in nine activities. Children scored each activity before and after performing it, and ticked "yes, maybe, no" in response to the question "Would you like to do it again?" A week after the task, children were asked to recall the separate activities that had made up the event. They were also asked to name the activity that they had liked the best. The results showed that 81% of the children recalled the activity that they had previously identified as worth repeating. This suggests that high QoE leads to endurability [124]: people remember enjoyable, useful, engaging experiences and want to repeat them.

2.4 Influencing Factors of QoE

As introduced in Sect. 2.3, QoE is a multifaceted quality, resulting from the interaction of multiple influencing factors. Besides factors that have been proven to influence QoE in the context of video, we also review elements of the experience found to be relevant in other research fields (e.g., psychology of gaming experience). It is important to remark that most of these factors are not independent. They may interact with each other, and as such, influence QoE in a complex way (see also Fig. 2.3). Following the model proposed in [99], the factors are arranged into three categories (shown in Table 2.1), namely System factors, User factors, and Contextual factors. Each group of factors is described in more detail in the remainder of this section.

2.4.1 System Factors

System factors are to a large extent comparable to Engeldrum's technological variables, and include all those characteristics of the system (or application or service) that contribute to determine the "technically produced quality" [78] of the eventual media presentation. As such, they also determine the presence of impairments in it. In the most general formulation, system factors can address characteristics of the device on which a video is viewed (e.g., a mobile phone, PC, tablet or television), of the technological signal variables (i.e., the video format or parameters in signal processing algorithms) and of the network configuration (i.e., the so-called QoS parameters). Each of these contributions to QoE is discussed here in some detail (although not fully encompassing all the existing literature).

Table 2.1 Factors influencing QoE discussed in this chapter

System factors (Sect. 2.4.1)	User factors (Sect. 2.4.2)	Contextual factors (Sect. 2.4.3)
Devices [85–87, 153]	Interest [67, 90, 104, 110, 124, 126, 159]	Physical environment [161, 183]
Signal and network variables [1, 56, 103, 175, 194]	Personality [26, 180]	Economic conditions [14, 83, 190]
	Age/gender [12, 13, 69, 117, 118, 185]	Social motivation [10, 16, 18, 23, 25, 50, 64, 91, 101, 108, 115, 124, 125, 150, 151]
	Affect/mood [124, 180]	

2.4.1.1 Devices

Nowadays, users watch videos through a diversity of devices. Studies showed that user acceptability of video quality varied with the type of device used to watch the video [85–87]. For example, See-To et al. [153] showed that user's QoE of the same video was significantly different on a desktop than on a mobile device. User expectations for mobile device performance were lower than for desktop performance, and as a consequence, a higher QoE for a video with the same amount of introduced impairments was found on the mobile device than on the desktop. In more general terms, the impact of display technology variables on image and video quality was investigated thoroughly during the last decades, at least with respect to artifactual quality [40, 81]. Pixel size and arrangement, static and dynamic contrast, white point and color gamut, motion blur and other motion artifacts, response characteristics and flicker, and finally viewing angle range are all elements known to highly impact perceived image quality [19, 93, 165]. The extension from artifactual quality to QoE is only limitedly addressed for device optimization. In 3D displays, stereoscopic depth has been shown to increase the appreciation for the viewing experience (see Sect. 2.5.1) but at the same time, disparity may generate visual discomfort [94]. McCarthy et al. [112] reported that a decrease of display resolution yielded user dissatisfaction.

2.4.1.2 Signal and Network Variables

Functional characteristics of video streaming (e.g., frame rate, resolution, and encoding) directly influence users' QoE [194], as also already discussed in Sect. 2.2.3. Gulliver et al. [56] conducted a subjective experiment to investigate the impact of different multimedia frame rates on the user's (impairment-centric) visual quality. Participants were asked to view video clips at different frame rates, and to answer for each clip some questions evaluating whether the participants understood the video content, and to give per clip an overall quality score and a score for the level of enjoyment. The results showed that the assimilation of video information was not significantly affected by frame rate, but the user's perceived visual quality and enjoyment were. In other words, higher frame rates improve overall user enjoyment and quality perception. Similar aspects have been investigated within the context of scalable video coding (SVC). The SVC specification of the H.264 coding scheme [1] adapts the video stream along the temporal, spatial, and signal-to-noise ratio (SNR) dimensions to obtain an optimal trade-off between frame-rate, spatial resolution, and (spatial) impairment visibility. For a given spatial resolution, the optimal trade-off between temporal and SNR quantization is known to depend, among other factors, on motion [175]. For fast motion videos, a decrease in SNR is preferred over a loss in smoothness resulting from low frame-rates. For static videos, the opposite happens. The trade-off between spatial resolution and frame-rate has instead been shown to depend on the video bitrate, with a preference for large spatial resolution at low bitrates (<800 kbps) [103].

2.4.2 User Factors

A user factor is defined as "any variant and invariant characteristic of a human user" that influences the viewing experience, such as demographic, personality, or interest related characteristics [99]. User factors determine for a large part the user "current state" mentioned in the QoE definition reported in Sect. 2.3. These factors were largely overlooked for a long time, because they were judged too difficult to quantify in both a subjective and objective way [81]. Nonetheless, lately researchers have started investigating them more systematically, also thanks to the large amount of personal data made available by the users themselves in Social Media. We review in the following some of the main findings with respect to the influence of user factors to QoE.

2.4.2.1 Interest

In psychology literature, interest has been considered as an emotion. Silvia [159] suggested that interest comes from two appraisals: novelty and coping potential. Novelty is the tendency to seek elements that are new, or unusual in one's environment, and evoke in the user a sense of curiosity. Huang et al. [67] showed that incorporating novel elements into a website attracted curious users and brought out enjoyable experiences. Coping potential is the ability to understand unfamiliar, complex objects, and as such is strongly user-dependent [159]. In the field of aesthetic appreciation of art, it has been shown, for example, that abstract, unfamiliar works of art were poorly appreciated by the average user [104]; nevertheless, the stronger the background art knowledge of the user, the higher the aesthetic appreciation would get [110]. O'Brien [124] indicated that QoE was often triggered when something resonated with a user's interest. Kortum and Sullivan [90] employed a total of 100 participants and 180 movie clips encoded at nine compression levels from 550 kbps up to DVD quality. After viewing the clips, participants were asked to rate the (impairment-related) visual quality and desirability of the movie content. The results showed a general increase in quality rating as the desire for content increased, at a given bitrate. Thus, personal interest in the video significantly influenced user judgments [90]. Palhais et al. [126] used videos of sport events, encoded in four different bitrate/resolution combinations. Participants chose three sports that they were more interested in and three sports that they liked less. Then they watched all videos and rated the (impairment-related) visual quality at the end of each video. The results demonstrated that the interest level had a strong influence on the subjective assessment of the visual quality: users tended to value a video with the same bitrate as higher in QoE when they were more interested in the content of the video.

2.4.2.2 Personality

Personality is “the particular combination of emotional, attitudinal, and behavioral response patterns of an individual.” One of the effective ways to determine personality is the five-factor model (FFM) [26] or “Big Five,” consisting of the dimensions openness (i.e., degree of intellectual curiosity and creativity), conscientiousness (i.e., tendency to show self-discipline), extraversion (i.e., the level of orientation towards other people), agreeableness (i.e., the tendency to be compassionate and cooperative), and neuroticism (i.e., the tendency to experience unpleasant emotions easily). Wechsung et al. [180] conducted an experiment asking 33 participants to perform a series of tasks (such as play, pause, and stop) in front of an IP-TV. The results showed that the personality of participants influenced their performance. For example, neuroticism was negatively correlated with performance, while agreeableness enhanced it. In contrast, the results also indicated no correlation of personality with impairment annoyance.

2.4.2.3 Age/Gender

Evidence exists that age influences QoE. Wolters et al. [185] found older adults to be more critical than younger users, which may suggest that elderly people have higher requirements for QoE. However, Naumann et al. [118] observed the opposite: they found that older users tended to rate the (impairment-centric) visual quality more positively than younger users. As far as gender is concerned, little has been done to investigate its effect on QoE appreciation. Males and females are known to react differently to emotional pictures [13] and to have different perception of olfactory and visual media synchronization [117]. As a consequence, it is reasonable to expect that optimal QoE settings may depend on gender too. An initial evidence in this sense can be found in [69]: within the context of 3D audio telephony and teleconferencing services, it was found that males and females have different preferences, in terms of experienced QoE, with respect to the size of the room where the teleconference is held. Closer to the field of image quality Campanella Bracken [12] showed that women viewing a video clip on either an HDTV or a (at that time) more standard resolution NTSC TV reported more perceived realism than men, which may imply that women evaluate at least part of the television content as more real than men.

2.4.2.4 Affect/Mood

During the interaction with online (video) services users may experience positive or negative affective states (or moods). The interaction itself may induce such a state, but it is also possible that people are already in a particular affective state, such that it may impact the way they experience the interaction with the video service. Positive affective states relate to enjoyment, satisfaction, and fun. For example, a lack of fun can act as a barrier to shop online or enjoyment during a webcast can draw the user in [124]. Negative affective states, such as frustration, anxiety, and boredom, may lead to low QoE. For example, participants that feel frustration towards a technology report lower QoE ratings [180].

2.4.3 Contextual Factors

Contextual factors describe all aspects of the environment within which the user consumes the media, e.g., physical location, economical aspects, or social context. The following are considered prominent contextual factors for QoE evaluation.

2.4.3.1 Physical Environment

Many aspects of the physical environment may affect QoE; these aspects may range from characteristics of the seating position (e.g., viewing distance and viewing height) to disturbances that occur in the environment a viewer is in. Viewing distance is a balancing act between two aspects: a shorter viewing distance increases the field of view, and makes the viewer more involved with the content, but may make impairment better visible as well [183]. Staelens et al. [161] investigated subjective quality under viewing conditions, in which television is typically watched. The results showed that the interruption of phone calls and SMS alerts could prevent a person from getting engaged into a video, which would result in low QoE.

2.4.3.2 Economic Aspects

The economic aspects relate to key concepts of marketing, such as the product and brand strategy, the pricing strategy, the positioning of the product in the market, and the market segmentation and identification of target groups [14]. These aspects are closely related to the notion of Quality of Customer Experience, introduced by Kilkki [83, 123]. He indicated that the economic aspects of a product or service, such as price and brand, can have a high impact on QoE, also due to customer loyalty (think about, e.g., Apple). According to [190], there is a positive correlation between the willingness to pay for a multimedia product/service and the (impairment-centric) visual quality of the video offered to the user. The study clearly showed that users were inclined to pay less if they were offered a video with a lower visual quality. When users felt they were overpaying for their service with regard to the quality they experienced, they reacted in different ways, all eventually leading to a decrease in revenues for the operator of those services.

2.4.3.3 Social Context

Social context refers to the fact that a user is affected by the interaction with a group of other people [151], being family, friends, or even strangers. In the past, many studies have reported the social impact of traditional TV watching [50, 91, 101]. Co-located co-viewing is a rather common way of consuming the more traditional media, such as TV programs [115], having great potential as a social activity and conversational topic [108]. Co-viewing when enjoying each other's company can increase user's overall satisfaction [115]. Recently, a concept of social TV—as implemented in "Amigo TV"—has emerged: it provides multiple viewers with a joint TV watching experience by adding communication features via audio conferencing, graphic symbols, and avatars [23,25]. User studies of social TV have confirmed the high acceptance of such technology, because it allows users to communicate with friends even when they are not physically co-located [125, 150]. Far less is known about the impact on QoE of a newer concept of social context, namely the one arising from recommendations and opinions (e.g., Facebook "likes") of friends and/or strangers. Watching online videos is not so often done by multiple people sitting together, nonetheless, a sort of social influence (or pressure) may manifest itself through the opinions of peers or friends gathered via, e.g., social networks. Having input from peers or friends is indeed already very common on shopping websites or in gaming communities. For example, O'Brien [124] noted that in one of his studies an interviewed person mentioned reading book reviews from "certain reviewers that I know I can trust that have similar taste to me" [124]. In other words, this interviewed person created his own social context to support him in deciding which books to read. In the gaming industry, social interaction is explicitly designed in the game. Lively virtual societies are built around multiplayer online games (e.g., World of Warcraft), and these games are highly successful [10, 16]. Several studies even claimed that digital games also can increase social interaction in gamers' real life (e.g., they talk to friends about the game strategy) [18]. Either playing games together or watching others play a game can bring enjoyment to gamers [64]. Although very little has been studied so far, all these studies point towards the impression that social context may strongly impact QoE.

2.5 Beyond Visual Quality: New Trends in Subjective QoE Assessment

Despite the body of work on influencing factors of QoE described in Sect. 2.4, it is clear that unveiling a reliable model of user QoE preference is still beyond reach. The profound transformation that media consumption underwent in the last decade opens countless questions and applications in which influencing factors and features of the viewing experience still have to be determined. In facing this major challenge, subjective assessments represent the core instrument to learn more about the interplay between perceptual, cognitive, and affective mechanisms that underlie the appreciation of viewing experience. Nevertheless, subjective QoE assessment methodologies need now to be integrated with knowledge developed from traditionally very different fields (e.g., human computer interaction, affective computing, behavioral psychology, but also media production, computer graphics, and lighting design) to overcome the traditional impairment sensitivity paradigm. We identify three major directions in which the subjective QoE assessment community should seek for the paradigm shift: the technological one, the psychological one, and the methodological one.

2.5.1 Beyond the Traditional Screen Technology: QoE of Immersive Viewing Experience

Imaging technologies are evolving quickly. In recent years, new display technologies have been developed that provide a more immersive viewing experience by enhancing specific experience features. High dynamic range (HDR) displays, for example, magnify perceived contrast by means of different backlight technologies and the usage of a larger number of bits to represent luminance information [4, 154]. Similarly, stereoscopic and autostereoscopic displays (3D) enhance perceived depth [119, 137], and upcoming 4k and 8k devices display images at a ultra-high resolution [17]. These features come with an undoubted added value for the viewing experience. Nevertheless, to ensure the full enjoyment of the enhanced experience, immersive technologies need optimization both at a display and at a signal level. In the case of HDR imaging for example, problems such as optimal design of the backlight dimming algorithm [89] as well as of tone mapping operators that can display a HDR image on a regular display [191] are still under investigation. Furthermore, to drive the optimization of such immersive technologies, it is essential to (1) properly understand the impact of an enhanced dimension on the eventual QoE and (2) assess whether such attribute enhancement modifies the impact of other attributes on QoE. In the following, we attempt at exemplifying why these two points are crucial for QoE optimization, by looking at that technology among the aforementioned ones that was most thoroughly studied in the last decade, i.e., stereoscopic displays.

Although introduced halfway the twentieth century, 3D displays became accessible to the general public within the last decade. Initially based on anaglyph projectors in movie theaters, stereoscopic display technology underwent major optimization efforts to finally enter consumers' living rooms. Initially, optimization was again driven by the concept of "image quality," of which a large body of knowledge was already available from 2D displays. Soon enough, it became clear that this concept was insufficient to properly quantify user satisfaction with respect to the overall experience provided by 3D displays. When asked to assess image/video quality of a stereoscopic display, users limited their judgment to the annoyance of

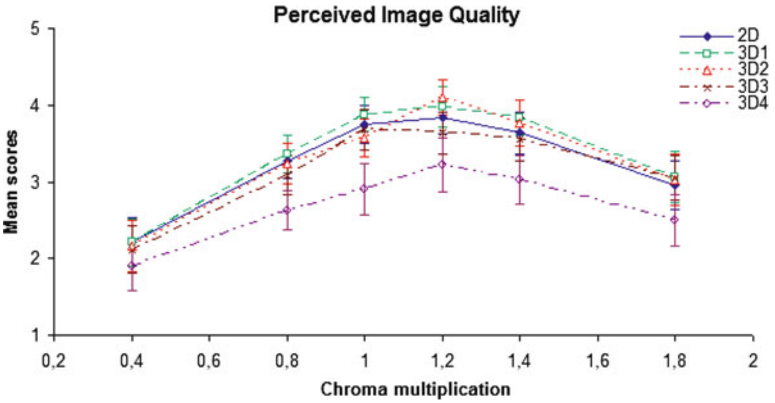


Fig. 2.4 Illustration of the effect of impairments (chroma multiplication) and stereoscopic depth on image quality scores for chroma affected images [92]. The different lines indicate mean scores for 2D content (*solid line*), and 3D content with a disparity of 1 mm (3D1), 2 mm (3D2), 3 mm (3D3), and 4 mm (3D4), respectively

the impairments introduced by the technology, not valuing the experience created by the stereoscopic depth [92, 95, 155, 163]. Seuntiens [155] showed, for example, how image quality scores decreased due to visible impairments as a consequence of compression similarly in 2D and 3D displays. A similar effect was found for blur, even for different camera-base distances [95]. Kuijsters [92] also showed that increasing stereoscopic depth in images of various levels of colorfulness didn't affect or slightly reduced the perceived image quality (as shown in Fig. 2.4).

At no point in the abovementioned results users judged instead the (added) value of the increased depth on the experience. Hence, to allow manufacturers to optimize, and where needed balance, the full experience of stereoscopic content, a higher level concept was needed. The concepts of naturalness, already introduced in the FUN model described in Sect. 2.3 [149], as well as that of "viewing experience" [157] were investigated to cover the user's perceptual and cognitive experience of stereoscopic displays. Based on the series of experiments described above, a higher level evaluation criterion EC was defined and modelled as a combination of image quality IQ and perceived depth D, i.e.,:

$$EC = \alpha \times IQ + \beta \times D \tag{2.2}$$

When using naturalness as EC, it was found to incorporate depth information for about 25 % (while 75 % of the judgment still depended on image quality), assessment of viewing experience consisted instead for about 82 % out of image quality and for about 18 % out of stereoscopic depth [95].

The importance of naturalness as an evaluation criterion for stereoscopic displays is in line with results from applying the IBQ method to stereoscopic content assessments. Häkkinen [59] showed that for many viewers stereoscopy changed the life-likeness of the content, although in some cases stereoscopy also introduced artificiality or "unrealness," depending on the specific content. Partly based on all these findings, the ITU [137] defined a new recommendation for the subjective evaluation of 3D-TV systems, prescribing that viewers should score three factors separately, i.e., picture quality, perceived depth, and visual comfort.

Immersive viewing technologies are these days evolving even beyond the display. Virtual and Augmented Reality technologies, for example, are currently used for a multitude of applications, ranging from Mental Health Computing [168] to Google Glasses [7], and an understanding of QoE with respect to those highly immersive contexts has to be achieved. Immersive experiences are also evolving by incorporating other types of technologies, traditionally uncorrelated to multimedia delivery. Starting with the Philips Ambilight TV for example, LED lights were incorporated in the TV display to increase the field of view and to give therefore viewers a more cinematic experience, also reducing the eye fatigue. As for stereoscopic displays, image quality was found to relate only to visual impairment, neglecting the added value of the light effect. The term viewing experience, on the contrary, was proven to cover both image quality and the added value of Ambilight, also in combination with increased depth (i.e., when mounted on stereoscopic displays) [156].

It is reasonable to expect that viewing experience will even further extend beyond the display in the future. Solid state lighting (SSL) technology, for example, is already being used to embed low-resolution "displays" in our whole environment. Because of their fast-switching and spectrally tuneable characteristics, LEDs can be spatially distributed and embedded in (semi)-transparent materials: this allows designers to present information on walls, floors, and/or ceilings around us. Although still used mainly towards functional purposes, we can foresee for the near future that such technology will provide enhanced entertainment experiences, based on the creation of (affective) atmospheres through the combination of visual content, sound and lighting. Ideas to use atmospheric light in combination with video or games (via scripting) or with music (e.g., by associating colors to terms from the lyrics) are indeed emerging. Nevertheless, assessment of the full experience of these systems will require—most probably—a new higher level concept, overarching aspects of image quality, sound quality, and light experience [172].

2.5.2 Beyond Perception: The Role of Aesthetics and Emotion in QoE Appreciation

A second important evolution in subjective QoE assessment is the inclusion of affective evaluations within QoE measurements. As mentioned in Sects. 2.3 and 2.4, the affective state of the user (i.e., his/her mood or specific emotional state) may impact the way a viewing experience is appreciated [180]. In turn, the potential for the viewing experience to impact on the affective state of the user (e.g., increase the arousal of the emotion as well as improve its valence) should be considered in QoE assessment paradigms. Efforts in this direction are currently growing (e.g., [135, 148, 181]), also reaching out to the affective computing community. Nevertheless, major challenges are ahead. First, appropriate methodologies to measure the affective impact of media in relation to QoE have yet to be determined. Both self-reporting instruments such as the Self-Assessment Manikin [96] or the Affect Button [15], and more "objective" tools such as physiological measurements (e.g., EEG or skin conductance) are being evaluated at the moment. Existing results are however scattered and more structured efforts are needed to identify a pool of affective measurements that can complement existing standards [73, 82, 137, 138] for subjective QoE measurement. A second important challenge lays in the need to decouple, within the QoE judgment, the effect of affective states pre-existing the visual experience from that of the emotional state induced by the viewing experience itself. The ability of the media to induce emotion, creating an empathic experience with the video content, may be positively taken into account in user QoE judgments. A valuable tool to this end would be to use stimuli of which the emotional impact on the user can be controlled (e.g., IAPS emotional slides, or standardized excerpts from movies with the potential to induce specific mood states [55,97]). They could be used to induce specific moods prior to the visual experience, to investigate the impact of pre-existing mood on QoE judgments; also, they could constitute test stimuli for the viewing experience, to allow an understanding of how induced emotional states alter QoE. Nevertheless, mood induction practices have to be carefully designed in order to carry out experiments that are still ethically acceptable.

At the same time, it is interesting to research which properties of the image have the potential to impact the affective state of the user, along with information on the changes in arousal and valence of this affective shift. Color, for example, is well known for having an impact on people's mood, both from a psychological and a physiological point of view [169, 192]. Similarly, contrast or content arrangement of an image may generate changes in the mood state. Understanding and quantifying the relationship between physical properties of the image, their perception and their impact on the user affective state is therefore a key challenge for upcoming QoE research.

Some work in this sense has been carried out within the scope of understanding the aesthetic appeal of media. Aesthetic appreciation is generally recognized to be related to both perceptual and affective mechanisms, and it has been for long studied independently from the concept of Quality of the Visual Experience. Mastered for a long time by artists and then also addressed by psychologists [11], lately it has started to attract the attention of the media engineering community. Predicting the aesthetic appeal of images has become interesting especially towards improving information retrieval, computer graphics, and automatic management of image collections [77]. As a consequence, studies have been carried out first to identify image and user attributes impacting on aesthetic appeal and then to model them. Perceptual image features (as per [51], see Sect. 2.3) such as color saturation, brightness, and amount of details (i.e., texture and visual crowding) were found to contribute to the final aesthetic quality judgment [27, 76, 109].

In particular, the deployment of visual attention has been shown to be related to image clutter [20], in turn negatively correlated with image aesthetic appeal scores [146]. The same study showed that visual importance [176] is to some extent predictive of compliance of the image to photographical compositional rules, which in turn has a beneficial effect on aesthetic appeal. Impairment generated by specific media configuration (System Influence Factors, see Sect. 2.4.1) such as blockiness [145] have been shown to negatively affect aesthetic appeal. Similarly, user Influence Factors such as experience, cognitive bias, and personal opinions and memories [132] have been found to strongly condition the appreciation of the aesthetic experience. Correlation between aesthetic ratings and familiarity with the image subject has been reported in [110], and content recognizability (i.e., the level of abstraction of the content) has been shown to have an influence on aesthetic appeal in works of art [98].

Interestingly, very little work has been carried out in trying to link the aesthetic appeal of an image to the overall Quality of the visual Experience. Nevertheless, initial evidence exists that the aesthetic appeal of an image does influence not only QoE, but also the judgment in terms of annoyance of impairments presented in the image itself [144]. In this study, Redi asked a pool of participants to judge "integrity" (namely, the traditional, impairment-related concept of visual quality) of a set of images, including (i) a group of pristine images and (ii) a group of images derived by those in group (i) by applying JPEG compression to them. As a result, the images in this second group presented the same content as those in the first one, but affected by visible compression artifacts. The pristine images of group (i) had already been evaluated in terms of aesthetic appeal in a separate study [145]. Redi correlated then the integrity scores of both groups of images with the aesthetic appeal scores of the pristine ones. It was found that the integrity judgments of the pristine images (group i) were influenced by the level of aesthetic appeal of the image, and that the two quantities were negatively correlated (see Fig. 2.5a). Conversely, when impairments were present (group ii, Fig. 2.5b) integrity judgments increased as the aesthetic appeal increased. It should be mentioned that, in instructing the participants to score aesthetic appeal, the concept of integrity was also explicitly mentioned, and distinguished from aesthetic appeal. This may have primed participants in (unconsciously) taking into account integrity in their aesthetic appeal judgments, partially explaining the results found in [144]. To check this, we repeated the analysis performed in [144], but by using a different set of aesthetic appeal scores, obtained from study [146]. There, participants were again asked to score aesthetic appeal, but now without reference to image integrity. The consistency in aesthetic appeal scores between experiments [145, 146] turned out to be 0.68. Although still acceptable in terms of predictive power of one set of scores for the other, this number is far from correlations typically found across experiments for e.g. impairment annoyance scores (typically, ∼0.9). There are several possible reasons for this discrepancy: (1) the highly personal component of the aesthetic appeal judgment (since different participants were used in both experiments), (2) the difference in experimental protocol, or (3) range effects (since only a subset of the images of the first experiment were used in the second) [149]. Despite these possible deviations, also the results of the second experiment showed

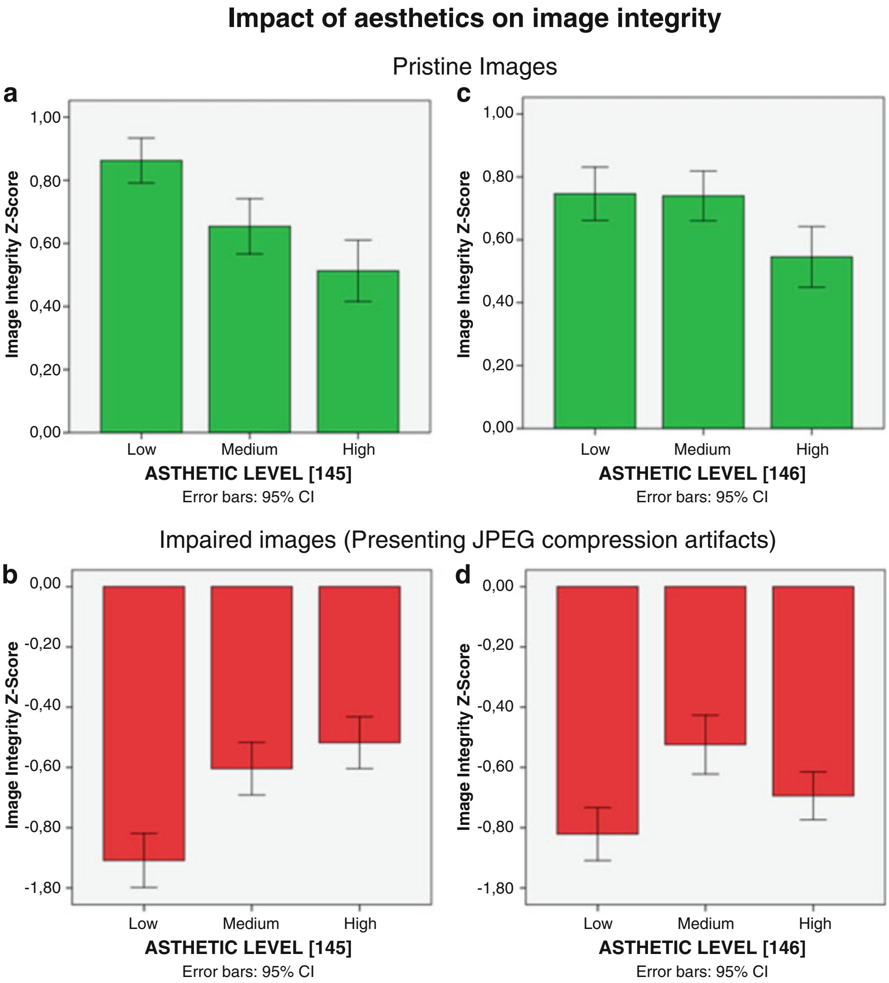


Fig. 2.5 Relationship between aesthetic appeal and image integrity [144–146]: dependency of integrity scores of pristine images on the aesthetic appeal level as assigned in experiment [145] (**a**) and in experiment [146] (**c**); dependency of integrity scores of impaired (JPEG-compressed) images on the aesthetic appeal level as assigned in experiment [145] (**b**) and in experiment [146] (**d**)

that the aesthetic appeal level had an impact on integrity [144] for both the pristine and impaired images. Pristine images that were highly aesthetically appealing were scored significantly lower in integrity than the others (Fig. 2.5c), whereas impaired images with a low aesthetic appeal were scored significantly lower than the others (Fig. 2.5d), confirming the trend identified in Fig. 2.5b. This consistency points out how aesthetic appeal plays a role in QoE and in tolerance to impairment; as a result, including aesthetic appeal information in QoE metrics may help in improving their accuracy and QoE optimization thereafter.

2.5.3 Beyond Lab-Based Studies: Methodological Shifts for Reliable QoE Quantification

Research on QoE has relied for long (and now more than ever) on determining user preferences with respect to the sensitivity to visual impairment through subjective studies. The main goal of subjective testing is to sort stimuli (i.e., media) according to their perceived properties or attributes [40] on a given scale.

Multiple psychometric methodologies have been developed with this purpose, and adapted for the measurement of QoE in standardized conditions [73, 82, 137, 138], and choosing the most appropriate one for a test is far from trivial. In discriminating among methodologies, Engeldrum [40] suggests to take into account aspects such as the confusion level within the set of stimuli and the effort required to the participant to complete his/her task. The confusion level is determined by how closely the test stimuli are spaced in quality. The narrower are the quality gaps among them, the higher is the probability of inducing confusion (disagreement, possibly inversions) in across-participants judgments. Methods able to accurately measure the quality of stimuli with high confusion (e.g., paired comparison) are typically unable to measure large quality gaps. The effort required to participants to complete their task depends on the number of judgments needed per stimulus, and hence it is related to the number of stimuli involved. Methods requiring a high number of judgments per stimulus are not suitable for experiments involving large datasets, as they could be prone to fatigue and learning errors. Other desirable properties of the methods, depending on the goal of the experiments, may be the minimization of inter-participant variability [62] or the robustness to range effects [32] (e.g., in case results of multiple, separate experiments need to be merged into a single set of data [141]).

2.5.3.1 Psychometric Methods for QoE Measurement

The Paired Comparison (PC) method [28, 166] is a classic psychometric technique that allows measuring distances among stimuli in terms of just noticeable differences (JNDs) [40]. The experimental procedure consists of asking subjects to compare each stimulus with all other stimuli in the set. As a result, even small differences between the stimuli can be detected. On the other hand, the judgment effort grows as the square of the number of stimuli, hence this number must be limited. Moreover, in analyzing the results, complications may arise due to the "zero and one problem" [116], or inconsistencies in the selected model [52]. Lately, the QoE has shown growing interest around PC, [102, 186, 187], and methods have been developed for establishing confidence intervals to the quality scores provided by the PC tests [186]. The double stimulus impairment scaling (DSIS) methodology [138] is also often chosen for the assessment of visual impairments. DSIS judgments are expressed on an interval scale (typically, a five-point categorical scale, ACR), as a (conscious) comparison of each impaired stimulus with its undistorted version.

Being a double stimulus method (i.e., reference and test stimuli are both shown during the judgment), the DSIS requires a moderate effort per judgment, but still allows the assessment of large datasets. A possible drawback of the method may be the categorical scale used for the assessment: the boundaries among categories (e.g., “good” and “fair”) are blurred and depend on the participant; this may result in low inter-participant agreement [40, 81]. The ACR scale is however to date the most used method for scaling stimuli, also in a Single Stimulus setting (i.e., without an explicit reference to be presented to the participant) [141]. Both DSIS and Single Stimulus scaling can be performed also with numerical scales, both discrete or continuous [68, 138]. In all these cases, the results of the tests are reported in terms of average score per stimulus (Mean Opinion Scores), expressed in the scale used for the experiment. These scores reflect human preference, though do not have a precise psychophysical meaning. Indeed, the obtained scores may vary with the definition of the scale [40], as well as with the quality range spanned by the stimuli (range effects [32]). This suggests that comparing results of different experiments may be problematic, possibly inducing inconsistencies when merging these data in a single, larger dataset.

Among classic scaling methodologies, The Quality Ruler (QR) method deserves a mention, as a middle-ground alternative between the direct scaling methodologies (DSIS, Single Stimulus) and PC. The QR method was first described by Keelan in [81], and subsequently adopted as an international ISO standard for psychometric experiments for image quality estimation [82]. The core idea of the QR method is to provide the participant with a set of reference images, anchored along a calibrated quality scale, to compare a test image with. The task of the participant is to find the reference image closest in quality to the test image by visual matching. Reference images (1) depict a single scene and vary in only one perceptual attribute (i.e., blur, blockiness, color saturation); (2) are closely spaced in quality, but altogether span a wide range of quality. They are presented in a way that easily allows detection of the quality difference between them, and their close spacing in quality should allow the participant to score with higher confidence, decreasing the risk of inversions and range effects. In practice, participants perform several comparisons reference-test stimuli to complete a single assessment, until they find the reference stimulus that matches the quality of the test one. The advantage of this procedure is that, as long as the referencer stimuli are kept the same, subjective scores obtained from a quality ruler experiment always refer to the ruler scale, and not to the quality range spanned by the test stimuli. This minimizes range effects. Furthermore, it has been shown that the visual matching procedure reduces inter-participant variability [141]. Unfortunately, this method has been successfully implemented for images [82, 141], but it is of hard applicability for video QoE assessment.

2.5.3.2 Subjective Testing Outside the Lab

To obtain reliable results, psychometric experiments have usually been performed in highly controlled, standardized environments [73, 82, 137, 138], typically within laboratory facilities. This allowed to control for lighting and viewing position, minimizing the effect of environmental contextual factors (see Sect. 2.4.3) and making visibility conditions homogeneous across participants.

The evolution of multimedia technology calls now for a shift in the traditional lab-based study paradigm. Since the advent of mobile technology (smartphones and tablets) and Internet-based video delivery, the visual experience is now consumed in very different environments, and should be studied within realistic usage conditions to be properly optimized [152, 161]. Furthermore, with the acceptance of the more encompassing definition of QoE presented in Sect. 3.3.1, where personal differences are taken into account and need to be understood, more attention should be given to the demographic composition of the pool of participants used in the subjective studies. This pool should be indeed as representative as possible of existing differences in terms of user Influencing Factors (Sect. 2.4.1). Recruiting such a diverse pool of participants may prove difficult, and the eventual amount of individuals to be involved in the study may explode, thus making a lab-based experiment unfeasible in terms of time consumption and cost. In this scenario, interest in using Crowdsourcing [66] for subjective tests of QoE has grown significantly [65]. Crowdsourcing was originally conceived to outsource small and repetitive tasks (so-called microtasks) to a multitude of people (so-called microworkers) who, online and for a small compensation, could perform these tasks in a time- and cost-effective way. Platforms such as Microworkers, Amazon Mechanical Turk, and CrowdFlower were created to facilitate the recruitment of microworkers, and soon enough it became clear that Crowdsourcing had an enormous potential for the performance of Human Intelligence Tasks (e.g., image labeling) in Multimedia research [36]. As a result, it became of interest for subjective QoE assessment as well. Compared to traditional subjective QoE assessment in a controlled lab environment, which is time-consuming and high-cost, Crowdsourcing tasks can be accomplished within a few minutes and do not require a long-term employment of the participants. More importantly, Crowdsourcing gives the opportunity to collect data from populations with very diverse demographics, enabling therefore the investigation of User Influencing Factors in QoE judgments [65, 170].

Nevertheless, Crowdsourcing still has some challenges ahead. Besides the reliability issues related to payment scheme, worker selection, and ease of sloppiness in carrying out the experimental task, which are extensively discussed in [65], there are a few other points that should be made. First, Crowdsourcing tasks should be fairly short (up to 10 min) to avoid boredom and unreliable behavior. Traditional QoE tests typically involve tens or hundreds of stimuli, requiring participants to score for much longer timespans (typically between 30 min and 1 h). Thus, to collect QoE scores for a large set of stimuli, experimenters usually have to decompose the scoring task in a set of smaller tasks (i.e., campaigns), each one including a subset of the stimuli. Although it is common practice to merge all QoE scores from different campaigns as if they were expressed on the same QoE scale, this may be a dangerous practice. Range as well as environmental effects [149] may occur, making the merging meaningless. A possible solution to this is to select a number of stimuli to be scored in all campaigns [147]. If carefully chosen (e.g., some with excellent quality, others with very low quality), these stimuli can function as anchors for the scoring scale, limiting range effects; furthermore, they can be used for realignment purposes.

It should also be considered that, by dividing the traditional design over many people, we intrinsically generate mixed-subjects designs. So, it may be the case that the measurements themselves lose in accuracy, because one cannot fully exploit within-subjects variance. An interesting solution to that has been recently proposed in [187, 188], which involves randomized paired comparison [38] to accommodate incomplete and imbalanced data. Paired Comparison has been found to be an effective methodology for measuring QoE via crowdsourcing, due to the simplicity of the task [102] and the availability of tools for the analysis of incomplete preference matrices. Furthermore, it has been shown to be a suitable methodology for embedding worker reliability checks [22, 102, 189]. This property is especially desirable for crowdsourcing, given that the trustworthiness of workers is often doubtful and that, due to the lack of supervision, workers responses may be inaccurate or erroneous (see also [65]). On the other hand, for scaling large sets of stimuli (in the order of hundreds), the applicability of paired comparison is still limited, as the number of pairs to be judged may be intractable also for such a far-outreaching methodology. Finally, although this may be partially compensated by the larger demographic spread, it should be considered that at the present stage Crowdsourcing attracts only a subset of the population, leaving out, e.g., elderly people (who, for the time being, cannot master the technology).

Conclusion

Although subjective assessment of QoE has been investigated for over 50 years, this field is in continuous expansion. In this chapter we documented the evolution of the impairment-sensitivity centric understanding of QoE to a more encompassing one, which takes into account both attributes of the experience and external influencing factors that modulate the user appreciation for a specific delivered media. In a world where media fruition is tightly related to social networks and social media, as well as to immersive but also mobile viewing systems, the user cannot be considered as a simple, passive observer anymore. Users are active agents which interact with the system, e.g., selecting the content and/or the modality with which they desire the media to be delivered. As a result, and as we documented in this chapter, elements such as visual semantics, user personality, preferences and intent, social and environmental context of media fruition also concur to the final experience assessment. While (theoretical) models of QoE appreciation exist that take these elements into account, in practice little is known, still, on how attributes of the experience combine into a final QoE judgment and how external factors influence this combination. Similarly, light still needs to be

(continued)

shed on the cognitive and affective processes that underlie viewing experience appreciation. Subjective quality assessment is therefore now more than ever a core element in pushing forward the research on QoE optimization. Empirical studies are needed to unveil and thoroughly understand the mechanisms underlying the interplay between attributes and influencing factors of viewing experiences, and existing as well as new methodologies will have to be deployed towards that goal. Methods designed for studying affect, human–machine interaction, as well as human cognitive processing will have to be adopted by the QoE community and integrated with existing subjective quality measurement tools to properly quantify viewing experiences in their context of usage. Moreover, these methods will have to be adapted to be deployed in large scale experiments that reach out to big amounts of users world-wide, by means of web-based technologies such as crowdsourcing. A different type of adaptation will also be necessary to perform subjective studies in less controllable but more realistic contexts of usage. Data analysis techniques will also have to be upgraded to allow the merging of heterogeneous data derived from lab-, real world-, and web-based studies. Content metadata as well as user/context information retrievable in Internet (e.g., social media profiles, or textual comments to media material) will also provide useful information to better understanding of user preferences. Finally, appropriate modeling tools will have to be deployed to construct, based on the abovementioned information, an accurate, encompassing model of Quality of Viewing Experience appreciation that could steer a better delivery of the richness of digital multimedia content that is available nowadays.

References

1. Nicola Adami, Alberto Signoroni, and Riccardo Leonardi. State-of-the-art and trends in scalable video compression with wavelet-based approaches. *Circuits and Systems for Video Technology, IEEE Transactions on*, 17(9):1238–1255, 2007.
2. Hani Alers, Judith Redi, Hantao Liu, and Ingrid Heynderickx. Studying the effect of optimizing image quality in salient regions at the expense of background content. *Journal of Electronic Imaging*, 22(4):043012–043012, 2013.
3. John Allnatt. *Transmitted-picture assessment.* Wiley Chichester, UK, 1983.
4. Munisamy Anandan. Progress of led backlights for lcds. *Journal of the Society for Information Display*, 16(2):287–310, 2008.
5. Javed Asghar, Francois Le Faucheur, and Ian Hood. Preserving video quality in iptv networks. *Broadcasting, IEEE Transactions on*, 55(2):386–395, 2009.
6. Simon Attfield, Gabriella Kazai, Mounia Lalmas, and Benjamin Piwowarski. Towards a science of user engagement (position paper). In *WSDM Workshop on User Modelling for Web Applications*, 2011.
7. Roberto Baldwin. Google glasses face serious hurdles, augmented-reality experts say. *Wired Magazine*, 2012.

8. Peter GJ Barten. *Contrast sensitivity of the human eye and its effects on image quality*, volume 72. SPIE press, 1999.
9. Soren Bech, Roelof Hamberg, Marco Nijenhuis, Kees Teunissen, Henny Looren de Jong, Paul Houben, and Sakti K Pramanik. Rapid perceptual image description (rapid) method. In *Electronic Imaging: Science & Technology*, pages 317–328. International Society for Optics and Photonics, 1996.
10. Steve Benford, Chris Greenhalgh, Tom Rodden, and James Pycock. Collaborative virtual environments. *Communications of the ACM*, 44(7):79–85, 2001.
11. Daniel E Berlyne. Aesthetics and psychobiology. 1971.
12. Cheryl Campanella Bracken. Presence and image quality: The case of high-definition television. *Media Psychology*, 7(2):191–205, 2005.
13. Margaret M Bradley, Maurizio Codispoti, Dean Sabatinelli, and Peter J Lang. Emotion and motivation ii: sex differences in picture processing. *Emotion*, 1(3):300, 2001.
14. Frances Brassington and Stephen Pettitt. *Principles of marketing*. FT Prentice Hall, 2005.
15. Joost Broekens and Willem-Paul Brinkman. Affectbutton: A method for reliable and valid affective self-report. *International Journal of Human-Computer Studies*, 71(6):641–667, 2013.
16. Barry Brown and Marek Bell. Play and sociability in there: Some lessons from online games for collaborative virtual environments. In *Avatars at Work and Play*, pages 227–245. Springer, 2006.
17. ITUR BT2020. Parameter values for ultra-high definition television systems for production and international program exchange, 2012.
18. Diane Carr, Gareth Schott, Andrew Burn, and David Buckingham. Doing game studies: A multi-method approach to the study of textuality, interactivity and narrative space. *Media International Australia, Incorporating Culture & Policy*, (110):19, 2004.
19. Joseph A Castellano. *Handbook of display technology*. Elsevier, 1992.
20. Cathleen D Cerosaletti, Alexander C Loui, and Andrew C Gallagher. Investigating two features of aesthetic perception in consumer photographic images: clutter and center. In *IS&T/SPIE Electronic Imaging*, pages 786507–786507. International Society for Optics and Photonics, 2011.
21. Damon M Chandler. Seven challenges in image quality assessment: past, present, and future research. *ISRN Signal Processing*, 2013, 2013.
22. Kuan-Ta Chen, Chen-Chi Wu, Yu-Chun Chang, and Chin-Laung Lei. A crowdsourceable qoe evaluation framework for multimedia content. In *Proceedings of the 17th ACM international conference on Multimedia*, pages 491–500. ACM, 2009.
23. Konstantinos Chorianopoulos and George Lekakos. Introduction to social tv: Enhancing the shared experience with interactive tv. *Intl. Journal of Human–Computer Interaction*, 24(2):113–120, 2008.
24. I Cisco. Cisco visual networking index: Forecast and methodology, 2011–2016. *CISCO White paper*, pages 2011–2016, 2012.
25. T Coppens, Frie Vanparijs, and K Handekyn. Amigotv: A social tv experience through triple-play convergence. *Alcatel Technology white paper*, 2005.
26. Paul T Costa and Robert R McCrae. The revised neo personality inventory (neo-pi-r). *The SAGE handbook of personality theory and assessment*, 2:179–198, 2008.
27. Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. Studying aesthetics in photographic images using a computational approach. In *Computer Vision–ECCV 2006*, pages 288–301. Springer, 2006.
28. Herbert Aron David. *The method of paired comparisons*, volume 12. DTIC Document, 1963.
29. Huib de Ridder. Subjective evaluation of scale-space image coding. In *Electronic Imaging'91, San Jose, CA*, pages 31–42. International Society for Optics and Photonics, 1991.
30. Huib de Ridder. Minkowski-metrics as a combination rule for digital-image-coding impairments. In *SPIE/IS&T 1992 Symposium on Electronic Imaging: Science and Technology*, pages 16–26. International Society for Optics and Photonics, 1992.

31. Huib de Ridder. Naturalness and image quality: saturation and lightness variation in color images of natural scenes. *Journal of imaging science and technology*, 40(6):487–493, 1996.
32. Huib de Ridder. Cognitive issues in image quality measurement. *Journal of Electronic Imaging*, 10(1):47–55, 2001.
33. Huib de Ridder, Frans JJ Blommaert, and Elena A Fedorovskaya. Naturalness and image quality: chroma and hue variation in color images of natural scenes. In *IS&T/SPIE's Symposium on Electronic Imaging: Science & Technology*, pages 51–61. International Society for Optics and Photonics, 1995.
34. Carlo Demichelis and Philip Chimento. Ip packet delay variation metric for ip performance metrics (ippm). 2002.
35. Robert Desimone and John Duncan. Neural mechanisms of selective visual attention. *Annual review of neuroscience*, 18(1):193–222, 1995.
36. Anhai Doan, Raghu Ramakrishnan, and Alon Y Halevy. Crowdsourcing systems on the world-wide web. *Communications of the ACM*, 54(4):86–96, 2011.
37. Florin Dobrian, Vyas Sekar, Asad Awan, Ion Stoica, Dilip Joseph, Aditya Ganjam, Jibin Zhan, and Hui Zhang. Understanding the impact of video quality on user engagement. *ACM SIGCOMM Computer Communication Review*, 41(4):362–373, 2011.
38. Alexander Eichhorn, Pengpeng Ni, and Ragnhild Eg. Randomised pair comparison: an economic and robust method for audiovisual quality assessment. In *Proceedings of the 20th international workshop on Network and operating systems support for digital audio and video*, pages 63–68. ACM, 2010.
39. S.N. Endrikhovskij. Image quality and colour categorization. *Colour image sci-ence: exploiting digital media*, pages 363–382, 2002.
40. Peter G Engeldrum. *Psychometric scaling: a toolkit for imaging systems development*. Imcotek Press, 2000.
41. Peter G Engeldrum. A theory of image quality: The image quality circle. *Journal of imaging science and technology*, 48(5):447–457, 2004.
42. Ulrich Engelke, Hagen Kaprykowsky, H Zepernick, and Patrick Ndjiki-Nya. Visual attention in quality assessment. *Signal Processing Magazine, IEEE*, 28(6):50–59, 2011.
43. Ulrich Engelke, Romuald Pepion, Patrick Le Callet, and Hans-Jürgen Zepernick. Linking distortion perception and visual saliency in h. 264/avc coded video containing packet loss. In *Visual Communications and Image Processing 2010*, pages 774406–774406. International Society for Optics and Photonics, 2010.
44. Mylene CQ Farias and Sanjit K Mitra. No-reference video quality metric based on artifact measurements. In *Image Processing, 2005. ICIP 2005. IEEE International Conference on*, volume 3, pages III–141. IEEE, 2005.
45. Mylène CQ Farias, Sanjit K Mitra, and John M Foley. Perceptual contributions of blocky, blurry and noisy artifacts to overall annoyance. In *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*, volume 1, pages I–529. IEEE, 2003.
46. Elena Fedorovskaya, Carman Neustaedter, and Wei Hao. Image harmony for consumer images. In *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*, pages 121–124. IEEE, 2008.
47. Elena A Fedorovskaya and Huib De Ridder. Subjective matters: from image quality to image psychology. In *IS&T/SPIE Electronic Imaging*, pages 86510O–86510O. International Society for Optics and Photonics, 2013.
48. Rony Ferzli and Lina J Karam. A no-reference objective image sharpness metric based on the notion of just noticeable blur (jnb). *Image Processing, IEEE Transactions on*, 18(4):717–728, 2009.
49. Markus Fiedler, Tobias Hossfeld, and Phuoc Tran-Gia. A generic quantitative relationship between quality of experience and quality of service. *Network, IEEE*, 24(2):36–41, 2010.
50. David Gauntlett and Annette Hill. *TV living: Television, culture and everyday life*. Routledge, 2002.
51. George Ghinea and JT Thomas. Quality of perception: user quality of service in multimedia presentations. *IEEE Transactions on Multimedia*, 7(4):786–789, 2005.

52. WILFRED A Gibson. A least-squares solution for case iv of the law of comparative judgment. *Psychometrika*, 18(1):15–21, 1953.
53. Bernd Girod. What's wrong with mean-squared error? In *Digital images and human vision*, pages 207–220. MIT press, 1993.
54. Lutz Goldmann, Francesca De Simone, Frederic Dufaux, Touradj Ebrahimi, Rudolf Tanner, and Mauro Lattuada. Impact of video transcoding artifacts on the subjective quality. In *Quality of Multimedia Experience (QoMEX), 2010 Second International Workshop on*, pages 52–57. IEEE, 2010.
55. James J Gross and Robert W Levenson. Emotion elicitation using films. *Cognition & Emotion*, 9(1):87–108, 1995.
56. Stephen R Gulliver and George Ghinea. Stars in their eyes: What eye-tracking reveals about multimedia perceptual quality. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 34(4):472–482, 2004.
57. Stephen R Gulliver and Gheorghita Ghinea. Defining user perception of distributed multimedia quality. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, 2(4):241–257, 2006.
58. Stephen R Gulliver and Gheorghita Ghinea. The perceptual and attentive impact of delay and jitter in multimedia delivery. *Broadcasting, IEEE Transactions on*, 53(2):449–458, 2007.
59. Jukka Häkkinen, Takashi Kawai, Jari Takatalo, Tuomas Leisti, Jenni Radun, Anni Hirsaho, and Göte Nyman. Measuring stereoscopic image quality experience with interpretation based quality methodology. In *Electronic Imaging 2008*, pages 68081B–68081B. International Society for Optics and Photonics, 2008.
60. Andrew M Haun and Eli Peli. Is image quality a function of contrast perception? In *IS&T/SPIE Electronic Imaging*, pages 86510C–86510C. International Society for Optics and Photonics, 2013.
61. Sheila S Hemami and Amy R Reibman. No-reference image and video quality estimation: Applications and human-motivated design. *Signal processing: Image communication*, 25(7):469–481, 2010.
62. T Hobfeld, Raimund Schatz, and Sebastian Egger. Sos: The mos is not enough! In *Quality of Multimedia Experience (QoMEX), 2011 Third International Workshop on*, pages 131–136. IEEE, 2011.
63. Donald Hoffman. The interface theory of perception: Natural selection drives true perception to swift extinction. *Object categorization: Computer and human vision perspectives*, pages 148–165, 2009.
64. Robyn M Holmes and Anthony D Pellegrini. Children's social behavior during video game play. *Handbook of Computer Game Studies. The MIT Press, Cambridge, Massachusetts*, 2005.
65. T Hoßfeld, C Keimel, M Hirth, B Gardlo, J Habigt, K Diepold, and P Tran-Gia. Crowdtesting: a novel methodology for subjective user studies and qoe evaluation. *University of Würzburg, Tech. Rep*, 486, 2013.
66. Jeff Howe. The rise of crowdsourcing. *Wired magazine*, 14(6):1–4, 2006.
67. Ming-Hui Huang. Designing website attributes to induce experiential encounters. *computers in Human Behavior*, 19(4):425–442, 2003.
68. Quan Huynh-Thu, M-N Garcia, Filippo Speranza, Philip Corriveau, and Alexander Raake. Study of rating scales for subjective quality assessment of high-definition video. *Broadcasting, IEEE Transactions on*, 57(1):1–14, 2011.
69. Mansoor Hyder, Noel Crespi, Michael Haun, Christian Hoene, et al. Are qoe requirements for multimedia services different for men and women? analysis of gender differences in forming qoe in virtual acoustic environments. In *Emerging Trends and Applications in Information Communication Technologies*, pages 200–209. Springer, 2012.
70. Selim Ickin, Katarzyna Wac, Markus Fiedler, Lucjan Janowski, Jin-Hyuk Hong, and Anind K Dey. Factors influencing quality of experience of commonly used mobile applications. *Communications Magazine, IEEE*, 50(4):48–56, 2012.

71. Wijnand A IJsselsteijn, Huib de Ridder, and Roelof Hamberg. Perceptual factors in stereoscopic displays: the effect of stereoscopic filming parameters on perceived quality and reported eyestrain. In *Photonics West'98 Electronic Imaging*, pages 282–291. International Society for Optics and Photonics, 1998.
72. Wijnand A IJsselsteijn, Huib de Ridder, and Joyce Vliegen. Subjective evaluation of stereoscopic images: effects of camera parameters and display duration. *Circuits and Systems for Video Technology, IEEE Transactions on*, 10(2):225–233, 2000.
73. P ITU-T RECOMMENDATION. Subjective video quality assessment methods for multimedia applications. 1999.
74. Ruud Janssen. *Computational image quality*, volume 101. SPIE press, 2001.
75. TJWM Janssen and FJJ Blommaert. Image quality semantics. *Journal of imaging science and Technology*, 41(5):555–560, 1997.
76. Wei Jiang, Alexander C Loui, and Cathleen Daniels Cerosaletti. Automatic aesthetic value assessment in photographic images. In *Multimedia and Expo (ICME), 2010 IEEE International Conference on*, pages 920–925. IEEE, 2010.
77. Dhiraj Joshi, Ritendra Datta, Elena Fedorovskaya, Quang-Tuan Luong, James Z Wang, Jia Li, and Jiebo Luo. Aesthetics and emotions in images. *Signal Processing Magazine, IEEE*, 28(5):94–115, 2011.
78. Satu Jumisko-Pyykkö. *User-centered quality of experience and its evaluation methods for mobile television*. PhD thesis, Doctoral thesis, Tampere University of Technology, Tampere, 2011.
79. Satu Jumisko-Pyykkö and Miska M Hannuksela. Does context matter in quality evaluation of mobile television? In *Proceedings of the 10th international conference on Human computer interaction with mobile devices and services*, pages 63–72. ACM, 2008.
80. Sandeep Kanumuri, Pamela C Cosman, Amy R Reibman, and Vinay A Vaishampayan. Modeling packet-loss visibility in mpeg-2 video. *Multimedia, IEEE Transactions on*, 8(2):341–355, 2006.
81. Brian Keelan. *Handbook of image quality: characterization and prediction*. CRC Press, 2002.
82. Brian W Keelan and Hitoshi Urabe. Iso 20462: A psychophysical image quality measurement standard. In *Electronic Imaging 2004*, pages 181–189. International Society for Optics and Photonics, 2003.
83. Kalevi Kilkki. Quality of experience in communications ecosystem. *J. UCS*, 14(5):615–624, 2008.
84. Hyun-Jong Kim, Dong Hyeon Lee, Jong Min Lee, Kyoung-Hee Lee, Won Lyu, and Seong-Gon Choi. The qoe evaluation method through the qos-qoe correlation model. In *Networked Computing and Advanced Information Management, 2008. NCM'08. Fourth International Conference on*, volume 2, pages 719–725. IEEE, 2008.
85. Hendrik Knoche and John McCarthy. Mobile users" needs and expectations of future multimedia services. Technical report, 2004.
86. Hendrik Knoche and M Angela Sasse. Getting the big picture on small screens: Quality of experience in mobile tv. Technical report, Information Science Reference, 2008.
87. Hendrik Knoche and M Angela Sasse. The big picture on small screens delivering acceptable video quality in mobile tv. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, 5(3):20, 2009.
88. Jan Koenderink. Vision as a user interface. In *IS&T/SPIE Electronic Imaging*, pages 786504–786504. International Society for Optics and Photonics, 2011.
89. Jari Korhonen, Claire Mantel, Nino Burini, and Soren Forchhammer. Searching for the preferred backlight intensity in liquid crystal displays with local backlight dimming. In *Quality of Multimedia Experience (QoMEX), 2013 Fifth International Workshop on*, pages 118–123. IEEE, 2013.
90. Philip Kortum and Marc Sullivan. The effect of content desirability on subjective video quality ratings. *Human factors: the journal of the human factors and ergonomics society*, 52(1):105–118, 2010.

91. Robert Kubey and Mihalyi Csikszentmihalyi. *Television and the quality of life: How viewing shapes everyday experience*. Routledge, 2013.
92. Andre Kuijsters, Wijnand A Ijsselsteijn, Marc TM Lambooij, and Ingrid EJ Heynderickx. Influence of chroma variations on naturalness and image quality of stereoscopic images. In *IS&T/SPIE Electronic Imaging*, pages 72401E–72401E. International Society for Optics and Photonics, 2009.
93. Wen-Hung Kuo, Po-Hung Lin, and Sheue-Ling Hwang. A framework of perceptual quality assessment on lcd-tv. *Displays*, 28(1):35–43, 2007.
94. Marc Lambooij, Marten Fortuin, Ingrid Heynderickx, and Wijnand IJsselsteijn. Visual discomfort and visual fatigue of stereoscopic displays: a review. *Journal of Imaging Science and Technology*, 53(3):30201–1, 2009.
95. Marc Lambooij, Wijnand IJsselsteijn, Don G Bouwhuis, and Ingrid Heynderickx. Evaluation of stereoscopic images: beyond 2d quality. *Broadcasting, IEEE Transactions on*, 57(2): 432–444, 2011.
96. Peter J Lang. Behavioral treatment and bio-behavioral assessment: Computer applications. 1980.
97. Peter J Lang, Margaret M Bradley, Bruce N Cuthbert, et al. *International affective picture system (IAPS): Affective ratings of pictures and instruction manual*. NIMH, Center for the Study of Emotion & Attention, 2005.
98. Julie Lassalle, Laetitia Gros, Thierry Morineau, and Gilles Coppin. Impact of the content on subjective evaluation of audiovisual quality: What dimensions influence our perception? In *Broadband Multimedia Systems and Broadcasting (BMSB), 2012 IEEE International Symposium on*, pages 1–6. IEEE, 2012.
99. Patrick Le Callet, Sebastian Möller, Andrew Perkis, et al. Qualinet white paper on definitions of quality of experience. *European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003)*, 2012.
100. Olivier Le Meur, Alexandre Ninassi, Patrick Le Callet, and Dominique Barba. Do video coding impairments disturb the visual attention deployment? *Signal Processing: Image Communication*, 25(8):597–609, 2010.
101. Barbara Lee and Robert S Lee. How and why people watch tv: Implications for the future of interactive television. *Journal of advertising research*, 35(6):9–18, 1995.
102. Jong-Seok Lee, Francesca De Simone, and Touradj Ebrahimi. Subjective quality evaluation via paired comparison: application to scalable video coding. *Multimedia, IEEE Transactions on*, 13(5):882–893, 2011.
103. Jong-Seok Lee, Francesca De Simone, Naeem Ramzan, Zhijie Zhao, Engin Kurutepe, Thomas Sikora, Jörn Ostermann, Ebroul Izquierdo, and Touradj Ebrahimi. Subjective evaluation of scalable video coding for content distribution. In *Proceedings of the international conference on Multimedia*, pages 65–72. ACM, 2010.
104. Congcong Li and Tsuhan Chen. Aesthetic visual quality assessment of paintings. *Selected Topics in Signal Processing, IEEE Journal of*, 3(2):236–252, 2009.
105. Weisi Lin and C-C Jay Kuo. Perceptual visual quality metrics: A survey. *Journal of Visual Communication and Image Representation*, 22(4):297–312, 2011.
106. Hantao Liu and Ingrid Heynderickx. A perceptually relevant no-reference blockiness metric based on local image characteristics. *EURASIP Journal on Advances in Signal Processing*, 2009:2, 2009.
107. Hantao Liu and Ingrid Heynderickx. Visual attention in objective image quality assessment: based on eye-tracking data. *Circuits and Systems for Video Technology, IEEE Transactions on*, 21(7):971–982, 2011.
108. James Lull et al. *Inside family viewing: ethnographic research on television's audiences.* Routledge, 1990.
109. Yiwen Luo and Xiaoou Tang. Photo and video quality evaluation: Focusing on the subject. In *Computer Vision–ECCV 2008*, pages 386–399. Springer, 2008.

110. Wendy Ann Mansilla, Andrew Perkis, and Touradj Ebrahimi. Implicit experiences as a determinant of perceptual quality and aesthetic appreciation. In *Proceedings of the 19th ACM international conference on Multimedia*, pages 153–162. ACM, 2011.
111. Claire Mantel, Nathalie Guyader, Patricia Ladret, Gelu Ionescu, and Thomas Kunlin. Characterizing eye movements during temporal and global quality assessment of h. 264 compressed video sequences. In *IS&T/SPIE Electronic Imaging*, pages 82910Y–82910Y. International Society for Optics and Photonics, 2012.
112. John D McCarthy, M Angela Sasse, and Dimitrios Miras. Sharp or smooth?: comparing the effects of quantization vs. frame rate for streamed video. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 535–542. ACM, 2004.
113. Ricky KP Mok, Edmond WW Chan, and Rocky KC Chang. Measuring the quality of experience of http video streaming. In *Integrated Network Management (IM), 2011 IFIP/IEEE International Symposium on*, pages 485–492. IEEE, 2011.
114. Anush Krishna Moorthy and Alan Conrad Bovik. Visual quality assessment algorithms: what does the future hold? *Multimedia Tools and Applications*, 51(2):675–696, 2011.
115. Margaret Morrison and Dean M Krugman. A look at mass and computer mediated technologies: Understanding the roles of television and computers in the home. *Journal of Broadcasting & Electronic Media*, 45(1):135–161, 2001.
116. JH Morrissey. New method for the assignment of psychometric scale values from incomplete paired comparisons. *JOSA*, 45(5):373–378, 1955.
117. Niall Murray, Yuansong Qiao, Brian Lee, Gabriel-Miro Muntean, and AK Karunakar. Age and gender influence on perceived olfactory & visual media synchronization. In *Multimedia and Expo (ICME), 2013 IEEE International Conference on*, pages 1–6. IEEE, 2013.
118. Anja B Naumann, Ina Wechsung, and Jörn Hurtienne. Multimodal interaction: A suitable strategy for including older users? *Interacting with Computers*, 22(6):465–474, 2010.
119. A Neil. Autostereoscopic 3d displays. *Computer*, 8:32–36, 2005.
120. MRM Nijenhuis. *Sampling, interpolation, images: a perceptual view*. PhD thesis, Eindhoven University of technology, Eindhoven, 1993.
121. MRM Nijenhuis and FJJ Blommaert. Perceptual error measure for sampled and interpolated images. *Journal of Imaging Science and Technology*, 41(3):249–258, 1997.
122. Alexandre Ninassi, Olivier Le Meur, Patrick Le Callet, and Dominique Barba. Considering temporal variations of spatial visual distortions in video quality assessment. *Selected Topics in Signal Processing, IEEE Journal of*, 3(2):253–265, 2009.
123. Alexandre Ninassi, Olivier Le Meur, Patrick Le Callet, Dominique Barba, Arnaud Tirel, et al. Task impact on the visual attention in subjective image quality assessment. In *Proceedings of European Signal Processing Conference*, 2006.
124. Heather L O'Brien and Elaine G Toms. What is user engagement? a conceptual framework for defining user engagement with technology. *Journal of the American Society for Information Science and Technology*, 59(6):938–955, 2008.
125. Lora Oehlberg, Nicolas Ducheneaut, James D Thornton, Robert J Moore, and Eric Nickell. Social tv: Designing for distributed, sociable television viewing. In *Proc. EuroITV*, volume 2006, pages 25–26, 2006.
126. Joana Palhais, Rui S Cruz, and Mário S Nunes. Quality of experience assessment in internet tv. In *Mobile Networks and Management*, pages 261–274. Springer, 2012.
127. Thrasyvoulos N Pappas, Robert J Safranek, and Junqing Chen. Perceptual criteria for image quality evaluation. *Handbook of image and video processing*, pages 669–684, 2000.
128. Fernando Pereira. Sensations, perceptions and emotions towards quality of experience evaluation for consumer electronics video adaptations. In *Int. Workshop on Video Processing and Quality Metrics for Consumer Electronics*, 2005.
129. Pablo Pérez, Jesús Macías, Jaime J Ruiz, and Narciso García. Effect of packet loss in video quality of experience. *Bell Labs Technical Journal*, 16(1):91–104, 2011.
130. Lawrence A Pervin and Oliver P John. *Handbook of personality: Theory and research*. Elsevier, 1999.

131. Kandaraj Piamrat, Cesar Viho, J Bonnin, and Adlen Ksentini. Quality of experience measurements for video streaming over wireless networks. In *Information Technology: New Generations, 2009. ITNG'09. Sixth International Conference on*, pages 1184–1189. IEEE, 2009.
132. Scott Plous. *The psychology of judgment and decision making.* Mcgraw-Hill Book Company, 1993.
133. Jordi Puig, Andrew Perkis, Frank Lindseth, and Touradj Ebrahimi. Towards an efficient methodology for evaluation of quality of experience in augmented reality. In *Quality of Multimedia Experience (QoMEX), 2012 Fourth International Workshop on*, pages 188–193. IEEE, 2012.
134. Jenni Radun, Tuomas Leisti, Jukka Häkkinen, Harri Ojanen, Jean-Luc Olives, Tero Vuori, and Göte Nyman. Content and quality: Interpretation-based estimation of image quality. *ACM Transactions on Applied Perception (TAP)*, 4(4):2, 2008.
135. Benjamin Rainer, Markus Waltl, Eva Cheng, Muawiyath Shujau, Christian Timmerer, Stephen Davis, Ian Burnett, Christian Ritz, and Hermann Hellwagner. Investigating the impact of sensory effects on the quality of experience and emotional response in web videos. In *Quality of Multimedia Experience (QoMEX), 2012 Fourth International Workshop on*, pages 278–283. IEEE, 2012.
136. JC Read, SJ MacFarlane, and Chris Casey. Endurability, engagement and expectations: Measuring children's fun. In *Interaction design and children*, volume 2, pages 1–23. Shaker Publishing Eindhoven, 2002.
137. ITU-R BT Recommendation. 2021, subjective methods for the assessment stereoscopic 3dtv systems. *International Telecommunication Union, Geneva, Switzerland*, 2012.
138. ITURBT Recommendation. 500-11, methodology for the subjective assessment of the quality of television pictures. *International Telecommunication Union, Geneva, Switzerland*, 4:2, 2002.
139. ITUT Recommendation. E. 800: Terms and definitions related to quality of service and network performance including dependability. *ITU-T 2008*, 2008.
140. Judith Redi, Ingrid Heynderickx, Bruno Macchiavello, and Mylene Farias. On the impact of packet-loss impairments on visual attention mechanisms. In *Circuits and Systems (ISCAS), 2013 IEEE International Symposium on*, pages 1107–1110. IEEE, 2013.
141. Judith Redi, Hantao Liu, Hani Alers, Rodolfo Zunino, and Ingrid Heynderickx. Comparing subjective image quality measurement methods for the creation of public databases. In *IS&T/SPIE Electronic Imaging*, pages 752903–752903. International Society for Optics and Photonics, 2010.
142. Judith Redi, Hantao Liu, Paolo Gastaldo, Rodolfo Zunino, and Ingrid Heynderickx. How to apply spatial saliency into objective metrics for jpeg compressed images? In *Image Processing (ICIP), 2009 16th IEEE International Conference on*, pages 961–964. IEEE, 2009.
143. Judith Redi, Hantao Liu, Rodolfo Zunino, and Ingrid Heynderickx. Interactions of visual attention and quality perception. In *IS&T/SPIE Electronic Imaging*, pages 78650S–78650S. International Society for Optics and Photonics, 2011.
144. Judith A Redi. Visual quality beyond artifact visibility. In *IS&T/SPIE Electronic Imaging*, pages 86510N–86510N. International Society for Optics and Photonics, 2013.
145. Judith A Redi and Ingrid Heynderickx. Image integrity and aesthetics: towards a more encompassing definition of visual quality. In *IS&T/SPIE Electronic Imaging*, pages 829115–829115. International Society for Optics and Photonics, 2012.
146. Judith A Redi and Isabel Povoa. The role of visual attention in the aesthetic appeal of consumer images: A preliminary study. In *Visual Communications and Image Processing (VCIP), 2013*, pages 1–6. IEEE, 2013.
147. Judith Alice Redi, Tobias Hoßfeld, Pavel Korshunov, Filippo Mazza, Isabel Povoa, and Christian Keimel. Crowdsourcing-based multimedia subjective evaluations: a case study on image recognizability and aesthetic appeal. In *Proceedings of the 2nd ACM international workshop on Crowdsourcing for multimedia*, pages 29–34. ACM, 2013.

148. Ulrich Reiter and Katrien De Moor. Content categorization based on implicit and explicit user feedback: combining self-reports with eeg emotional state analysis. In *Quality of multimedia experience (QoMEX), 2012 fourth international workshop on*, pages 266–271. IEEE, 2012.
149. Huib Ridder and Serguei Endrikhovski. 33.1: Invited paper: image quality is fun: reflections on fidelity, usefulness and naturalness. In *SID Symposium Digest of Technical Papers*, volume 33, pages 986–989. Wiley Online Library, 2002.
150. Raimund Schatz, Siegfried Wagner, Sebastian Egger, and Norbert Jordan. Mobile tv becomes social-integrating content with communications. In *Information Technology Interfaces, 2007. ITI 2007. 29th International Conference on*, pages 263–270. IEEE, 2007.
151. Jose A Scheinkman. Social interactions. *The New Palgrave Dictionary of Economics*, 2, 2008.
152. Dimitri Schuurman, Katrien De Moor, Lieven De Marez, and Tom Evens. A living lab research approach for mobile tv. *Telematics and Informatics*, 28(4):271–282, 2011.
153. Eric WK See-To, Savvas Papagiannidis, and Vincent Cho. User experience on mobile video appreciation: How to engross users and to enhance their enjoyment in watching mobile video clips. *Technological Forecasting and Social Change*, 79(8):1484–1494, 2012.
154. Helge Seetzen, Wolfgang Heidrich, Wolfgang Stuerzlinger, Greg Ward, Lorne Whitehead, Matthew Trentacoste, Abhijeet Ghosh, and Andrejs Vorozcovs. High dynamic range display systems. In *ACM Transactions on Graphics (TOG)*, volume 23, pages 760–768. ACM, 2004.
155. Pieter Seuntiens, Lydia Meesters, and Wijnand Ijsselsteijn. Perceived quality of compressed stereoscopic images: Effects of symmetric and asymmetric jpeg coding and camera separation. *ACM Transactions on Applied Perception (TAP)*, 3(2):95–109, 2006.
156. Pieter Seuntiens, Ingrid Vogels, and Arnold van Keersop. Visual experience of 3d-tv with pixelated ambilight. *Proceedings of PRESENCE*, 2007, 2007.
157. Pieter J Seuntiëns, Ingrid E Heynderickx, Wijnand A IJsselsteijn, Paul MJ van den Avoort, Jelle Berentsen, Iwan J Dalm, Marc T Lambooij, and Willem Oosting. Viewing experience and naturalness of 3d images. In *Optics East 2005*, pages 601605–601605. International Society for Optics and Photonics, 2005.
158. Mario Siller and John Woods. Improving quality of experience for multimedia services by qos arbitration on a qoe framework. In *in Proc. of the 13th Packed Video Workshop 2003*. Citeseer, 2003.
159. Paul J Silvia. Interest – the curious emotion. *Current Directions in Psychological Science*, 17(1):57–60, 2008.
160. Joshua A Solomon, Andrew B Watson, and Albert Ahumada. Visibility of dct basis functions: Effects of contrast masking. In *Data Compression Conference, 1994. DCC'94. Proceedings*, pages 361–370. IEEE, 1994.
161. Nicolas Staelens, Stefaan Moens, Wendy Van den Broeck, Ilse Marien, Brecht Vermeulen, Peter Lambert, Rik Van de Walle, and Piet Demeester. Assessing quality of experience of iptv and video on demand services in real-life environments. *Broadcasting, IEEE Transactions on*, 56(4):458–466, 2010.
162. Nicolas Staelens, Glenn Van Wallendael, Karel Crombecq, Nick Vercammen, Jan De Cock, Brecht Vermeulen, Rik Van de Walle, Tom Dhaene, and Piet Demeester. No-reference bitstream-based visual quality impairment detection for high definition h. 264/avc encoded video sequences. *Broadcasting, IEEE Transactions on*, 58(2):187–199, 2012.
163. Wa James Tam, Lew B Stelmach, and Philip J Corriveau. Psychovisual aspects of viewing stereoscopic video sequences. In *Photonics West'98 Electronic Imaging*, pages 226–235. International Society for Optics and Photonics, 1998.
164. KT Tan, Mohammed Ghanbari, and Donald E Pearson. An objective measurement tool for mpeg video quality. *Signal Processing*, 70(3):279–294, 1998.
165. Cornelis Teunissen. Flat panel display characterization: a perceptual approach. 2009.
166. Louis L Thurstone. A law of comparative judgment. *Psychological review*, 34(4):273, 1927.
167. International Telecommunication Union. Itu-t rec. 109: Definition of quality of experience (qoe). *Liaison Statement, Ref.: TD 109rev2 (PLEN/12)*, 2007.

168. Vanessa Vakili, Willem-Paul Brinkman, and Mark A Neerincx. Lessons learned from the development of technological support for ptsd prevention: a review. *Stud Health Technol Inform*, 181:22–6, 2012.
169. Patricia Valdez and Albert Mehrabian. Effects of color on emotions. *Journal of Experimental Psychology: General*, 123(4):394, 1994.
170. Martin Varela, Toni Maki, Lea Skorin-Kapov, and Tobias Hossfeld. Towards an understanding of visual appeal in website design. In *Quality of Multimedia Experience (QoMEX), 2013 Fifth International Workshop on*, pages 70–75. IEEE, 2013.
171. Elena Vicario, Ingrid Heynderickx, Giulio Ferretti, and Paola Carrai. 17.1: Design of a tool to benchmark scaling algorithms on lcd monitors. In *SID Symposium Digest of Technical Papers*, volume 33, pages 704–707. Wiley Online Library, 2002.
172. Ingrid MLC Vogels. How to make life more colorful: from image quality to atmosphere experience. In *Color and Imaging Conference*, volume 2009, pages 123–128. Society for Imaging Science and Technology, 2009.
173. ECL Vu and DM Chandler. Visual fixation patterns when judging image quality: Effects of distortion type, amount, and subject experience. In *Image Analysis and Interpretation, 2008. SSIAI 2008. IEEE Southwest Symposium on*, pages 73–76. IEEE, 2008.
174. Tero Vuori, Maria Olkkonen, Monika Pölönen, Ari Siren, and Jukka Häkkinen. Can eye movements be quantitatively applied to image quality studies? In *Proceedings of the third Nordic conference on Human-computer interaction*, pages 335–338. ACM, 2004.
175. Demin Wang, Filippo Speranza, Andre Vincent, Taali Martin, and Phil Blanchfield. Toward optimal rate control: a study of the impact of spatial resolution, frame rate, and quantization on subjective video quality and bit rate. In *Visual Communications and Image Processing 2003*, pages 198–209. International Society for Optics and Photonics, 2003.
176. Junle Wang, Damon M Chandler, and Patrick Le Callet. Quantifying the relationship between visual salience and visual importance. In *IS&T/SPIE Electronic Imaging*, pages 75270K–75270K. International Society for Optics and Photonics, 2010.
177. Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *Image Processing, IEEE Transactions on*, 13(4):600–612, 2004.
178. Zhou Wang and Xinli Shang. Spatial pooling strategies for perceptual image quality assessment. In *Image Processing, 2006 IEEE International Conference on*, pages 2945–2948. IEEE, 2006.
179. Andrew B Watson. Efficiency of a model human image code. *JOSA A*, 4(12):2401–2417, 1987.
180. Ina Wechsung, Matthias Schulz, Klaus-Peter Engelbrecht, Julia Niemann, and Sebastian Möller. All users are (not) equal-the influence of user characteristics on perceived quality, modality choice and performance. In *Proceedings of the Paralinguistic Information and its Integration in Spoken Dialogue Systems Workshop*, pages 175–186. Springer, 2011.
181. Amaya Becvar Weddle and Hua Yu. How does audio-haptic enhancement influence emotional response to mobile media? In *Quality of Multimedia Experience (QoMEX), 2013 Fifth International Workshop on*, pages 158–163. IEEE, 2013.
182. J.H.D.M. Westerink. Influences of subject expertise in quality assessment of digitally coded images. In *SID International Symposium Digest of Technical Papers*, volume 20, pages 124–127. SID, 1989.
183. Joyce HDM Westerink and Jacques AJ Roufs. Subjective image quality as a function of viewing distance, resolution, and picture size. *SMPTE journal*, 98(2):113–119, 1989.
184. Stefan Winkler and Subramanian Ramanathan. Overview of eye tracking datasets. In *QoMEX*, pages 212–217, 2013.
185. K Maria Wolters, Klaus-Peter Engelbrecht, Florian Gödde, Sebastian Möller, Anja Naumann, and Robert Schleicher. Making it easier for older people to talk to smart homes: the effect of early help prompts. *Universal Access in the Information Society*, 9(4):311–325, 2010.

186. Chen-Chi Wu, Kuan-Ta Chen, Yu-Chun Chang, and Chin-Laung Lei. Crowdsourcing multimedia qoe evaluation: A trusted framework. *IEEE transactions on multimedia*, 15(5):1121–1137, 2013.
187. Qianqian Xu, Qingming Huang, Tingting Jiang, Bowei Yan, Weisi Lin, and Yuan Yao. Hodgerank on random graphs for subjective video quality assessment. *Multimedia, IEEE Transactions on*, 14(3):844–857, 2012.
188. Qianqian Xu, Qingming Huang, and Yuan Yao. Online crowdsourcing subjective image quality assessment. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 359–368. ACM, 2012.
189. Qianqian Xu, Jiechao Xiong, Qingming Huang, and Yuan Yao. Robust evaluation for quality of experience in crowdsourcing. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 43–52. ACM, 2013.
190. Kyoko Yamori and Yoshiaki Tanaka. Relation between willingness to pay and guaranteed minimum bandwidth in multiple-priority service. In *Communications, 2004 and the 5th International Symposium on Multi-Dimensional Mobile Communications Proceedings. The 2004 Joint Conference of the 10th Asia-Pacific Conference on*, volume 1, pages 113–117. IEEE, 2004.
191. Akiko Yoshida, Volker Blanz, Karol Myszkowski, and Hans-Peter Seidel. Perceptual evaluation of tone mapping operators with real-world scenes. In *Electronic Imaging 2005*, pages 192–203. International Society for Optics and Photonics, 2005.
192. Ai Yoto, Tetsuo Katsuura, Koichi Iwanaga, and Yoshihiro Shimomura. Effects of object color stimuli on human brain activities in perception and attention referred to eeg alpha band response. *Journal of Physiological Anthropology*, 26(3):373–379, 2007.
193. H van Zee and DW Kaandorp. Kwaliteit en degradatie. ipo rapport no. 853. Technical report, Institute for Perception Research, Eindhoven, 1992.
194. Thomas Zinner, Oliver Hohlfeld, Osama Abboud, and Tobias Hoßfeld. Impact of frame rate and resolution on objective qoe metrics. In *Quality of Multimedia Experience (QoMEX), 2010 Second International Workshop on*, pages 29–34. IEEE, 2010.

Chapter 3
Recent Advances in Image Quality Assessment

Guangtao Zhai

3.1 Subjective Quality Assessment

Approximately 900 billion digital images will be taken in 2014 and this number is expected to keep increasing every year in the future. A natural problem follows is that the visual quality of such a great amount of photographs is hard to guarantee, possibly due to the limitations in camera device, lighting condition, and shooting skills. Therefore the systems to monitor, control, and improve the visual quality of digital photographs are highly desirable [1]. Image quality assessment (IQA), due to its capability of rating image quality in a way that approximates human judgement, is an ideal solution to this problem.

At the most general level, IQA can be classified into subjective assessment and objective assessment [2]. Subjective IQA is widely accepted as the most accurate quality gauge since human eyes are the final receiver of most, if not all, visual communication systems. Moreover, subjective IQA is also of great importance for direct optimization of coding and other algorithms used in visual communication systems [3, 4].

The most significant contribution of subjective IQA over the last decades is probably the construction of the IQA databases, which are consisted of digital images with various kinds of distortions and their subjective ratings. Those databases have greatly facilitated the research of objective IQA metrics in recent years. Examples include Laboratory for Image & Video Engineering (LIVE) database [5], Tampere Image Database 2008 (TID2008) [6], and Categorical Subjective Image Quality (CSIQ) database [7], as well as four recent ones, including Tampere Image Database

G. Zhai (✉)
Shanghai Jiao Tong University, Shanghai, China
e-mail: zhaiguangtao@sjtu.edu.cn

C. Deng et al. (eds.), *Visual Signal Quality Assessment*,
DOI 10.1007/978-3-319-10368-6_3

2013 (TID2013) [8], LIVE multiply distorted image database (LIVEMD) [9], contrast-changed image quality database (CID2013) [10], and high dynamic range image quality database (HDR2014) [11].

3.1.1 Popular IQA Databases

The LIVE database [5] was developed at University of Texas at Austin. It is consisted of five sub-sets of 982 subject-rated images, which include 779 distorted images created from 29 pristine ones with five types of distortions at different distortion levels. The distortion types are: (a) JPEG2000 compression; (b) JPEG compression; (c) White noise contamination; (d) Gaussian blur; and (e) fast fading channel distortion of JPEG2000 compressed bitstream. The subjective test was carried out with each data set individually. A cross-comparison set that mixes images from all distortion types is then used to help align the subject scores across data sets [13]. The subjective scores of the overall images are then adjusted accordingly. Those realigned differential mean opinion score (DMOS) values, ranging from -3 to 112, are used because they are more precise than the original scores.

The TID2008 database [6] was developed as a joint international effort between Finland, Italy, and Ukraine. It includes 1,700 distorted images generated from 25 reference images with 17 distortion categories and 4 distortion levels. The types of distortions include: (a) Additive Gaussian noise; (b) Additive noise in color channels; (c) Spatially correlated noise; (d) Masked noise; (e) High frequency noise; (f) Impulse noise; (g) Quantization noise; (h) Gaussian blur; (i) Image denoising; (j) JPEG compression; (k) JPEG2000 compression; (l) JPEG transmission errors; (m) JPEG2000 transmission errors; (n) Noneccentricity pattern noise; (o) Local block-wise distortions of different intensity; (p) Mean shift (intensity shift); (q) Contrast change. The mean opinion score (MOS) values of those images are from 0.2 to 7.3.

The categorical image quality (CSIQ) database [7] was developed at Oklahoma State University and consists of 866 images which are derived from 30 original versions. Six distortion types (with four to five levels) were used in CSIQ, namely JPEG compression, JPEG2000 compression, additive Gaussian white noise, additive Gaussian pink noise, Gaussian blurring, and global contrast decrements. The DMOS of each image ranges from 0 to 1.

3.1.2 New Quality Database

Recently, the TID2013 database [8] was released as an extension of the TID2008 database. It contains total number of 3,000 images, which were generated by corrupting 25 original images with 24 types of distortion at 5 different levels. The distortions include the abovementioned 17 types [(a)–(q) in TID 2008] and (r) Change of color saturation; (s) Multiplicative Gaussian noise; (t) Comfort noise;

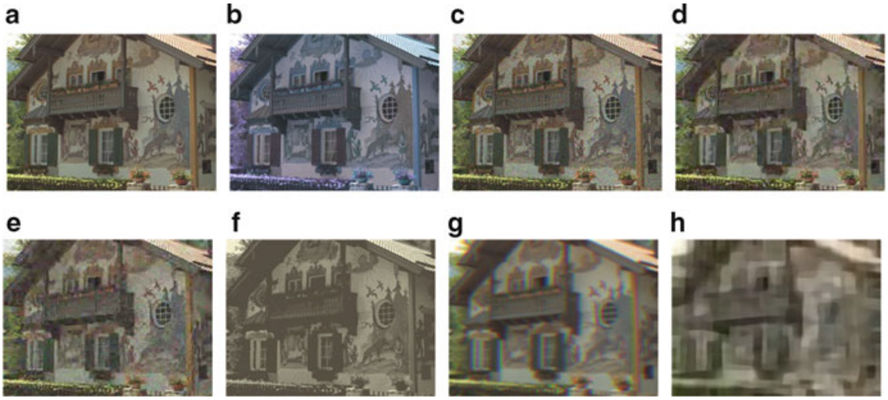


Fig. 3.1 The sample image "mountain chalet" in the TID2013 database: (**a**) original; (**b**) change of color saturation; (**c**) Multiplicative Gaussian noise; (**d**) comfort noise; (**e**) lossy compression of noisy images; (**f**) image color quantization with dither; (**g**) chromatic aberrations; (**h**) sparse sampling and reconstruction

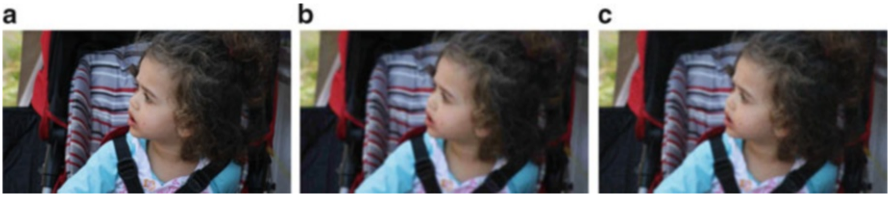


Fig. 3.2 The sample image "babygirl" in the LIVEMD database: (**a**) original; (**b**) blur & JPEG; (**c**) blur & noise

(u) Lossy compression of noisy images; (v) Image color quantization with dither; (w) Chromatic aberrations; (x) Sparse sampling and reconstruction. In Fig. 3.1, we show an example image and its distorted versions with the latterly introduced seven distortion types. The MOS values of the whole images were acquired via paired comparison assessment and range from 0.2 to 7.3.

Those above introduced IQA databases covered many types of distortions. However, in real-world image processing systems, different distortions tend to occur together, e.g. noisy and blurry images. This multiple distortion problem causes difficulty even for many IQA metrics that are very successful for single type of distortion. To investigate this problem, a subjective study in [9] was conducted to obtain human judgements on images corrupted by two distortion scenarios: (1) image storage: where images are first blurred and then compressed by a JPEG encoder; (2) camera image acquisition: where images are first blurred due to defocus and then corrupted by white Gaussian noise. In each of the two scenarios, a group of 225 images (135 multiply distorted images and 90 singly distorted images) were generated from 15 original ones. Examples of multiply corrupted images are given in Fig. 3.2.

If we study the average performance of existing IQA metrics for each type of those mentioned distortions, it can be found that one specific type of distortion, namely contrast change, poses the most prominent challenge. TID2008 was the first database to include contrast related image subsets (contrast change and mean luminance shift) and the original images were taken from the Kodak database [12]. The CSIQ database also contains contrast-changed images created from 30 sources spanning a wide range of contents and scenes. It is easy to imagine that for a given original image, proper contrast enhancement can lead to improved perceptual quality, which however conflicts with the increased distance between the original and enhanced images. In other words, if it is assumed that the original image has the best quality, then no contrast change can further enhance the quality. On the other hand, most of those original images are not always of perfect contrast. Therefore, IQA of contrast changed images proves to be difficult for most state-of-the-art quality metrics. The contrast related image subsets in TID2008, CSIQ, and TID2013 are relatively small and this further restricts the study of the topic. Recently, a dedicated and more comprehensive database for contrast changed images CID2013 [10] were introduced. CID2013 consists of 15 natural images taken from Kodak database and 400 contrast-changed versions. Two types of distortions were used, namely mean luminance shift and contrast change. In mean luminance shift, the original image I_o is added with a positive or negative value ($+\triangle I$ or $-\triangle I$). The offset $\triangle I$ has six levels of $\{20, 40, 60, 80, 100, 120\}$. In contrast change, images undergo luminance mapping, using either concave arc, convex arc, cubic or logistic function. The transfer curves and some example images are shown in Figs. 3.3 and 3.4.

High dynamic range (HDR) imaging has attracted a lot of attention and enthusiasm in the last decade. With quick advances of sensor technologies, even consumer level digital cameras are capable of capturing HDR images. However, a vast majority of nowadays displays still only support 8-bit color depth, and this leads to the widely studied problem of showing HDR images on low dynamic range (LDR) devices or tone mapping. On the other hand, the emergence of 10- or more bit display devices brings the possibility of direct visualization of those HDR images. So a natural question to ask is whether existing popular image quality metrics designed for and validated on LDR (8-bit) images perform equally well for HDR (10-bit) images. In [11] a new and dedicated HDR image quality database (HDR2014)

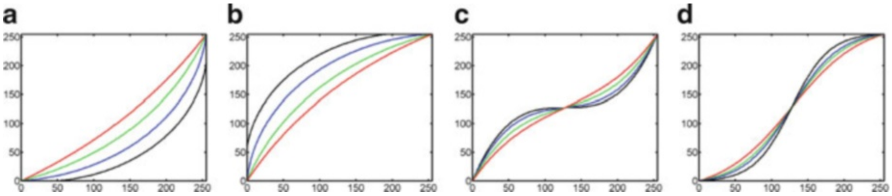


Fig. 3.3 Four kinds of transfer mappings in CID2013: (**a**) concave arcs; (**b**) convex arcs; (**c**) cubic functions; (**d**) logistic functions

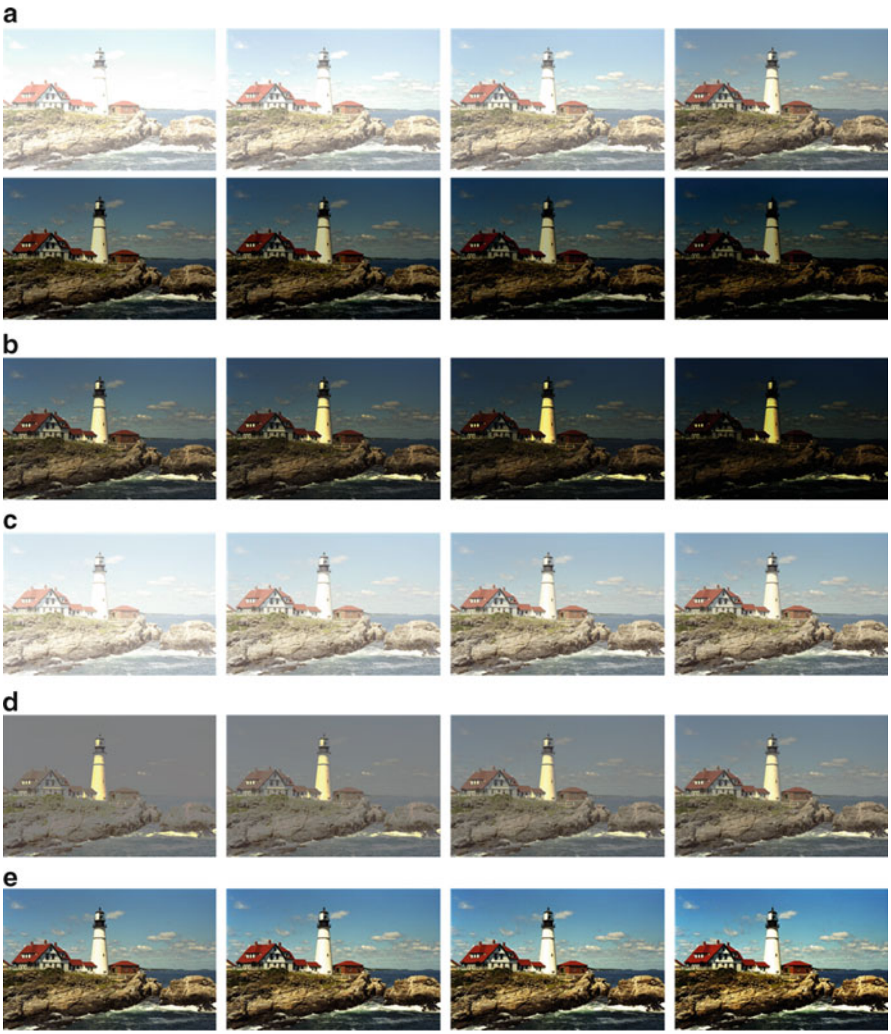


Fig. 3.4 Representative images in the CID2013 database. (**a**) Mean-shifted images. (**b**) Concave arcs transferred images. (**c**) Convex arcs transferred images in CID2013. (**d**) Cubic functions transferred images. (**e**) Logistic functions transferred images

was proposed. That HDR2014 database consists of 192 images with four kinds of distortions applied on six reference images. More specifically, eight distortion levels for the artifacts of JPEG/JPEG2000 compression, white noise injection, and Gaussian blurring were used. Twenty-five inexperienced viewers were involved in the subjective viewing test. Images were displayed on a pair of carefully calibrated 8-bit LDR and 10-bit HDR monitors and the subjective scores on both of which were

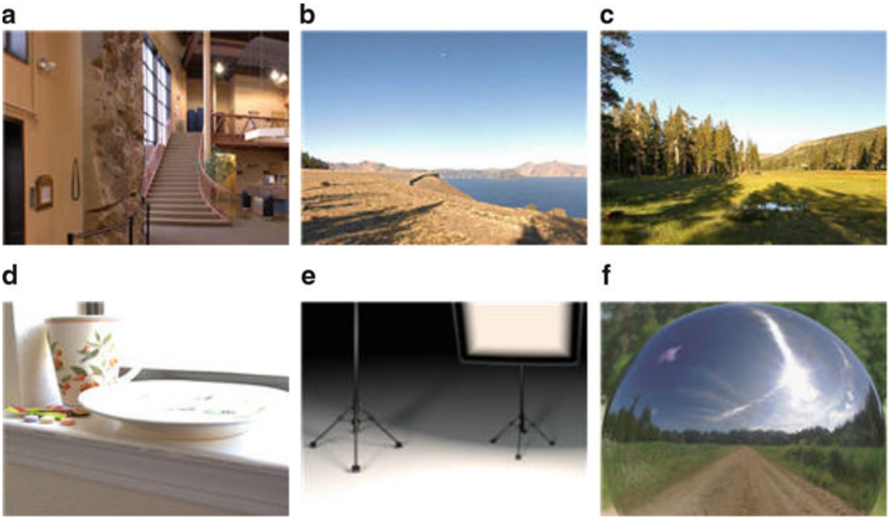


Fig. 3.5 The six original HDR images from websites

recorded. Then some ubiquitous and state-of-the-art IQA metrics were tested on that database. Experimental results show that HDR monitor indeed improved perceptual quality of the visual stimuli, as compared to LDR ones. And several existing IQA metrics are still doing well on HDR images, yet performance of some metrics drops significantly. In Fig. 3.5, six original HDR images are exhibited.

3.2 Objective Quality Assessment

Subjective IQA, despite being the most accurate, is often slow, expensive, and laborious, and thus not suitable for the quantification and optimization of real-world visual communication systems. Therefore, many researchers have devoted to the exploration of objective IQA algorithms. Objective metrics can be further divided into three types depending on the availability of the reference image to be compared with during the tests, namely (1) full-reference, FR-IQA; (2) reduced-reference, RR-IQA; (3) no-reference, NR-IQA. Some representative FR/RR/NR-IQA algorithms will be briefly reviewed in this section.

3.2.1 Classic Quality Metrics

The mean-squared error (MSE) and peak signal-to-noise ratio (PSNR) are the most widely known quality measures, and have been used ubiquitously as quality criterion in visual communication system up to date. The MSE essentially measures the energy difference between a distorted visual signal $\mathbf{d}$ and the ideal one $\mathbf{r}$, and PSNR is a close relative of MSE,

$$\text{MSE} = \frac{1}{N}||\mathbf{r} - \mathbf{d}||^2 \qquad \text{PSNR} = 10\log_{10}(\frac{L^2}{\text{MSE}}) \tag{3.1}$$

where N is the pixel number in the image and L is the maximum dynamic range. For standard 8-bit images/videos, L equals to $255(= 2^8 - 1)$. It is easy to find that MSE/PSNR is of very clear physical meanings and easy to compute, but they completely ignore the substantially influence of the pixel location and image contents on quality evaluation, which makes them poorly correlate with the human judgement of image quality, or the MOS [2].

To address the weakness of MSE/PSNR, Wang et al. considered location information during quality assessment, and introduced the universal quality index (UQI) [14]. Later Wang et al. improved UQI and proposed a simple yet effective IQA metric called structural similarity (SSIM) index [15]. SSIM was based on a reasonable hypothesis that human visual perception is highly adapted for extracting structural information from a scene. The SSIM metric compares the luminance, contrast, and structural similarities as follows:

$$\begin{aligned} \text{SSIM} &= \frac{1}{M}\sum_{i=1}^{M}\text{SSIM\_MAP}(r_i, d_i) \\ &= \frac{1}{M}\sum_{i=1}^{M} l(r_i, d_i)\cdot c(r_i, d_i)\cdot s(r_i, d_i) \end{aligned} \tag{3.2}$$

where M is the number of local windows in the image and l, c, and s stands for luminance, contrast, and structural similarity, respectively. SSIM was shown to outperform MSE/PSNR with a seizable margin on the LIVE database [5].

3.2.2 FR-IQA

Despite the successfulness of SSIM on the LIVE database, its performance is far from ideal, especially on those new IQA databases such as TID2008, CSIQ, and TID2013. A large quantity of FR-IQA approaches, including many improved SSIM-type of methods [16–39], have been proposed in the last decade and have achieved remarkable improvement in terms of prediction accuracy.

3.2.2.1 Scale Transform-Based FR-IQA

It is easy to imagine that the perceived quality of an image heavily depends upon the viewing distance. So IQA will also benefit from multi-scale analysis, which is an effective tool for various image processing tasks. Multi-scale SSIM (MS-SSIM) [16] proposed to perform SSIM on different levels and fuse the results with psychophysically determined weights. In [17], it was found that using different scale transform coefficients for each component (luminance, contrast, and structure) in SSIM achieves further performance gain. Also, it was noticed that the information from contrast (variance) and structure (covariance) have greater importance than the luminance term in determining the final quality. This can be explained by the fact that human eyes adapt well to luminance changes [18].

Obviously, the ideal scale or level for IQA depends on both the viewing distance and image resolution. A simply yet effective self-adaptive scale transform (SAST) [19] was proposed to simulate the spatial filtering mechanism of the human visual system (HVS). The basic idea is to estimate the suitable scaling parameter from image resolution and viewing distance. Instead of operating in the spatial domain, another recent work [20] relies on discarding part of image details by adaptive high-frequency clipping (AHC) in the discrete wavelet transform (DWT) domain.

3.2.2.2 Saliency-Based FR-IQA

Visual attention is a fundamental property of the HVS and therefore integration of a saliency detection stage into IQA metrics usually leads to improved performance. Early attempts used eye fixation or visual region-of-interest detection data during the pooling stage. WSSIM [21] weights SSIM with saliency map for the LIVE database. And recently the eye fixation maps for other popular image quality databases were provided in [22].

It is apparent that distortion or artifacts also attract visual attention. So using the original image alone for saliency detection is not enough. A newly proposed metric S_NW-SSIM [23] combined saliency features from both the original and distorted images with a nonlinear model [24]. This combined saliency map also leads to performance gain of IQA metrics.

Recently proposed statistical information-content weighted SSIM (IW-SSIM) [25] achieved good performance and become currently the *de facto* benchmark for pooling-type of IQA methods. The information content weighting map is computed from the natural scene statistics (NSS) model [26]. A recent interesting finding is that local similarity estimated using IQA metrics directly is also good for weighting IQA metrics themselves [27]. For example, the structural similarity weighted SSIM (SW-SSIM) [27] has outperformed SSIM substantially.

3.2.2.3 Gradient Magnitude-Based FR-IQA

In recent years, many researchers have realized the significance of low-level features in the IQA. For example, the feature similarity index (FSIM) [28] uses phase congruency and gradient magnitude to characterize the local image quality. And the gradient similarity index (GSIM) [29] measures the changes of gradient similarity of images. Gradient magnitude similarity deviation (GMSD) was proposed in [30]. It was pointed out that the spatial distribution of distortion levels has impact on perceptual quality, i.e. unevenly distributed of distortion degrades visual quality more severely and a local-tuned-global model using gradient information was proposed in [31]. The metric in [31] adopted the Scharr operator [32], which is essentially convolution and the gradient magnitude (GM) is computed as

$$G = \sqrt{G_h^2 + G_v^2} \tag{3.3}$$

where G_h and G_v are the partial derivatives of the input image along horizontal and vertical directions using the Scharr operator. A similarity measure that has the merits of being symmetric, bounded and having unique maximum [15] is then used to quantify the difference between GM maps of the original image $\mathbf{r}$ and its contaminated version $\mathbf{d}$

$$G_m(\mathbf{r}, \mathbf{d}) = \frac{2G_r \cdot G_d + C_1}{G_r^2 + G_d^2 + C_1} \tag{3.4}$$

where G_r and G_d indicate the GM of the original and distorted images, and C_1 is a positive constant introduced for numerical stability. Simple global average pooling can be used

$$G_g(\mathbf{r}, \mathbf{d}) = \Phi(G_m) = \frac{1}{M}\sum_{i=1}^{M} G_m(r_i, d_i) \tag{3.5}$$

where M is the total number of pixels in the image, and Φ computes the mean value. The local distortion-based pooling is then applied in a similar fashion, which is defined by

$$G_l(\mathbf{r}, \mathbf{d}) = \Phi(G_s) = \frac{1}{M_s}\sum_{i=1}^{M_s} G_s(r_i, d_i) \tag{3.6}$$

where G_s indicates the highest $s\%$ values in G_m, and M_s is the pixel numbers in G_s. s was selected as 15 in [31]. For color images, before computing the GM, the simple and widely used YIQ color space transform [33] can be adopted

$$\begin{bmatrix} Y \\ I \\ Q \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ 0.596 & -0.274 & -0.322 \\ 0.211 & -0.523 & 0.312 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{3.7}$$

where Y conveys the luminance information, and I and Q contain the chrominance information. Y is used to compute G_l and G_g based on Eqs. (3.5)–(3.6), I and Q are used to measure the distinction of chrominance between the original and distorted images

$$I_m(\mathbf{r},\mathbf{d}) = \frac{2I_{\mathbf{r}} \cdot I_{\mathbf{d}} + C_2}{I_{\mathbf{r}}^2 + I_{\mathbf{d}}^2 + C_2} \qquad Q_m(\mathbf{r},\mathbf{d}) = \frac{2Q_{\mathbf{r}} \cdot Q_{\mathbf{d}} + C_2}{Q_{\mathbf{r}}^2 + Q_{\mathbf{d}}^2 + C_2} \tag{3.8}$$

where $I_{\mathbf{r}}$ and $I_{\mathbf{d}}$ (and $Q_{\mathbf{r}}$ and $Q_{\mathbf{d}}$) represent I (Q) chromatic channels of images $\mathbf{r}$ and $\mathbf{d}$, and C_2 is similar to C_1. Finally, the LTG model combines those components as

$$\mathrm{LTG}(\mathbf{r},\mathbf{d}) = \frac{\Phi(G_s^{\phi_1})}{\Phi(G_m^{\theta_2})} \cdot \Phi(I_m^{\phi_3} \cdot Q_m^{\phi_3}) \tag{3.9}$$

where ϕ_1, ϕ_2, and ϕ_3 are model parameters. Eq. (3.9) can be approximated as

$$\mathrm{LTG}(\mathbf{r},\mathbf{d}) \approx \frac{\Phi(G_s^{\phi_1})}{\Phi(G_m^{\phi_1})} \cdot \Phi(G_m^{\phi_2'}) \cdot \Phi(I_m^{\phi_3} \cdot Q_m^{\phi_3}) \tag{3.10}$$

where $\phi_2' = \phi_1 - \phi_2 > 0$. The first term indicates that, for different images having the same G_g value, more uneven distribution of distortions will result in worse quality. The second term represents the global quality. And the last term is the measure of difference in the chrominance information.

3.2.2.4 Other Model-Based FR-IQA

Visual information fidelity (VIF) [34] was defined as the ratio of the mutual information between the original and distorted images to the information content of the original image itself. In [35, 36], IQA metrics GES and LGPS were proposed in the Gabor transform domain. Most apparent distortion (MAD) [37] works with the detection- and appearance-based strategies. The brain theory and neuroscience were found to be effective in the IQA design, e.g. internal generative mechanism (IGM) [38]. This model classifies an input image into the predictable and disorder regions, before using psychophysical parameters [16] to pool the modified PSNR and SSIM values of two regions above. Very recently, Zhang et al. proposed the Image quality model based on phase and amplitude differences (IPAD) [39] through the analysis of both amplitude and phase.

3.2.3 RR-IQA

As a tradeoff between FR and NR IQA, RR IQA supposes that only partial information of the original image is available. Friston et al. proposed the free energy principle to explain and unify several brain theories in biological and physical sciences about human action, perception, and learning [40, 41]. Similar to the Bayesian brain hypothesis [42] that has been widely used in ensemble learning, the free energy principle makes a basic assumption that the cognitive process is controlled by an internal generative model in the human brain. With this generative model, the human brain can predict those encountered scenes in a constructive manner.

The internal generative model is essentially a probabilistic model that can be separated into a likelihood term and a prior term. Visual perception is then to invert this likelihood term, in order to infer the posterior possibilities of the given scene. It is natural that there always exists a gap between the encountered scene and brain's prediction, because the internal generative model cannot be universal. The gap between the external input and its generative-model-explainable part is closely related to the visual quality of perceptions, and can be used in IQA. A free energy based distortion metric (FEDM) that simulates the internal generative model of human brain was proposed in [43].

It is assumed that the internal generative model $\mathscr{G}$ is parametric for visual perception, and the perceived scenes can be explained by adjusting the vector θ of parameters. Given an input signal I, its "surprise" (determined by entropy) is evaluated by integrating the joint distribution $P(I, \theta|\mathscr{G})$ over the space of model parameters θ

$$-\log P(I|\mathscr{G}) = -\log \int P(I, \theta|\mathscr{G}) d\theta. \tag{3.11}$$

We then introduce a dummy term $Q(\theta|I)$ into both the denominator and numerator in Eq. (3.11) to derive:

$$-\log P(I|\mathscr{G}) = -\log \int Q(\theta|I) \frac{P(I, \theta|\mathscr{G})}{Q(\theta|I)} d\theta. \tag{3.12}$$

Using the Jensen's inequality, we can easily obtain the following relationship from Eq. (3.12):

$$-\log P(I) \leq -\int Q(\theta|I) \log \frac{P(I, \theta)}{Q(\theta|I)} d\theta. \tag{3.13}$$

The upper bound of the right-hand side in Eq. (3.13) is called "free energy"

$$F(\theta) = -\int Q(\theta|I) \log \frac{P(I, \theta)}{Q(\theta|I)} d\theta. \tag{3.14}$$

It is clear that the free energy is a discrepancy measure between the input image and its best explanation inferred by the internal generative model, and it thereby presents itself as a natural proxy for psychically quality of images. A perceptual distance between the reference image **r** and its distorted counterpart **d** can be defined as the absolute difference of the two images in free energy as

$$\mathrm{FEDM}(\mathbf{r}, \mathbf{d}) = \left| F(\hat{\theta}_r) - F(\hat{\theta}_d) \right| \tag{3.15}$$

with

$$\begin{aligned} \hat{\theta}_r &= \arg\min_{\theta_r} F(\theta|\mathscr{G}, \mathbf{r}), \\ \hat{\theta}_d &= \arg\min_{\theta_d} F(\theta|\mathscr{G}, \mathbf{d}). \end{aligned}$$

The $\mathscr{G}$ was chosen to be the linear AR model for its ability to approximate a wide range of natural scenes by varying its parameters and for its simplicity. The AR model is defined as

$$x_n = \chi^k(x_n) \cdot \lambda + \varepsilon_n \tag{3.16}$$

where x_n is a pixel in question, $\chi^k(x_n)$ is a vector consisting of k nearest neighbors of x_n, $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_k)^T$ is a vector of AR coefficients, and ε_n is additive Gaussian noise term with zero mean. So the free energy of the reference image **r** is quantified by the entropy between itself and the predicted version

$$R(x_n) = \chi^k(x_n) \cdot \lambda_{est} \tag{3.17}$$

where λ_{est} is the optimal solution of AR coefficients for x_n estimated with the least square method. The free energy of the distorted image **d** can be computed in a similar fashion.

RR entropic-difference indexes (RRED) [44] and Fourier transform based quality measure (FTQM) [45] were proposed in DWT and discrete Fourier transform (DFT) domains. There also exist some SSIM based methods, e.g. RR-SSIM [46] and structural degradation model (SDM) [47]. The SDM was developed according to an observation that, for most images with various types of distortions, their low-pass filtered version will have different degrees of spatial frequency decrease. This observation reveals one limitation of SSIM that it is not able to distinguish different distortion types. The SDM solves this problem by measuring the similarity between the structural degradation information of original and distorted images, and thus achieves higher IQA performance. Following the definition of local statistics in SSIM [15], μ_I and σ_I denote local mean and variance of **d** with a 2D circularly symmetric Gaussian weighting function $\mathbf{w} = \{w(k,l)|k = -K, \ldots, K, l = -L, \ldots, L\}$, which satisfies $\mathrm{sum}(\mathbf{w}) = 1$ and $\mathrm{var}(\mathbf{w}) = 1.5$ ($\mathrm{sum}(\cdot)$ and $\mathrm{var}(\cdot)$ compute sum and variance). $\bar{\mu}_{\mathbf{d}}$ and $\bar{\sigma}_{\mathbf{d}}$ have the same definitions except that the impulse function was used instead of the Gaussian weighting function. Then, the structural degradation information is given by

$$S_a(\mathbf{d}) = E(\frac{\sigma_{(\mu_{\mathbf{d}}\bar{\mu}_{\mathbf{d}})} + C_1}{\sigma_{(\mu_{\mathbf{d}})}\sigma_{(\bar{\mu}_{\mathbf{d}})} + C_1}) \tag{3.18}$$

$$S_b(\mathbf{d}) = E(\frac{\sigma_{(\sigma_{\mathbf{d}}\bar{\sigma}_{\mathbf{d}})} + C_1}{\sigma_{(\sigma_{\mathbf{d}})}\sigma_{(\bar{\sigma}_{\mathbf{d}})} + C_1}) \tag{3.19}$$

where $E(\cdot)$ is a direct average pooling, $\sigma_{(\mu_{\mathbf{d}}\bar{\mu}_{\mathbf{d}})}$ and $\sigma_{(\sigma_{\mathbf{d}}\bar{\sigma}_{\mathbf{d}})}$ represent the local covariance similar to the definition in SSIM [15], and C_1 is a small constant to avoid dividing by zero.

3.2.4 NR-IQA

Both FR and RR IQA rely on information of the original images, which would be difficult, if not impossible to get at the user-end of the real world visual communication systems. To address this problem, many NR or blind IQA metrics were proposed during the last decade. Wang et al. designed a simple yet effective JPEG quality estimator [48] by quantifying the blockiness and blur level through checking the zero-crossing rate and the average absolute diversity between in-block image samples. For the artifact of additive noise, these years have witnessed the emergence of a number of blind noise estimation algorithms [49–51]. In [52], a quality model was proposed with a pair of edge detectors for vertical and horizontal directions. In [53], the authors computed the edge width in 8×8 blocks to measure the just-noticeable blur (JNB) factor. Inspired by the successfulness of JNB, the cumulative probability of detecting blur (CPDB) algorithm [54] predicts image sharpness by calculating the probability of blurriness at each edge location.

NR perceptual sharpness metrics also exist. In [55], the authors combined spatial and transform-based features to induce a hybrid approach, called spectral and spatial sharpness (S_3). Specifically, the slope of the local magnitude spectrum and total variation is used to form a sharpness map, and then the scalar index of (S_3) is computed as the average of the 1 % highest values in that sharpness map. A transform-based fast image sharpness (FISH) method [56] was introduced based on the evaluation of log-energies in high-frequency DWT subbands followed by a weighted average. Recently, Feichtenhofer et al. developed a perceptual sharpness index (PSI) [57] by analyzing the edge slopes before integrating an acutance measure to model the influence of local contrast information on the perception to image sharpness. In [58], Wang et al. analyzed the local phase coherence (LPC) and pointed out that the phases of complex wavelet coefficients constitute a highly predictable pattern in the scale space in the vicinity of sharp image features, and furthermore, the LPC structure was found to be disrupted by image blur. Based on this idea, Hassen et al. designed a valid LPC-based sharpness index (LPC-SI) [59].

Note that those above reviewed blind IQA metrics are distortion-specific. General-purpose NR IQA has also been intensively studied in recent years. The general-purpose NR IQA can be mainly categorized into two types. The first type extracts effective features from distorted images then adopts a regression process. Examples include NSS based DIIVINE [60], BLIINDS-II [61], and BRISQUE [62] which conducted IQA in DWT, DCT, and spatial domain, respectively.

The NFSDM metric was proposed with an alternative features extraction method [63] that systematically integrates two effective RR FEDM [43] and SDM [47] to eliminate the demand of reference images. Specifically, it will be shown that there exists an approximate linear dependence between the structural degradation information and the free energy feature of natural images. Thirty randomly selected images from the Berkeley database [64] have been used to validate the linear dependence [63]. The advantage of using Berkeley database is that the contents are different from existing IQA databases [5–8] which will be used to testify the IQA metrics. The scatter plot of structural degradation information $\hat{S}_s(\mathbf{r})$ $\check{S}_s(\mathbf{r})$ ($s = \{a1, a3, a5, b1, b3, b5\}$) vs. the free energy feature $F(\mathbf{r})$ of those 30 test images are shown in Fig. 3.6. The linear dependence between the free energy feature and the structural degradation information provides an opportunity to characterize distorted images without original image information. A linear regression model can be used

$$F(\mathbf{r}) = \alpha_s \cdot \hat{S}_s(\mathbf{r}) + \beta_s \tag{3.20}$$

$$F(\mathbf{r}) = \theta_s \cdot \check{S}_s(\mathbf{r}) + \phi_s \tag{3.21}$$

where α_s, β_s, θ_s, and ϕ_s can be estimated with least square method, and the results are listed in Table 3.1.

Then we utilize $\widehat{SS}_s = F(\mathbf{d}) - (\alpha_s \cdot \hat{S}_s(\mathbf{d}) + \beta_s)$ and $\check{S}S_s = F(\mathbf{d}) - (\theta_s \cdot \check{S}_s(\mathbf{d}) + \phi_s)$ to reduce the dependence of original references, due to the fact that both $\widehat{SS}_s$ and $\check{S}S_s$ values of high-quality images (with few distortions) are close to zero, whereas they will be far from zero when distortions become larger. Consequently, we define the first set of twelve features as follows:

$$\begin{cases} f_{01} - f_{06} : \widehat{SS}_s & s = \{a1, a3, a5, b1, b3, b5\} \\ f_{07} - f_{12} : \check{S}S_s & s = \{a1, a3, a5, b1, b3, b5\} \end{cases}.$$

Additionally, the NFEQM correlates well with human ratings on noise and blur images, so we use $F(\mathbf{d})$ as the last feature f_{13} for NR IQA.

The second class of general-purpose NR IQA metrics operates without human ratings. For instance, natural image quality evaluator (NIQE) [65] was developed to estimate the deviations from statistical regularities observed in natural images without any prior knowledge of image contents or distortion types. And quality-aware clustering (QAC) [66] works by learning a set of quality-aware centroids to act as a codebook to compute the quality levels of image patches and infer the quality score of the overall image.

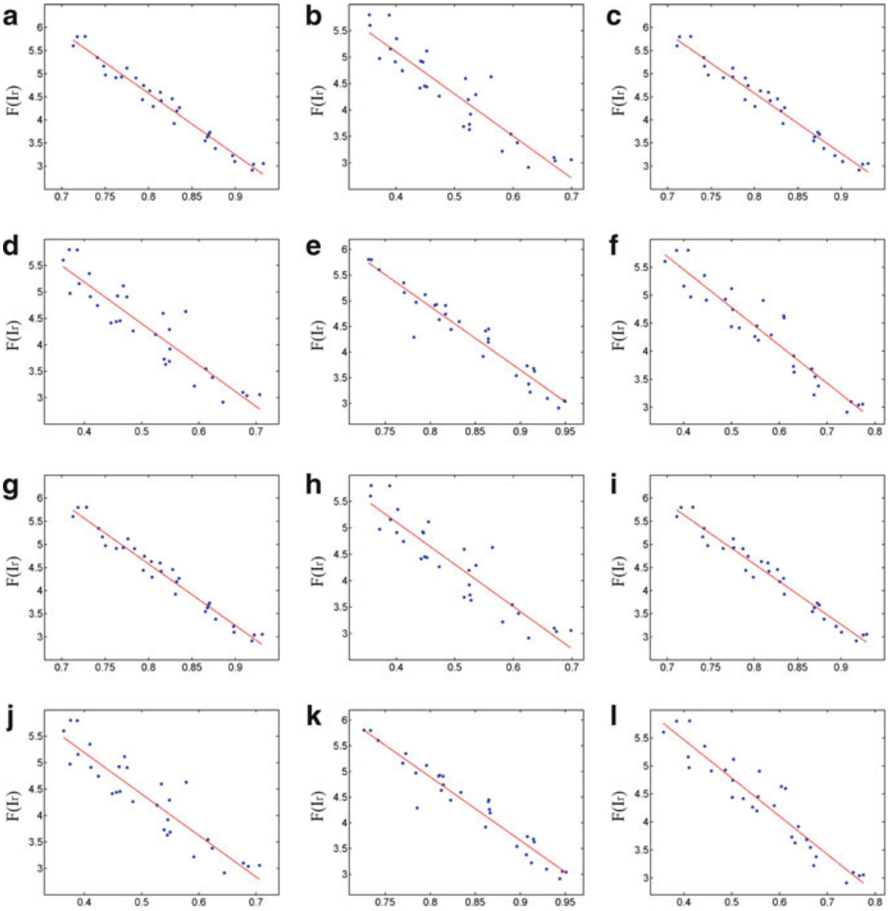


Fig. 3.6 Scatter plots of the structural degradation information $\hat{S}_s(\mathbf{r})$ and $\check{S}_s(\mathbf{r})$ ($s = \{a1, a3, a5, b1, b3, b5\}$) vs. the free energy feature $F(\mathbf{r})$ on 30 images in Berkeley database [64]. **(a)** $\hat{\mathrm{S}}_{\mathrm{a1}}(\mathbf{r})$; **(b)** $\check{\mathrm{S}}_{\mathrm{a1}}(\mathbf{r})$; **(c)** $\hat{\mathrm{S}}_{\mathrm{a3}}(\mathbf{r})$; **(d)** $\check{\mathrm{S}}_{\mathrm{a3}}(\mathbf{r})$; **(e)** $\hat{\mathrm{S}}_{\mathrm{a5}}(\mathbf{r})$; **(f)** $\check{\mathrm{S}}_{\mathrm{a5}}(\mathbf{r})$; **(g)** $\hat{\mathrm{S}}_{\mathrm{b1}}(\mathbf{r})$; **(h)** $\check{\mathrm{S}}_{\mathrm{b1}}(\mathbf{r})$; **(i)** $\hat{\mathrm{S}}_{\mathrm{b3}}(\mathbf{r})$; **(j)** $\check{\mathrm{S}}_{\mathrm{b3}}(\mathbf{r})$; **(k)** $\hat{\mathrm{S}}_{\mathrm{b5}}(\mathbf{r})$; **(l)** $\check{\mathrm{S}}_{\mathrm{b5}}(\mathbf{r})$

Table 3.1 The estimates of parameters α_s, β_s, θ_s and ϕ_s for $\hat{S}_s$ and $\check{S}_s$ ($s = \{a1, a3, a5, b1, b3, b5\}$) using the least square method

	α_s	β_s		θ_s	ϕ_s
$\hat{S}_{a1}$	−13.279	15.194	$\hat{S}_{b1}$	−13.326	15.236
$\hat{S}_{a3}$	−7.9861	8.2961	$\hat{S}_{b3}$	−8.0013	8.3093
$\hat{S}_{a5}$	−13.019	14.988	$\hat{S}_{b5}$	−13.096	15.051
$\check{S}_{a1}$	−7.8427	8.3219	$\check{S}_{b1}$	−7.8451	8.3282
$\check{S}_{a3}$	−12.399	14.808	$\check{S}_{b3}$	−12.378	14.795
$\check{S}_{a5}$	−6.768 7	8.1662	$\check{S}_{b5}$	−6.8255	8.1973

3.3 Emerging Direction in Quality Assessment

3.3.1 Comparative IQA

It is a straightforward task for human observers to judge the relative quality of two visual signals of the same content, but subject to different type/level of distortions. We call this process the comparative IQA (C-IQA). In existing study, the FR and RR IQA methods both need the prior knowledge of the original images while the NR algorithms usually work with a single input image. The CP-IQA approach is inherently different from FR, RR, and NR methods in that it takes as input an image pair and predicts their relative quality without using any knowledge about the original image, as shown in Fig. 3.7.

This C-IQA problem remains a difficult challenge for the current IQA research. To solve this problem, a free energy model based C-IQA approach was proposed to predict the relative perceptual quality of a pair of images with different artifact types/levels [67]. The C-IQA model is designed to emulate the process of comparing the relative quality of two visual stimuli as performed by the HVS within the framework of free energy minimization. The brain's generative models initialized on the inputs are used to explain the two images. And their relative quality can then be determined through comparing the free energy level of this model-data fitting process. As exemplified in Fig. 3.8, $F_{I_i \rightarrow I_j}$ represents the free energy that is computed between the image I_i and the restored image using the generative model derived from the image I_j for restoration. A computationally efficient solution to the proposed C-IQA scheme based on a linear autoregressive image model was also introduced and has shown to achieve about 98 % accuracy in line with the

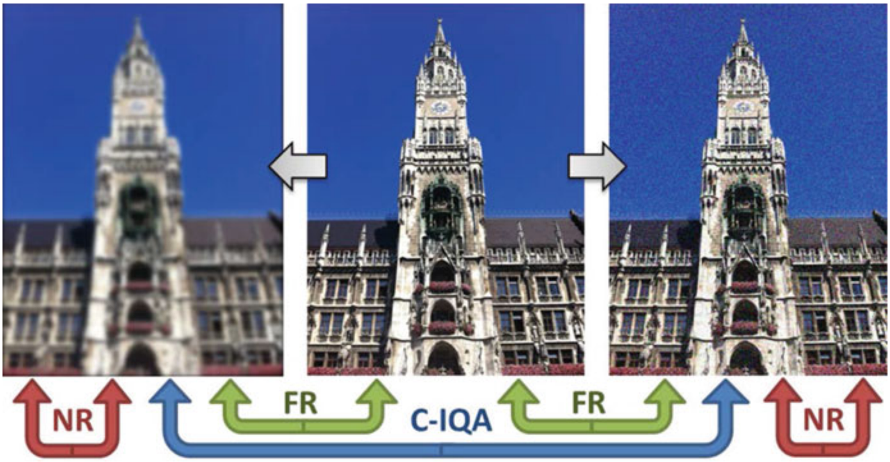


Fig. 3.7 An example to show the difference and connection between C-IQA and FR-/NR-IQA

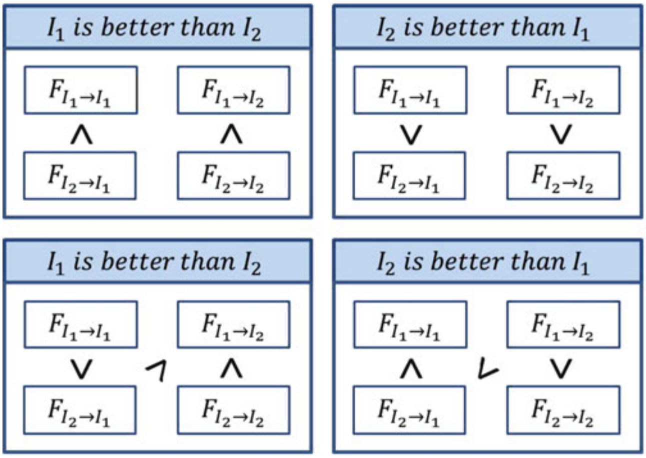

I_1 is better than I_2
$F_{I_1\to I_1}$ $F_{I_1\to I_2}$
$\wedge$ $\wedge$
$F_{I_2\to I_1}$ $F_{I_2\to I_2}$
I_2 is better than I_1
$F_{I_1\to I_1}$ $F_{I_1\to I_2}$
$\vee$ $\vee$
$F_{I_2\to I_1}$ $F_{I_2\to I_2}$
I_1 is better than I_2
$F_{I_1\to I_1}$ $F_{I_1\to I_2}$
$\vee$ $\wedge$
$F_{I_2\to I_1}$ $F_{I_2\to I_2}$
I_2 is better than I_1
$F_{I_1\to I_1}$ $F_{I_1\to I_2}$
$\wedge$ $\vee$
$F_{I_2\to I_1}$ $F_{I_2\to I_2}$

Fig. 3.8 Comparative image quality assessment of I1 and I2 via free energy comparison

subjective ratings when applied on over 300,000 image pairs sampled from the LIVE database, outperforming FR PSNR, SSIM, and some of the most advanced NR IQA algorithms, see [67] for more details.

3.3.2 Multiply Distorted Quality Assessment

As mentioned, though many successful quality metrics, such as SSIM, were reportedly to achieve very high accuracy for various kinds of image distortions, in practice, multiple image distortions tend to occur together and this leads difficulty to previous works of IQA including SSIM and variations. This problem is even more prominent for NR IQA. The LIVEMD database [9] was released with two groups of multiply distorted images, blur followed by JPEG compression and blur followed by noise contamination. In [68] a FIve-Step BLInd Metric (FISBLIM) for quality assessment of multiply distorted images was proposed using several common image processing blocks to simulate the image perceiving process of the human eyes. As presented in Fig. 3.9, the building blocks include scale invariant based noise estimator (SINE) [49] for noise estimation, block-matching and 3D filtering (BM3D) [69] for image denoising, a blur metric [52], a JPEG quality evaluator [48], and a HVS based fusion model. It is worth highlighting that the FISBLIM method is not training based and the performance is robust and not database-dependent.

A linear fusion model of FISBLIM is used to combine the measures of noise, blur, and JPEG:

$$FISBLIM = \alpha \cdot Q_N + \beta \cdot \lambda \cdot Q_B + (1 - \lambda) \cdot Q_J \tag{3.22}$$

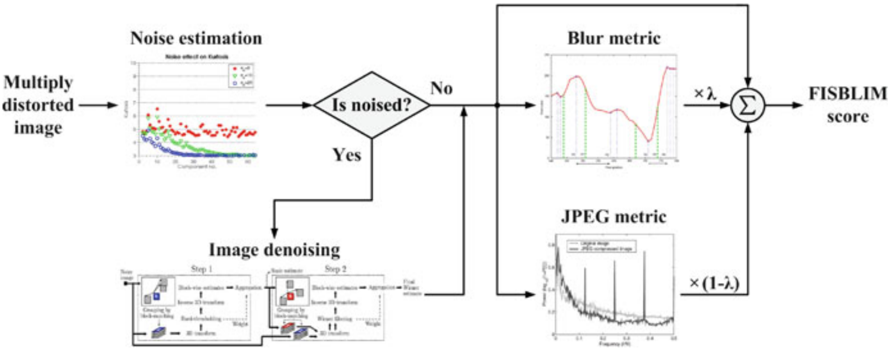


Fig. 3.9 The illustration of primary flowchart of the FISBLIM algorithm

where α, β, λ are model parameters, and Q_N, Q_B, Q_J are objective quality predictions for noise, blur, JPEG.

It is assumed that the HVS has the ability to directly extract noise from an image even under multiple distortions. In other words, the measure of noise is relatively independent of the other two distortion categories. Therefore, SINE based noise estimation is believed to be immune to the influence of blur and JPEG. As shown in Fig. 3.10, all the images in [9] with the four various levels of noise are represented by red, green, blue, and black scatter plots, showing that the noise estimation is largely robust against blur and JPEG compression.

For blur and JPEG compression, the HVS can easily distinguish one from the other but this task is still not easy for computer algorithms. Luckily it was found that the ratio of B_h and A_J can validly separate the JPEG and JPEG plus blurring images from other distortion types. More specifically, B_J is computed as the mean of $|d_h|$ and $|d_v|$ values located in the edge of all the blocks, while A_J is computed for the 6×6 interior part. It is not difficult to conjecture that B_J is nearly equal to A_J for a clean image. On the other hand, B_J is larger than A_J for a JPEG or JPEG plus blurring image. Figure 3.11 displays the relationship between B_J and A_J for all the multiply distorted images in [9]. Among them, red, green, blue, and black scatter plots indicate four different levels of JPEG compression.

Note that the blur metric in [52] can be easily influenced by blockiness, because the basic idea is to measure the spread of the edges in an image. Considering the fact that A_J and Z_J are proposed to measure blur, it is possible to only adopt Q_J

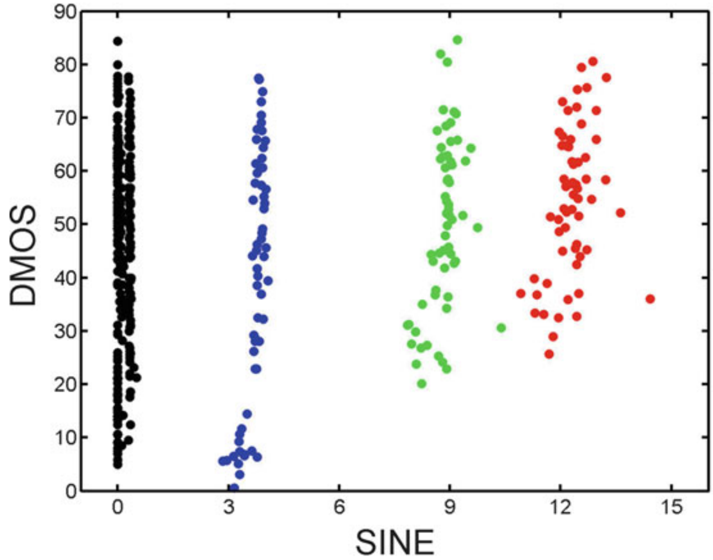


Fig. 3.10 Scatter plots of differential MOS (DMOS) vs. SINE on LIVE multiply distorted database. *Red*, *green*, *blue*, and *black scatter plots* represent four various levels of noise

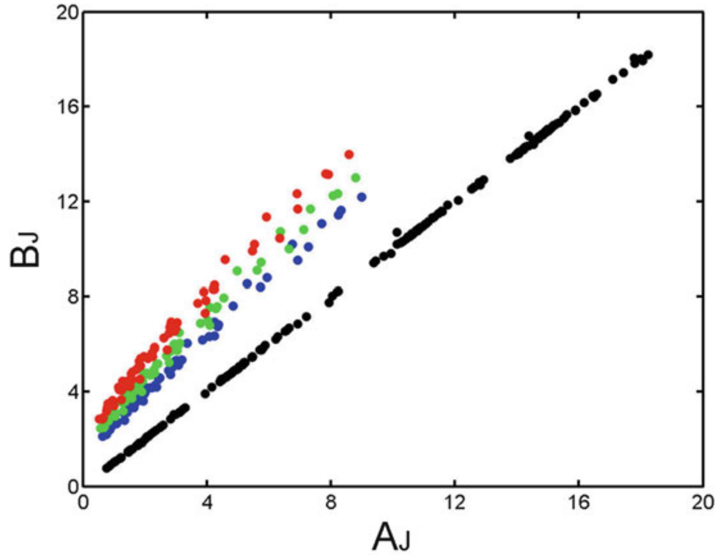


Fig. 3.11 Scatter plots of B_J vs. A_J on LIVE multiply distorted database. *Red*, *green*, *blue*, and *black scatter plots* correspond to four different degrees of JPEG compression

with updated newer values of $\{\kappa_1, \kappa_2, \theta_1, \theta_2, \theta_3\}$ to predict the qualities of JPEG and JPEG plus blur images. The parameter λ that manipulates the choice of using blur metric or JPEG metric is given by

$$\lambda = \begin{cases} 0 & \text{if} \quad B_J/A_J \geq thr \\ 1 & \text{otherwise} \end{cases}. \tag{3.23}$$

where thr is set as an empirical value of 1.5. As shown in Fig. 3.9, JPEG metric will be only applied if $B_J/A_J \geq thr$ is satisfied, otherwise blur metric will be employed.

3.3.3 Contrast-Changed IQA

It is widely known that, for most natural images, appropriate contrast enhancement can usually lead to improved subjective quality. However, contrast change has largely been overlooked in the current research of IQA. To fill this void, a dedicated contrast-changed image database CID2013 was proposed in [10]. As mentioned earlier, the CID2013 database is composed of 400 contrast-changed images of 15 original natural images and the MOSs were recorded from 22 inexperienced viewers.

A novel reduced-reference image quality metric for contrast change (RIQMC) was proposed [10]. RIQMC used entropies and order statistics of the image histograms and outperformed some mainstream IQA methods on existing contrast change related IQA databases.The RIQMC algorithm works in two steps: the computation of entropy and order statistics, and a linear combination. The entropy of an image $\mathbf{d}$ is computed as

$$H(\mathbf{d}) = -\sum_{i=0}^{255} p_i(\mathbf{d}) \cdot \log \ p_i(\mathbf{d}) \tag{3.24}$$

where $p_i(\mathbf{d})$ indicates the probability density of grayscale i in the image $\mathbf{d}$. And the entropy of original image $\mathbf{r}$ is denoted as $H(\mathbf{r})$ following the same definition. Then, the first order statistic or the mean of an image $\mathbf{d}$ is defined using a Gaussian kernel as

$$F_1(\mathbf{d}) = a_1 \cdot exp[-(\frac{E(\mathbf{d}) - a_2}{a_3})^2] \tag{3.25}$$

where a_1, a_2, a_3 are model parameters. Besides, as inspired by the concept of expected contrast in OCTM [70], the second order statistic term was defined in the RIQMC algorithm as

$$F_2(\mathbf{d}) = \sigma^2(\tilde{\mathbf{d}}) = E(\tilde{\mathbf{d}}^2) - E(\tilde{\mathbf{d}})^2 \tag{3.26}$$

where $\tilde{\mathbf{d}}$ is the image histogram. According to the neural mechanism of surface quality perception [71], the on-center and off-center cells and an accelerating nonlinearity in the HVS compute the subband skewness to estimate the perceptual quality of surface. So the third order statistic (skewness) in RIQMC is computed as

$$F_3(\mathbf{d}) = skewness(\mathbf{d}) = \frac{E[(\mathbf{d} - E(\mathbf{d}))^3]}{\sigma^3(\mathbf{d})}. \tag{3.27}$$

Finally, the last fourth order statistic (kurtosis) of the histogram used in the RIQMC algorithm can be evaluated by

$$F_4(\mathbf{d}) = kurtosis(\mathbf{d}) = \frac{E[(\mathbf{d} - E(\mathbf{d}))^4]}{\sigma^4(\mathbf{d})} - 3. \tag{3.28}$$

In the second step, a linear combination was adopted to integrate the aforementioned entropies and order statistics as

$$\text{RIQMC} = \sum_{i=1}^{4} f_i \cdot F_i + f_5 \cdot [H(\mathbf{d}) - H(\mathbf{r})] \tag{3.29}$$

where $f_1 \dots f_5$ and $a_1 \dots a_3$ are parameters controlling the relative importance of each component and can be optimized for each database.

Conclusion

This chapter reviewed some representative method in subjective and objective IQA, with emphasis on recently proposed algorithms. Some classic and new databases for IQA were introduced in the first section. Section two focused on objective quality metrics in the categories of full reference, reduced reference, and no reference. Section three discussed some emerging research topics including comparative quality assessment and quality assessment for contrast changed images.

References

1. A. C. Bovik, “Automatic prediction of perceptual image and video quality,” *Proceedings of the IEEE*, vol. 101, no. 9, pp. 2008–2024, September 2013,
2. Z. Wang and A. C. Bovik, “Mean squared error: Love it or leave it?-A new look at signal fidelity measures,” *IEEE Signal Process. Mag.*, vol. 26, no. 1, pp. 98–117, January 2009.
3. G. Zhai, J. Cai, W. Lin, X. Yang, and W. Zhang, “Three dimensional scalable video adaptation via user-end perceptual quality assessment,” *IEEE Trans. Broadcasting*, vol. 54, no. 3, pp. 719–727, September 2008.

4. G. Zhai, J. Cai, W. Lin, X. Yang, W. Zhang, and M. Etoh, “Cross-dimensional perceptual quality assessment for low bitrate videos,” *IEEE Trans. Multimedia*, vol. 10, no. 7, pp. 1316–1324, November 2008.
5. H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik, “LIVE image quality assessment Database Release 2,” [Online]. Available: http://live.ece.utexas.edu/research/quality
6. N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti, “TID2008-A database for evaluation of full-reference visual quality assessment metrics,” *Advances of Modern Radioelectronics*, vol. 10, pp. 30–45, 2009.
7. E. C. Larson and D. M. Chandler, “Categorical image quality (CSIQ) database,” [Online], Available: http://vision.okstate.edu/csiq
8. N. Ponomarenko, O. Ieremeiev, V. Lukin, K. Egiazarian, L. Jin, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, and C.-C. Jay Kuo, “Color image database TID2013: Peculiarities and preliminary results,” *4th European Workshop on Visual Information Processing EUVIP2013*, pp.106–111, June 2013.
9. D. Jayaraman, A. Mittal, A. K. Moorthy, and A. C. Bovik, “Objective quality assessment of multiply distorted images,” *Proc. IEEE Asilomar Conference on Signals, Systems and Computers*, pp. 1693–1697, November 2012.
10. K. Gu, G. Zhai, X. Yang, W. Zhang, and M. Liu, “Subjective and objective quality assessment for images with contrast change,” *Proc. IEEE Int. Conf. Image Process.*, pp. 383–387, September 2013.
11. M. Liu, G. Zhai, S. Tan, Z. Zhang, K. Gu, and X. Yang, “HDR2014 - A high dynamice range image quality database,” *Proc. IEEE Int. Conf. Multimedia and Expo Workshops*, 2014.
12. Kodak Lossless True Color Image Suite: http://r0k.us/graphics/kodak/
13. H. R. Sheikh, M. F. Sabir, and A. C. Bovik, “A statistical evaluation of recent full reference image quality assessment algorithms,” *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440–3451, November 2006.
14. Z. Wang and A. C. Bovik, “A universal image quality index,” *IEEE Signal Processing Letters*, vol. 9, pp. 81–84, March 2002.
15. Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, April 2004.
16. Z. Wang, E. P. Simoncelli, and A. C. Bovik, “Multi-scale structural similarity for image quality assessment,” *IEEE Asilomar Conference Signals, Systems and Computers*, pp. 1398–1402, November 2003.
17. M. Liu, G. Zhai, K. Gu, Q. Xu, X. Yang, X. Sun, W. Chen, and Y. Zuo, “A new image quality metric based on MIx-Scale transform,” *Proc. IEEE Workshop on Signal Processing Systems*, pp. 266–271, October 2013.
18. E. Peli, “Contrast in complex images,” *Journal of Optical Society of America*, vol. 7, pp. 2032–2040, October 1990.
19. K. Gu, G. Zhai, X. Yang, and W. Zhang, “Self-adaptive scale transform for IQA metric,” *Proc. IEEE Int. Symp. Circuits and Syst.*, pp. 2365–2368, May 2013.
20. K. Gu, G. Zhai, M. Liu, Q. Xu, X. Yang, and W. Zhang, “Adaptive high-frequency clipping for improved image quality assessment,” *Proc. IEEE Visual Communications and Image Processing*, pp. 1–5, November 2013.
21. H. Liu and I. Heynderickx, “Visual attention in objective image quality assessment: Based on eye-tracking data,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 7, pp. 971–982, April 2011.
22. X. Min, G. Zhai, Z. Gao, and K. Gu, “Visual attention data for image quality assessment databases,” *Proc. IEEE Int. Symp. Circuits and Syst.*, 2014.
23. K. Gu, G. Zhai, X. Yang, L. Chen, and W. Zhang, “Nonlinear additive model based saliency map weighting strategy for image quality assessment”, *IEEE International Workshop on Multimedia Signal Processing*, pp. 313–318, September 2012.

24. L. Itti, C. Koch, and E. Niebur, “A model of saliency-based visual attention for rapid scene analysis,” *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 1254–1259, November 1998.
25. Z. Wang and Q. Li, “Information content weighting for perceptual image quality assessment,” *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1185–1198, 2011.
26. E. P. Simoncelli and B. A. Olshausen, “Natural image statistics and neural representation,” *Annual Review of Neuroscience*, vol. 24, no. 1, pp. 1193–1216, 2001.
27. K. Gu, G. Zhai, X. Yang, W. Zhang, and M. Liu, “Structural similarity weighting for image quality assessment,” *Proc. IEEE Int. Conf. Multimedia and Expo Workshops*, pp. 1–6, July 2013.
28. L. Zhang, L. Zhang, X. Mou, and D. Zhang, “FSIM: A feature similarity index for image quality assessment,” *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, August 2011.
29. A. Liu, W. Lin, and M. Narwaria, “Image quality assessment based on gradient similarity,” *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 1500–1512, April 2012.
30. W. Xue, L. Zhang, X. Mou, and A. C. Bovik, “Gradient magnitude similarity deviation: A highly efficient perceptual image quality index,” *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 684–695, February 2014.
31. K. Gu, G. Zhai, X. Yang, J. Zhou, X. Gu, and W. Zhang, “An efficient color image quality metric with local-tuned-global model,” *Proc. IEEE Int. Conf. Image Process.*, 2014.
32. B. J*ä*hne, H. Haubecker, and P. Geibler, *Handbook of Computer Vision and Applications.* New York: Academic, 1999.
33. C. Yang and S. H. Kwok, “Efficient gamut clipping for color image processing using LHS and YIQ,” *Optical Engineering*, vol. 42, no. 3, pp. 701–711, March 2003.
34. H. R. Sheikh, and A. C. Bovik, “Image information and visual quality,” *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, February 2006.
35. G. Zhai, W. Zhang, X. Yang, S. Yao, and Y. Xu, “GES: a new image quality assessment metric based on energy features in Gabor transform domain,” *Proc. IEEE Int. Symposium on Circuits and Systems*, pp. 1715–1718, 2006.
36. G. Zhai, W. Zhang, Y. Xu, and W. Lin, “LGPS: Phase based image quality assessment metric,” *IEEE Workshop on Signal Processing Systems*, pp. 605–609, 2007.
37. E. C. Larson and D. M. Chandler, “Most apparent distortion: Full-reference image quality assessment and the role of strategy,” *Journal of Electronic Imaging*, vol. 19, no. 1, March 2010.
38. J. Wu, W. Lin, G. Shi, and A. Liu, “Perceptual quality metric with internal generative mechanism,” *IEEE Trans. Image Process.*, vol. 22, no. 1, pp. 43–54, January 2013.
39. F. Zhang, W. Jiang, F. Autrusseau, and W. Lin, “Exploring V1 by modeling the perceptual quality of images,” *Journal of Vision*, vol. 14, no. 1, pp. 1–14, 2014.
40. K. Friston, J. Kilner, and L. Harrison, “A free energy principle for the brain,” *Journal of Physiology Paris*, vol. 100, pp. 70–87, 2006.
41. K. Friston, “The free-energy principle: A unified brain theory?” *Nature Reviews Neuroscience*, vol. 11, pp. 127–138, 2010.
42. D. C. Knill and A. Pouget, “The Bayesian brain: The role of uncertainty in neural coding and computation,” *Trends Neurosci.*, vol. 27, no. 12, pp. 712–719, 2004.
43. G. Zhai, X. Wu, X. Yang, W. Lin, and W. Zhang, “A psychovisual quality metric in free-energy principle,” *IEEE Trans. Image Process.*, vol. 21, no. 1, pp. 41–52, January 2012.
44. R. Soundararajan and A. C. Bovik, “RRED indices: Reduced-reference entropic differencing for image quality assessment,” *IEEE Trans. Image Process.*, vol. 21, no. 2, pp. 517–526, February 2012.
45. M. Narwaria, W. Lin, I. V. McLoughlin, S. Emmanuel, and L. T. Chia, “Fourier transform-based scalable image quality measure,” *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3364–3377, August 2012.
46. A. Rehman and Z. Wang, “Reduced-reference image quality assessment by structural similarity estimation,” *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3378–3389, August 2012.

47. K. Gu, G. Zhai, X. Yang, and W. Zhang, “A new reduced-reference image quality assessment using structural degradation model,” *Proc. IEEE Int. Symp. Circuits and Syst.*, pp. 1095–1098, May 2013.
48. Z. Wang, H. R. Sheikh, and A. C. Bovik, “No-reference perceptual quality assessment of JPEG compressed images,” *Proc. IEEE Int. Conf. Image Process.*, pp. 477–480, September 2002.
49. D. Zoran and Y. Weiss, “Scale invariance and noise in natural images,” *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 2209–2216, September 2009.
50. G. Zhai and X. Wu, “Noise estimation using statistics of natural images,” *Proc. IEEE Int. Conf. Image Process.*, pp. 1857–1860, September 2011.
51. X. Liu, M. Tanaka, and M. Okutomi, “Single-image noise level estimation for blind denoising,” *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 5226–5237, December 2013.
52. P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi, “A no-reference perceptual blur metric,” *Proc. IEEE Int. Conf. Image Process.*, vol. 3, pp. 57–60, 2002.
53. R. Ferzli and L. Karam, “A no-reference objective image sharpness metric based on the notion of just noticeable blur (JNB),” *IEEE Trans. Image Process.*, vol. 18, no. 4, pp. 717–728, April 2009.
54. N. D. Narvekar and L. J. Karam, “A no-reference image blur metric based on the cumulative probability of blur detection (CPBD),” *IEEE Trans. Image Process.*, vol. 20, no. 9, pp. 2678–2683, September 2011.
55. C. Vu, T. Phan, and D. Chandler, “S_3: A spectral and spatial measure of local perceived sharpness in natural images,” *IEEE Trans. Image Process.*, vol. 21, no. 3, pp. 934–945, March 2012.
56. P. Vu and D. Chandler, “A fast wavelet-based algorithm for global and local image sharpness estimation,” *IEEE Signal Process. Lett.*, vol. 19, no. 7, pp. 423–426, July 2012.
57. C. Feichtenhofer, H. Fassold, and P. Schallauer, “A perceptual image sharpness metric based on local edge gradient analysis,” *IEEE Signal Process. Lett.*, vol. 20, no. 4, pp. 379–382, April 2013.
58. Z. Wang and E. Simoncelli, “Local phase coherence and the perception of blur,” *in Advances in Neural Information Processing Systems*, vol. 16, pp. 1–8. Cambridge, MA, USA: MIT Press, May 2004.
59. R. Hassen, Z. Wang, and M. Salama, “Image sharpness assessment based on local phase coherence,” *IEEE Trans. Image Process.*, vol. 22, no. 7, pp. 2798–2810, July 2013.
60. A. K. Moorthy and A. C. Bovik, “Blind image quality assessment: From scene statistics to perceptual quality,” *IEEE Trans. Image Process.*, pp. 3350–3364, vol. 20, no. 12, December 2011.
61. M. A. Saad, A. C. Bovik, and C. Charrier, “Blind image quality assessment: A natural scene statistics approach in the DCT domain,” *IEEE Trans. Image Process.*, pp. 3339–3352, vol. 21, no. 8, August 2012.
62. A. Mittal, A. K. Moorthy, and A. C. Bovik, “No-reference image quality assessment in the spatial domain,” *IEEE Trans. Image Process.*, pp. 4695–4708, vol. 21, no. 12, December 2012.
63. K. Gu, G. Zhai, X. Yang, W. Zhang, and L. Liang, “No-reference image quality assessment metric by combining free energy theory and structural degradation model,” *Proc. IEEE Int. Conf. Multimedia and Expo*, pp. 1–6, July 2013.
64. D. Martin, C. Fowlkes, D. Tal, and J. Malik, “A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics,” *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 416–423, 2001.
65. A. Mittal, R. Soundararajan, and A. C. Bovik, “Making a ‘completely blind’ image quality analyzer,” *IEEE Signal Process. Letters*, pp. 209–212, vol. 22, no. 3, March 2013.
66. W. Xue, L. Zhang, and X. Mou, “Learning without human scores for blind image quality assessment,” *Proc. IEEE Int. Conf. Comput. Vis. and Pattern Recognition*, pp. 995–1002, June 2013.
67. G. Zhai and A. Kaup, “Comparative image quality assessment using free energy minimization,” *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 1884–1888, 2013.

68. K. Gu, G. Zhai, M. Liu, X. Yang, W. Zhang, X. Sun, W. Chen, and Y. Zuo, "FISBLIM: A five-step blind metric for quality assessment of multiply distorted images," *Proc. IEEE Workshop on Signal Processing Systems*, pp. 241–246, October 2013.
69. K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-D transform-domain collaborative filtering," *IEEE Trans. Image Process.*, vol. 16, no. 8, pp. 2080–2095, August 2007.
70. X. Wu, "A linear programming approach for optimal contrast-tone mapping," *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1262–1272, May. 2011.
71. I. Motoyoshi, S. Nishida, L. Sharan, and E. H. Adelson, "Image statistics and the perception of surface qualities," *Nature*, vol. 447, pp. 206–209, May 2007.

Chapter 4
Quality Assessment of Mobile Videos

Manri Cheon, Soo-Jin Kim, Chan-Byoung Chae, and Jong-Seok Lee

4.1 Introduction

Recently, video consumption using mobile devices has been very popular due to the technological advances of mobile devices capable of producing and consuming high quality videos and high speed wireless communication networks. Many mobile devices are now able to capture and display high definition (HD) videos. Moreover, high speed communication technologies such as long-term evolution (LTE) are available in service. One of the most critical issues in mobile video delivery services is how to maximize the quality of experience (QoE) of the users for the delivered video contents. Traditionally, quality of service (QoS) has been primarily used for designing video communication systems, which measures the performance of the systems using technical parameters at the application and network levels, such as delay, bit error rate, etc. However, when systems are optimized in the perspective of human observers, it is more appropriate to consider perceptual aspects related to QoE rather than QoS parameters, since there exists a gap between the perceived quality and QoS [14].

In general, various factors are involved in determining QoE of mobile videos. Above all, since video compression is usually performed prior to transmission in order to reduce the data rate, compression artifacts including blockiness, blurring, and ringing are included in the delivered videos, which may degrade the perceived quality of the videos.

M. Cheon • S.-J. Kim • C.-B. Chae (✉) • J.-S. Lee (✉)
School of Integrated Technology, Yonsei University, Incheon 406-840, South Korea
e-mail: manri.cheon@yonsei.ac.kr; Soojin.Kim@yonsei.ac.kr; cbchae@yonsei.ac.kr; jong-seok.lee@yonsei.ac.kr

C. Deng et al. (eds.), *Visual Signal Quality Assessment*,
DOI 10.1007/978-3-319-10368-6_4

Then, failure in smooth data transmission may cause visible artifacts at the user side, which results in different types of artifacts according to the transmission protocol. When a protocol that does not guarantee transmission reliability, e.g., real-time transport protocol (RTP) with user datagram protocol (UDP), is used, loss of a data packet is typically compensated for during decoding of the video at the user terminal via error concealment. In most cases, perfect concealment is difficult, and thus artifacts due to imperfect concealment may become visible at the spatio-temporal location corresponding to the lost packet and its neighbor in the video. In comparison with the compression artifacts appearing globally in the video, packet loss artifacts tend to remain local; the range that the artifacts appear depends on the predictive coding structure configured during encoding. On the other hand, when a protocol ensuring transmission reliability, e.g., transmission control protocol (TCP), is used, retransmission is requested for the erroneous packet until it is properly transmitted, which may make the video buffer empty. As a result, artifacts in the temporal domain such as freezing and jitter may occur.

For adaptive video transmission, video scalability may be employed, meaning spatial, temporal, or quality resolution can be adaptively changed during transmission [43]. When the temporal resolution, i.e., frame rate, is reduced in order to reduce the data rate (e.g., from 30 fps to 15 fps or even lower), jerkiness artifacts may be experienced by the user.

The communication channel status tends to change over time, and the amount of the aforementioned artifacts specific to mobile video communications also tends to vary over time according to the channel status. Employment of adaptive streaming (e.g., dynamic adaptive streaming over HTTP (DASH) [72]) may also cause temporal changes of video quality, as it allows adaptive switching between different versions of the same content according to the channel status. Such temporal variations of the amount and type of quality degradation may also affect perceived quality negatively.

In summary, the quality of mobile videos is influenced by diverse factors, which have quite different characteristics and thus affect QoE in different ways. Moreover, several factors may be involved in simultaneously, and thus their interplay also needs to be considered during quality assessment of mobile videos.

In order to deal with this extremely complicated issue, there have been many researches regarding subjective and objective quality assessment of mobile videos, which are reviewed in this article. The results of the efforts toward understanding perceptual effects of various quality factors are surveyed in Sect. 4.2. Then, objective quality metrics estimating perceived quality based on the subjective quality assessment results are introduced in Sect. 4.3. In Sect. 4.4, publicly available databases for mobile video quality assessment are summarized. Finally, conclusion is made along with future challenges in Sect. 4.5.

4.2 Subjective Quality Assessment of Mobile Videos

4.2.1 Methodology and Environment

As human subjects usually act as end users of digital contents, subjective tests are performed to measure the perceived quality in the context of multimedia services and applications.

Subjective experiments have to be carried out with scientific rigor. They must be conducted in controlled environments with a significantly large number of subjects by following a methodology suitable for the test objective, in order to ensure reproducibility and reliability of the results. Also, the test material needs to be carefully selected, including diverse contents spanning all quality levels evenly, if possible. International standards provide guidelines for subjective test activities (e.g., International Telecommunication Union Radiocommunication Sector [ITU-R] BT.500-13). Many existing researches for quality assessment of mobile videos also followed these guidelines, using desktop screens in controlled laboratory environments. However, the mobile environments are different in various aspects in terms of the types of used devices, the degree of concentration of users, lighting conditions, etc., which affect the quality perception. Recent researches have been trying to consider these in subjective quality assessment.

In [33], user-oriented subjective quality evaluation for mobile videos in its usage contexts and comparison with a laboratory experiment were performed. The quality evaluation was conducted in terms of acceptability (acceptable or not) and satisfaction of quality (11-grade scale) under three different tasks and environments following selected typical mobile videos usage situations: waiting or killing time at the railway station, relaxing in a cafe, and taking a local bus to the predefined location including transitions by foot. Test video sequences were encoded using H.264/AVC and corrupted by four different transmission error rates. It was shown that the evaluations were more favorable and less discriminative in the mobile contexts compared to the laboratory environment: In terms of acceptability, people accepted higher transmission error rates in the real contexts; in terms of satisfaction, there was no significant difference of overall satisfaction ratings between the two studies, but the score differences between the packet loss rates were smaller in the mobile contexts than in the laboratory experiment.

Thus, the importance of considering uncontrolled experimental environments has been increased. In [58], subjective evaluation of mobile videos watched in a natural environment was performed. In addition to subjective video quality evaluation, the work focused on usage patterns in a natural research context, which includes the behavioral factors of watching mobile videos. The participants were able to watch the videos using mobile phones when they wanted and where they wanted. User evaluations were gathered by means of questionnaires on the device, complemented with traditional pen and paper diaries. It was shown that most videos were watched at home and in the afternoon and evening. In terms of the acceptability of the video quality, there were no significant differences according to the physical context of the users.

Effects of the device have been investigated in several studies. Khan et al. [35] assessed the impact of devices on video quality assessment. Subjective tests were carried out on two devices, PC and mobile phone, using QCIF video sequences. Test sequences were generated with a combination of parameters associated with the access network (block error rates and mean burst lengths), H.264/AVC codec related parameters (sender bit rates), and content types. The results showed that MOS values obtained from PC-based tests are relatively higher than those from mobile phone based tests in nearly all test scenarios. In [48], quality perception on a mobile phone and a laptop was compared. Test sequences were encoded with the H.264 baseline profile, and streamed through an emulated network with packet loss and packet delay variation. Unlike [35], the results did not reveal any strong evidence to conclude that devices have any impact on user perception when the spatial resolution is fixed. Two types of devices, mobile phone and tablet, were used for subjective quality assessment in the study of Moorthy et al. [50]. They compared the DMOS scores for the two devices. Statistical hypothesis tests indicated that while the results for the two cases are correlated and statistically indistinguishable, the degree of correlation is a function of the distortion category. Specifically, for the frame-freeze case, the perception of visual quality varies significantly as a function of the display resolution. However, for the other types of distortion such as compression, rate adaptation, temporal dynamics, and wireless packet loss, the perception varies insignificantly. In [46], it was observed that subjects had different reactions to the videos depending on the type of video scalability and the type of devices. For example, when the video has smaller spatial resolution by video scalability, perceived scores are better for watching in mobile phones than in tablets.

There exist studies about the quality assessment methodology for mobile videos. Tominaga et al. [73] compared eight popular subjective assessment methods, namely, the double-stimulus continuous quality-scale (DSCQS), double-stimulus impairment scale (DSIS), absolute category rating method with a 5-grade scale (ACR5), ACR5 with hidden reference (ACR5-HR), ACR 11-grade scale (ACR11), ACR11 with hidden reference (ACR11-HR), subjective assessment of multimedia video quality (SAMVIQ), and SAMVIQ with hidden reference (SAMVIQ-HR), under diverse mobile video scenarios. They evaluated the total assessment time and difficulty/easiness, as well as the characteristics of different rating scales and their statistical reliability. They concluded that ACR5 is the most suitable method considering rating scale, time, and ease of evaluation for subjective quality assessment of mobile video services. In [57], ACR and SAMVIQ subjective methodologies were compared for different spatial resolutions. It was shown that the correlation between their result scores is weaker when the spatial resolution increases; since SAMVIQ allows multiple viewing unlike ACR, the former helps subjects to examine test videos more thoroughly than the latter and, consequently, the gap of quality perception between the two methodologies could become larger for larger resolutions.

4.2.2 Quality Perception of Mobile Videos

4.2.2.1 Quality of Compression and Transmission Artifacts

Generally, compression is an indispensable element for video services. Thus, there exist many researches about subjective quality assessment of codecs or compression techniques in mobile scenarios. For instance, in [16], subjective quality assessment was conducted to compare the two latest codecs, H.264/AVC and HEVC, under two cellular bit rate conditions. It was shown that better quality is obtained by using the latest codec, i.e., HEVC, for mobile videos under the same bit rate conditions.

With compression artifacts, transmission artifacts, e.g., packet loss, delay, buffering, etc., are important factors affecting the quality of mobile videos. Mobile environments generally refer networked situations and transmission errors occurred occasionally. Thus, many researches were performed to consider both compression and transmission artifacts.

Minhas et al. [48] studied effects of packet loss and packet delay variation on QoE. It was found that quality of videos encoded with the H.264/AVC baseline profile is affected sensitively by both types of the network disturbance. The results showed that for packet loss rates above 5 %, the users rated the videos as "bad." And, delays above ± 8 ms were rated as "bad." It was also shown that the sensitivity in quality perception was higher for delay variation than packet loss. In [30], subjective tests were performed in the case of broadcast digital television (DTV). It was shown that there is no correlation between the percentage of packet loss and the subjective ratings. Rather, the lost slice type (I, P, or B) has a significant impact on the perceived quality degradation.

Interaction between compression and packet loss artifacts in quality perception was investigated in [38], where subjective quality assessment of H.264/AVC video streaming with packet loss was performed by using the paired comparison methodology. The results showed that, in general, both artifacts significantly degrade the perceived quality unless the coding artifacts are already so severe that addition of packet loss artifacts does not cause further quality degradation. Similar results were obtained in [44]. The work examined the impact of error length, peak signal-to-noise ratio (PSNR) drop (due to H.264/AVC encoding), and loss location on the perceptual quality of the decoded video with a single transmission loss. It was shown that an error is visible only if the error length or PSNR drop exceeds a certain threshold. And, perceptual quality was approximately linear with respect to the PSNR drop and error length. Additionally, it was also shown that, when a video sequence contains multiple losses, the perceptual quality depends on the sum of PSNR drops of individual transmission losses, as well as loss pattern. However, unlike [38], it was not shown that which type of artifacts, compression or packet loss, degrades the perceived quality more significantly.

There exist researches studying effects of temporal patterns of transmission errors on perceived quality. In [3], the authors performed a series of subjective video quality tests that evaluated the quality of H.264/AVC videos distorted by several packet loss patterns. The results showed that for a fixed loss rate, multiple bursty losses are more damaging than a single contiguous long loss. The influence of rebuffering interruptions on the user's QoE was investigated by De Pessemier et al. [59]. In this study, six scenarios combining three (low, medium, and high) simulated bandwidths representing a universal mobile telecommunication system (UMTS), high speed packet access (HSPA), and WiFi communication channel, respectively, and two quality levels (high and low) were considered. Although video interruptions due to rebufferings were experienced as disturbing, users accepted a (limited) number of these rebufferings in a mobile context. The high quality video sources that are sent over a low bandwidth connection and thereby require numerous rebufferings during video playback were in general evaluated as "unacceptable." And subjective quality assessment results were highly correlated to the objectively measured parameters of the video session, such as the number of rebufferings, the rebuffer time, and the loading time of the video. Moreover, it was shown that the users preferred a fluent playback of the video to a higher resolution, frame rate, and bit rate.

Moorthy et al. [50] performed video quality assessment including subjective, behavioral, and objective studies. Test video sequences had a HD resolution, which is popular for mobile videos in these days, and included diverse distortion types reflecting mobile scenarios, e.g., compression, wireless packet loss, and dynamically varying distortions such as frame freezes and temporally varying compression rates. These test sequences were displayed on two types of mobile devices, mobile phone, and tablet. It was shown that time-varying quality has a definite negative impact on human subjective judgments of quality, and this impact is a function of the frequency of significant distortion changes and of the differences in quality levels between segments.

4.2.2.2 Quality of Video Scalability

Scalability is useful to deal with the limit in the network capacity and user heterogeneity in terms of the network environment and terminal capability (e.g., decoding and display capabilities). It allows the same content to be efficiently delivered in different formats simultaneously to multiple users. Generally, scalability considers three dimensions, spatial dimension, temporal dimension, and quality dimension.

It is challenging to understand and model human quality perception of video scalability in a multidimensional space involving spatial, temporal, and quality variations (and their corresponding artifacts) as well as application- and content-dependent expectations of users. Moreover, expressing and comparing quality across scalability dimensions on a unified scale is not straightforward. A thorough survey of existing studies for quality perception of video scalability is given in [43].

Among the scalability dimensions, subjective assessment of temporal scalability has been conducted the most extensively. As a general conclusion of the related studies, the threshold of subjective acceptability seems to be around 15 Hz, but its exact value varies with content, application, viewers, and so on [8]. In [54], the subjective quality assessment was performed using a mobile phone considering the impact of the three dimensions of scalability. It was found that the temporal resolution affects the quality independently of spatial resolution and quantization step size, while there is significant interaction between spatial resolution and quantization step size. In [20], the scope of validity of PSNR as a video quality metric was examined using subjective experimental data. The work showed that PSNR is inaccurate in measuring video quality of a video content encoded at different frame rates because it is not capable of assessing the perceptual trade-off or interaction between the spatial and temporal qualities.

Regarding the relationship between the temporal and signal-to-noise ratio (SNR) scalability dimensions, it is traditionally believed that a high frame rate is more important than high frame quality for contents containing fast motion. Thus, reduction of the frame rate decreases subjective quality only slightly for slow motion contents [74]. However, other studies showed results contradicting this belief. In [75], it was shown that the preference of frame rate against frame quality varies according to the bit rate condition. The boundaries between the bit rate ranges were higher for complex scenes. It implies that reaching a certain satisfiable level of frame quality has priority over increasing the frame rate under a limited bit rate budget. It should be noted that the conclusion in [74] was not based on analysis for fixed bit rate conditions, which explains its inconsistency with that in [75]. The relative importance of spatial quality and frame rate on perceived quality was examined in [39] via pairwise comparisons to find the preferred path from bad quality to good quality, or vice versa. It was shown that there is a strong correlation between temporal complexity of content and perceived importance of frame rate.

The trade-off relation between the spatial and temporal dimensions was investigated in [10]. Sequences with different combinations of spatial and temporal resolutions were produced for each fixed bit rate condition, and their relative subjective preferences were obtained. It was shown that for fast motion contents, the frame rate is more important than the frame size.

The relation of the spatial resolution and frame quality was studied in [74]. It was shown that a small frame size with high frame quality is preferable to a large size with low quality when the bit rate is not sufficiently high. In this work, smaller spatial resolutions were not upscaled to the original size. Thus, it is not straightforward to compare these results with those in other studies using spatial upscaling. A similar study was also performed in [79], focusing on the effect of the spatial resolution alone and the combined effect of the spatial resolution and quantization artifacts. Videos having different spatial resolutions were displayed at the full screen size of a mobile device. The results showed that the dropping rate of the perceived quality due to reduction of the spatial resolution increases as QP increases and the dropping rate of the perceived quality due to increase of QP increases as the spatial resolution reduces.

Subjective quality assessment for all three scalability dimensions has been considered recently [42, 54, 84]. In [84], an extensive subjective experiment was conducted for low bit rate videos. The results showed that for a fixed bit rate, the frame size should be kept low, while a low frame rate is preferable for fast motion contents, which supports the aforementioned results of [75]. When the frame rate is high (e.g., 30 Hz) while the frame size is small for low bit rate conditions, improvement in the SNR dimension is usually the most efficient to enhance perceived quality rather than improvement in the spatial dimension. Similarly, when the frame size is large (e.g., CIF) while the frame rate is low, perceived quality is enhanced more efficiently by improving picture quality in the SNR dimension than by increasing the frame rate. In [42], subjective quality assessment of scalable video coding was performed via a paired comparison methodology, which investigates the influence of the combination of scalability options on perceived quality for an adaptive strategy that selects the optimal combination for a given bandwidth constraint. It was shown that the priority between the spatial resolution and frame rate depends on the bit rate condition and content type, which was considerably consistent in the used two types of codecs, scalable extension of H.264/AVC and wavelet-based scalable video coding. For low bit rate conditions, the spatial resolution was important for perceived quality, whereas for higher bit rate conditions, a high frame rate was preferable. The results of [54] show that the rate of quality degradation along the temporal dimension is independent of spatial resolution and quantization step size, and vice versa. The rate of quality degradation along the spatial dimension is a linear function of quantization step size.

Although it is not easy to directly compare the aforementioned studies, their results can be roughly summarized as follows. The trade-off is basically between frame rate and frame quality, considering that low spatial resolutions are upscaled to a maximum resolution. Thus, the frame quality is affected by both coding artifacts and blurring due to upscaling. It seems that there is a bit rate threshold at which the preference of scalability options is switched. Below the threshold, enhancing the frame quality has the priority by improvement in either the SNR or spatial dimension. Above the threshold, the frame quality reaches a certain satisfactory level and, thus, the frame rate becomes more critical for perceived quality. Here, the threshold is mainly dependent on the content characteristics. It is higher for contents containing faster motion because more bits are needed to achieve an acceptable level of quality for this kind of contents, but it is also affected by the encoder type, viewing environment, user expectation, and so on.

4.3 Objective Quality Assessment of Mobile Videos

4.3.1 General Objective Metrics

Traditionally, signal-based metrics such as PSNR and mean square error (MSE) are widely used for evaluating loss of image quality due to its simplicity and mathematical convenience. However, it is also known that the correlation between these metrics and human judgment of quality is limited [78]. There have been attempts to consider the human visual system (HVS) for obtaining better correlation with human judgment. Examples include multi-scale structural similarity (MS-SSIM) [76], video quality metric (VQM) [23], and motion-based video integrity evaluation (MOVIE) [68]. The structural similarity (SSIM) image quality paradigm is based on the assumption that the HVS is highly adapted for extracting structural information from the scene, and therefore a measure of structural similarity can provide a good approximation to perceived image quality. MS-SSIM supplies more flexibility than the single-scale SSIM by incorporating the variations of viewing conditions. The National Telecommunications and Information Administration implemented the general VQM as a means for quantifying perceptual quality degradation in video systems that utilize compression. This is standardized in ITU-T J.144 [23]. MOVIE evaluates dynamic video fidelity that integrates both spatial and temporal aspects of distortion assessment.

In [50], various general objective image and video quality metrics were compared through an experiment using a tablet and a mobile phone. Five different distortion types, namely compression, rate adaptation, temporal dynamics, wireless channel packet loss, and all, were considered. When a mobile phone was used, visual information fidelity (VIF) [70], which measures the mutual information between the input and the output of the HVS channel for the test image in comparison with the mutual information for the reference image, was the top performer in compression, packet loss, and all, MOVIE showed the highest correlation with subjective quality scores in the case of rate adaptation, and visual signal-to-noise ratio (VSNR) [7], which estimates visual fidelity by computing contrast thresholds, showed the highest correlation for temporal dynamics. In the tablet study, VIF recorded the highest correlations in compression, MOVIE was the best in rate adaptation, packet loss, and all, and SNR showed the highest correlation in temporal dynamics. The hypothesis testing and statistical analysis confirmed the correlation results.

Those perceptual quality metrics are developed for general image/video quality assessment. On the other hand, there are objective quality metrics for the mobile scenario. Since they consider mobile-specific factors such as network distortions, they are expected to have better performance than general metrics.

4.3.2 Objective Metrics for Mobile Videos

Table 4.1 summarizes representative objective metrics for mobile videos. In general, objective metrics can be classified according to the availability of the original video besides the test video. Full-reference (FR) metric is used, when the original video is accessible, reduced-reference (RR) metrics is used when description or parameters of the original signal are available, and no-reference (NR) metrics is used when the original signal is not available. FR can be used in offline scenarios for designing and optimizing video processing algorithms as replacements of or conjunction with subjective tests. On the other hand, RR and NR metrics are useful for in-service quality monitoring to adapt transmission and coding strategies to bandwidth fluctuations and packet losses.

Table 4.1 Objective quality metrics for mobile videos

Ref.	Method	Parameters	Type of information
[81]	NR	QP, bit rate, number of lost packets, display duration	Bitstream-based
[80]	NR	Temporal complexity, frame type, bits per pixel, number of lost packets	Packet-based
[65]	NR	QP, motion vectors, bit rate, packet loss	Bitstream-based
[1]	NR	QP, motion vectors, bit rate, packet loss visibility, error propagation	Bitstream and packet (hybrid)
[5]	FR	Distorted frame ratio, frame loss rate	Pixel-based
[45]	FR	Packet loss error length, severity, and location	Pixel-based
[56]	FR	Spatial and temporal pooling, egomotion detection	Pixel-based
[18]	NR	Packet loss, rebuffering, bit rate	Bitstream-based
[82]	RR	Packet loss, frame rate, bit rate	Bitstream-based
[6]	RR	Temporal index	Pixel-based
[64]	NR	Frame rate	Bitstream-based

Objective quality metrics can also be categorized depending on the type of information, such as operation parameter-based, packet-based, bitstream-based, pixel-based, and hybrid approaches.

The most common approach for mobile quality assessment is to measure the impacts of coding and packet losses at the same time [1, 5, 15, 45, 80, 81].

In [81], an NR metric using information extracted from the bitstream header was developed. It measures the quality by subtracting the impact of packet loss from initial quality determined by the compression error. The coded frame quality, Q_n is estimated from the QP value, the spatial complexity and the temporal complexity. The transmission error, d_n, is calculated with direct distortion due to the lost packets, $d_{e,n}$, and distortion due to packet error propagation, $d_{p,n}$.

$$d_n = d_{e,n} + d_{p,n}$$

$$d_{e,n} = \left(\frac{num_A - num_R}{num_A}\right) \cdot Q_n \cdot \left(\frac{\sigma_{T,n}}{a}\right)^b \tag{4.1}$$

$$d_{p,n} = d_r \cdot \left(1 + \left(\frac{\sigma_{T,n}}{c}\right)^d\right) \tag{4.2}$$

where $\sigma_{T,n}$ is the temporal complexity of the nth frame. a, b, c, and d are model parameters. num_A is the total number of packets related to the nth frame and num_R is the number of packets valid for decoding the frame. d_r is the quality degradation of the reference (rth) frame, from which the distortion will propagate to the current frame. The final quality is given by subtracting the transmission error to the initial quality as follows.

$$Q_{F,n} = Q_n - d_n \tag{4.3}$$

A similar approach was used in [81] but, unlike the above bitstream-based model, the one in [81] is a packet-layer model exploiting information available only at the packet level. [80] investigated the relationship between the spatial quality and the average number of bits for coding a frame and display duration of a frame. The central part of the developed model, $Q_{v,GOF}$, is the quality of a group of frames (GOF), which is used as the basic unit of temporal pooling.

$$Q_{v,GOF} = \frac{\sum_{n \in GOF} C(n) \cdot T(n)}{\sum_{n \in GOF} T(n)} \tag{4.4}$$

where $C(n)$ is the contribution of the nth frame. The central part of the developed model, $\sigma'_T(n)$, is the normalized temporal complexity, $T(n)$ is display duration of the nth frame, and d_1 to d_3 are model parameters.

$$C(n) = Q_s(n)(d_1 + d_2\sigma'_T(n) + d_3\sigma'_T(n)\log(T(n))) \tag{4.5}$$

In [65], the "T-V Model," a parametric model for video quality estimation by considering bit rate, packet loss, and video content characteristic is presented. The model is expressed as follows:

$$Q_v = Q_{max} - Q_c - Q_t \tag{4.6}$$

$$Q_c = a_0 + a_1 MV - a_1 e^{-a_2 b + a_3 QP_1} \tag{4.7}$$

$$Q_t = (b_0 - Q_c)\frac{p}{b_1 + p} \tag{4.8}$$

where Q_v is the predicted quality, Q_{max} is the maximum achievable quality, Q_c is the quality degradation by coding, and Q_t is the quality degradation by transmission error. b is the bit rate, p is the percentage of packet loss, MV is the average of the standard deviation of the horizontal components of the motion vectors, QP_1 is the averaged QP value over each I-frame, and a_1 to a_3 and b_0 to b_1 are coefficients that need to be calculated for each codec and display size. The extended study [1] presented a modification to the transmission quality, Q_t, based on the evaluation of the visibility of each lost packet:

$$Q_t = a_4 d_e + d_{e.p}^{a_5} + a_6 \tag{4.9}$$

where d_e is the amount of noticeable distortion in the frame where the loss occurred, and $d_{e.p}$ is the amount of impaired pixels due to error propagation. a_4 to a_6 are coefficients.

In [5], two FR metrics based on PSNR were proposed, namely PSNR-based objective MOS (POMOS) and rates-based objective MOS (ROMOS). The former only considers the average PSNR of frames ($aPSNR$), whereas distorted frame rate (d), averaged PSNR of distorted frames ($dPSNR$), and frame loss rate (ℓ) are used for the latter.

$$\text{POMOS} = \beta_0 + \beta_1 aPSNR \tag{4.10}$$

$$\text{ROMOS} = \beta_0 + \beta_1 \frac{d}{dPSNR} + \beta_2 \ell \tag{4.11}$$

ROMOS showed a higher correlation with subjective rating than POMOS.

In [45], another FR method based on PSNR was proposed. In order to consider transmission errors, it includes network impairment factors such as packet loss, packet loss pattern, duration of a loss-affected segment, severity of loss, and loss location. The DMOS value is determined by summation of compression error (Q_c), and network impairment factors (Q_t):

$$Q_v = Q_c + Q_t \tag{4.12}$$

$$Q_c = \frac{Q_{c,max}}{1 + e^{s(PSNR-PSNR_T)}} \tag{4.13}$$

$$Q_t = CD \frac{1}{L} \sum_{i=1}^{N} W(D_i) MPDS_i \tag{4.14}$$

where $PSNR_T$ is the transition value of the PSNR curve over time. $Q_{c,max}$ is the maximum possible perceptual quality degradation due to coding artifacts and s is the roll-off factor of sigmoid function. N is the number of losses, CD is the length (in terms of frame) of video clip. $W(D_i)$ is the exponential decay function that simulates the effect of multiple losses. $MPDS_i$ is the sum of PSNR drops in the video segment affected by the packet loss in the i-th frame. It was shown to outperform PSNR, VQM, and SSIM.

In [56], it was recognized that severe and highly annoying distortions that occur locally in space or time heavily influence an observer's judgment of quality. The authors performed experiments using SSIM as an indicator of spatio-temporally local quality of a distorted video and observed the distributions of these scores. The SSIM scores are classified into two regions, lower quality region G_L and higher quality region G_H. These two regions are classified by using different thresholds that are determined using egomotion detection. As a result, a content adaptive spatial and temporal pooling strategy was proposed. The overall quality is expressed below:

$$Q_v = \frac{\sum_{f \in G_L} s_f + w \cdot \sum_{f \in G_H} s_f}{|G_L| + w \cdot |G_H|} \tag{4.15}$$

where $|G_L|$ and $|G_H|$ denote the cardinalities of G_L and G_H, respectively. The weight w is the ratio between the scores in G_L and G_H and s_f is the spatial quality for frame f.

There exist approaches that take into account the limited computation power and memory in mobile devices for estimating perceived quality by only using temporal information of videos. The network artifacts such as delay, freezing, jerkiness, blockiness, and blackout can be captured by measuring temporal information. In [6], an RR metric was developed where temporal information was determined as the differences of the corresponding pixel values in the two neighboring frames in the video. Video quality Q_v is expressed as

$$Q_v = \beta_0 + \beta_1 I_t \tag{4.16}$$

where I_t is the temporal information, β_0 and β_1 are model coefficients. Despite its simplicity, it showed a high correlation with MOS values.

The metric developed in [64] is another example exploiting the impact of temporal artifacts on video quality. It is given as a logistic function of the frame rate of the received video stream in order to take into account the saturation of perceived quality for received frame rates higher than a sufficient value, which can capture the impact of jerkiness and jitter artifacts. In other words,

$$Q_v = a_1 + \frac{a_2 - a_1}{1 + \exp(a_3 \cdot f + a_4)} \tag{4.17}$$

where f is the actual frame rate of the received video and a_1 to a_4 are model coefficients. It was shown that the model predicts subjective quality well when jerkiness is the dominant video impairment factor.

In [18] the combined effects of packet loss and buffering are taken into account. During the buffering time the image freezes, producing an annoying effect that affects the perceived quality. The proposed model estimates the video quality based on the MOS for the original video clip, the buffer size in the receiver, the re-buffering time during reproduction and the packet loss in the network, and was evaluated for MPEG-4 in QCIF display size with bit rates up to 256 kb/s. The model is given as

$$Q_v = 1 + (Q_c - 1)Q_t - Q_b \tag{4.18}$$

$$Q_c = c_0 - c_1 e^{-\lambda b} \tag{4.19}$$

$$Q_t = k \frac{p_u - p_m}{p_u - p_l} \tag{4.20}$$

$$Q_b = C_0 + C_1 InitP + C_2 BufP + C_3 BufF \tag{4.21}$$

where b is the bit rate, p_u and p_l are the upper and lower packet loss rate limits, respectively, p_m is the average packet loss rate of the current logging window, *InitP* is the initial buffer time, *BufP* is the re-buffering time, *BufF* is the number of re-buffering events per minute and $k, c_o, c_1, \lambda, C_0, C_1, C_2$, and C_3 are model coefficients. Video content characteristics are not taken into account in this model.

In [82], Q_t in the ITU-T G.1070 [22] model (see Sect. 4.3.4) was modified and extended in order to take into account burst packet loss.

$$Q_t = e^{-\frac{p}{B_{P_\ell} D_{P_\ell}}} \tag{4.22}$$

$$B_{P_\ell} = 1 + \alpha \frac{D_B}{N_{BP}} + \beta \frac{D_B D}{Loss} \tag{4.23}$$

where D_B is the density of burst, D is the burst duration, and $Loss$ is the total loss. N_{BP} is the number of burst periods. Coefficients α and β are dependent on codec, distortion concealment, and other factors related to content. It was shown that the model achieves better accuracy than the G.1070 [22] video model under burst loss conditions.

4.3.3 Objective Metrics for Video Scalability

Table 4.2 summarizes state-of-the art objective metrics for video scalability. The table shows the considered scalability dimensions, the used codec for model development, the way incorporating content-dependence of perceived quality, and the detailed formula for each metric. Detailed description follows below.

In [13], a quality metric accounting for quantization, frame rate, and motion speed was developed for mobile video broadcasting applications. It is based on the observation that the quality is dominated by PSNR if there is no motion in the source sequence but the frame rate reduction and the motion speed are influencing factors to the perceived quality. The metric was shown to have higher correlation with subjective ratings than PSNR.

The metric proposed in [52] considers temporal and quality scalability dimensions by formulating the perceived quality as a product of a PSNR-based spatial quality factor and a temporal correction factor using the frame rate. It was compared with the metrics proposed in [13, 64] that consider the temporal aspect in quality estimation, and show slightly higher correlation. It was also demonstrated that the model parameters can be estimated from content dependent spatial and temporal features such as frame difference, motion direction, and Gabor texture features.

Inspired by Feghali et al. [13], the metric proposed in [37] for the three-dimensional scalability is expressed as a weighted sum of three terms: PSNR, the motion activity-modulated effect of the frame rate, and the effect of the frame size. The third term, which is the major difference from the model in [13], accounts for the observation that the perceived quality increases with the increasing spatial resolution of stimuli.

The method in [37] was further modified in [71]. The developed metric is formulated as a weighted sum of normalized PSNR, the effect of the temporal scalability, and the effect of the spatial scalability. Normalization of PSNR was done to reflect the reduced influence of coding artifacts on the perceived quality for contents having high spatial complexity and the saturation of perceived quality for PSNR over 45 dB. In addition, the normalized PSNR was weighted by a spatial complexity measure in order to account for the fact that the quality degradation due to reduction of the spatial resolution is severe for contents having high spatial complexity.

In [54], three separate experiments were carried out for evaluating the influence of the spatial scalability to the perceived quality, the combinational effects of spatial resolution and QP, and all the combinational impact of spatial/temporal resolutions

Table 4.2 Objective quality metrics for video scalability

Ref.	Scalability[a]	Considered ranges[b]	Codec	Content-dependence	Formula[c]
[13]	T, R	S:130 × 192 T: 7.5–30 Hz	H.263+	Motion information	$\mathrm{PSNR} + \alpha_1 m^{\alpha_2}(f_{\max} - f)$
[52]	T, R	S:QCIF, CIF T: 7.5–30 Hz	SVC	Model parameters	$Q_{max}\left(1 - \frac{1}{1+e^{(\beta_1 \mathrm{PSNR} - \beta_2)}}\right)\left(\frac{1-e^{\beta_3 \frac{f}{f_{max}}}}{1-e^{\beta_3}}\right)$
[37]	S, T, R	S:QCIF, CIF T: 7.5–30 Hz	SVC	MPEG-7 motion activity	$\gamma_1 \mathrm{PSNR} + \gamma_2 M(f_{\max} - f)\frac{\gamma_3}{1+e^{-\gamma_4(h-h_0)}} + \gamma_5$
[71]	S,T, R	S:QCIF, CIF T: 7.5–30 Hz	SVC	MPEG-7 motion activity, edge histogram	$\delta_1\left(\frac{\mathrm{PSNR}-20}{25}\right)^{\delta_2 - \delta_2 S} + \delta_4 M(f_{\max} - f)$ $+\delta_5 e^{-0.5\left(\frac{h-h_{\max}}{\delta_6 - \delta_7 S}\right)^2} + \gamma_8$

(continued)

Table 4.2 (continued)

Ref.	Scalability[a]	Considered ranges[b]	Codec	Content-dependence	Formula[c]
[54]	S, T, R	S:QCIF, CIF, 4CIF T: 7.5–30 Hz	SVC	Quantization step size spatial/temporal information	$\frac{1-e^{-\upsilon_q(\frac{q}{q_{\min}})}}{1-e^{-\upsilon_s}}$ $\cdot\frac{1-e^{-\upsilon_s(q)(\frac{s}{s_{\max}})^{\nu_s}}}{1-e^{-\upsilon_s(q)}}\cdot\frac{1-e^{-\upsilon_t(\frac{t}{t_{\max}})^{\nu_t}}}{1-e^{-\upsilon_t}}$
[83]	T, R	S:130 × 192 T: 7.5–30 Hz	SVC	Squared difference of slope (SDS), Sobel filter information	$\left(\frac{1}{XY}\sum_{\text{allboundarypoints}} SDS\right)^{\sigma_1}$ $\cdot\left(\frac{1}{N_e}\sum_{\text{alledgepoints}}\lvert xp_1-xp_2\rvert\right)^{\sigma_2}$
					$\cdot\frac{f_{max}}{f}\left(\sqrt{\frac{1}{XY}\sum_{x=1}^{X}\sum_{y=1}^{Y}\lvert P_i(x,y)-P_{i-1}(x,y)\rvert}\right)^{\sigma_3}$

[a] S: spatial dimension; T: temporal dimension; R: SNR dimension

[b] QCIF: 176 × 144; CIF: 352 × 288; 4CIF: 704 × 576

[c] f: frame rate; $f_{\max}$: maximum frame rate; h: image height; $h_{\max}$: maximum image height; h_0: mean of the minimum and maximum image heights; m: average magnitude of the top 25 % largest motion vectors; M: MPEG-7 motion activity; S: MPEG-7 edge histogram; $\{\alpha_i\}$; $\{\beta_i\}$; $\{\gamma_i\}$; $\{\delta_i\}$; $\{\sigma_i\}$: model parameters; υ_q, υ_s, υ_t, ν_q, ν_s, and ν_t: content dependent parameters depending on QP, spatial resolution, and temporal resolution; N_e: number of edge points; xp_1, xp_2: edge points; X×Y frame; P_i: i th picture

and QP. Separate quality models reflecting the quality impact of spatial resolution, temporal resolution, and SNR were derived, which are multiplied for the final quality model.

A low-complexity NR algorithm, called quality impairment score, was proposed to assess video quality under different spatial, temporal, and SNR combinations in [83]. The metric is a weighted product of three factors, a blur metric, a blockiness metric, and a jerkiness metric, each of which measures the quality in the spatial, SNR, and temporal scalability dimensions, respectively.

It is noteworthy that the aforementioned two studies noted that temporal resolution has a huge impact on perceived quality, but if the temporal resolution is reasonably high then the other two scalabilities have mild influences to perceived quality.

4.3.4 Objective Metrics in Standardization

The study groups 9 (SG9) and 12 (SG12) in ITU-T have been at the forefront in defining and validating subjective and objective quality assessment methods [9]. SG12 is to provide adequate QoE and QoS for new multimedia services and applications such as IPTV. Especially, question 13 in SG12 has considered end-user expectations, quality management and assurance, QoS/QoE monitoring methodologies, etc. Question 14 focuses on the development of parametric models and tools for multimedia quality assessment. Table 4.3 summarizes major objective metrics in ITU-T standardization.

An early method, J.144 [23], contains a three-layered (object, texture, and noise) picture quality assessment model as seen by the human eyes. Generally, the human eyes cannot watch a whole frame at a glance, but can watch only a local spot area in a frame, which is around the gaze point, and recognize the texture and also quality of the area depending on the degrees and characteristics of noise mixed in this texture. Nine proponents of video quality metrics are introduced in J.144 [23].

The four models in BT.1683 [21] are based on the observation that the HVS is sensitive to degradation around the edges. They are FR models that include their own edge analysis methods used for computing video quality.

Recommendation ITU-T J.247 [24] also provides objective perceptual video quality measurement methods when a full reference signal is available. It takes into account distortion in the form of block distortion by calculating the ratio between horizontal and vertical edges, temporal variance of partial spatial distortion, and freeze length.

Recommendation ITU-T J.341 [25] describes an FR model for HD resolutions. This recommendation focuses on audiovisual quality by considering time alignment and spatial frame alignment between audio and video signals. Spatial quality features are computed by using a local similarity and a local difference measure. A jerkiness feature takes into account motion intensity as temporal quality. The overall quality is represented as weighted combination of these quality features.

Table 4.3 Objective metrics in standardization

Rec.	Method	Name	Parameters
J.144 [23]	FR	Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference	PSNR, content complexity
BT.1683 [21]	FR	Objective perceptual video quality measurement techniques for standard definition digital broadcast television in the presence of a full reference	PSNR, degradation around edges, motion distortion
J.247 [24]	FR	Objective perceptual multimedia video quality measurement in the presence of a full reference	PSNR, horizontal and vertical edges, motion distortion
J.341 [25]	FR	Objective perceptual multimedia video quality measurement of HDTV for digital cable television in the presence of a full reference	Local motion intensity and local similarity
J.342 [26]	RR	Objective multimedia video quality measurement of HDTV for digital cable television in the presence of a reduced reference	Edge PSNR, blocking metric, freezed frames
G.1070 [22]	RR	Opinion model for video-telephony applications	Bit rate, frame rate, packet-loss rate
P.1201 [27]	NR	Parametric non-intrusive assessment of audiovisual media streaming quality	Loss magnitude, burstiness, bitrate, content complexity, freezing ratio
P.1202 [28]	NR	Parametric non-intrusive bitstream assessment of video media streaming quality	Frame rate, rebuffering duration, packet error concealment mode

Recommendation ITU-T J.342 [26] targets similar applications with J.341 [25] (i.e., HDTV) but it provides an RR model. This model measures the visibility of edges for extracting the blockiness feature and computes a jerkiness feature by averaging the product of display time and motion intensity for temporal quality. The features are transformed into a perceptual scale by using a parameterized function.

The latest methods such as G.1070 [22], P.1201 [27], and P.1202 [28] are parametric models that analyze packet headers and bitstream information rather than decoded pixel information.

The G.1070 model takes into account the packet loss, assuming a random loss distribution. Video quality Q_v is expressed as:

$$Q_v = 1 + Q_c \exp\left(-\frac{P_\ell}{D_{P_\ell}}\right) \tag{4.24}$$

$$Q_c = Q_{f_{opt}} \exp\left\{\frac{(\ln(f) - \ln(f_{opt}))^2}{2D_{Fr}^2}\right\} \tag{4.25}$$

where Q_c represents the video quality affected by the coding distortion and $Q_{f_{opt}}$ is the maximum video quality with an optimal frame rate, f_{opt}. The packet loss robustness factor D_{P_ℓ} expresses the degree of video quality robustness against packet loss, P_ℓ represents the packet loss rate, p is the percentage of packet loss, and f is the frame rate. D_{Fr} represents the degree of video quality robustness due to frame rate.

Models in P.1201 [27] provide audio, video, and audiovisual quality estimates, and they use only packet header information. In contrast, the models in ITU-T P.1202 [28] provide only video quality estimates using further payload information, thus they can be more accurate in their quality predictions but more complex than those in P.1201. The two standards include two different versions, i.e., lower resolution models (ITU-T P.1201.1 and ITU-T P.1202.1), and higher resolution models (ITU-T P.1201.2 or ITU-T P.1202.2). The model in P.1201.2 exploits bit rate, packet loss, concealment type, etc. Video quality estimate Q_v is given as:

$$Q_v = 100 - Q_c - Q_t \tag{4.26}$$

Here, Q_c is the estimated video quality due to video compression artifacts, which is given by

$$Q_c = a1 \cdot e^{a2 \cdot BitPerPixel} + a3 \cdot ContentComplexity + a4 \tag{4.27}$$

where a1 to a4 are model coefficients. Q_t is the estimated video quality due to transmission artifacts. In case of freezing due to skipping of erroneous frames, it is given by:

$$Q_t = b1 \cdot \log(b2 \cdot B_{Fr} + 1) \tag{4.28}$$

where B_{Fr} is the bit per pixel multiplied by the freezing ratio. $b1$ and $b2$ are model coefficients. When the decoder tries to repair erroneous frames (e.g., motion copy or frame copy), Q_t is given by:

$$Q_t = c1 \cdot \log(c2 \cdot \frac{LossMagnitude}{Q_{c_n}} + 1) \tag{4.29}$$

where $c1$ to $c2$ are model coefficients and Q_t is represented as transmission artifacts Q_{c_n} is weighted value of Q_c. The model in P.1202.2 uses an additional parameter upon the parameters used in P.1201.2 such as rebuffering. The video quality Q_v is expressed as a linear relationship:

$$\begin{aligned} Q_v = \alpha_1 &\times compressionartifactvalue \\ &+ \alpha_2 \times slicingartifactvalue \\ &+ \alpha_3 \times freezingartifactvalue \\ &+ \alpha_4 \times rebufferingartifactvalue + \alpha_5 \end{aligned} \tag{4.30}$$

where α_1 to α_5 are model coefficients.

ITU-T study groups continuously search quality models for various applications such as TCP-based multimedia streaming, UDP-based streaming, adaptive streaming, etc. ITU-T also carries out studies supporting advanced capabilities such as ultra high-definition (UHD) and 3D TV.

4.4 Databases for Mobile Quality Assessment

For researches about mobile video quality assessment, it is useful to refer to existing relevant databases that are publicly available. Most databases for video quality assessment offer a wide variety of test video sequences and the resulting subjective test data. Based on the subjective quality data, additional researches reflecting subjective quality, e.g., development and evaluation of objective quality models, can be performed in practice.

Table 4.4 summarizes representative publicly available video databases that can be used for researches about mobile video quality assessment. They cover a wide variety of conditions, e.g., coding artifacts, spatial resolutions, temporal resolutions, transmission errors, and so on. Generally, these databases can be classified into two cases, one that considers network artifacts (packet loss) (e.g., [2,3,11,45,63,69,85]) and another one that considers scalability (e.g., [11, 42, 45, 51, 53–55, 60–62, 85]). As can be seen in the table, the recent databases tend to deal with relatively higher spatial and temporal resolutions.

A thorough survey of recent public databases for image and video quality assessment can be found in [77], where the databases including those in Table 4.4 are analyzed in various aspects.

Table 4.4 Databases about mobile video quality assessment

Ref.	Year	Codec[a]	Scalability[b,c]	Transmission error	#Video	Method[d]	#Subject
[51]	2008	No codec	S: CIF/QCIF T: 6/7.5/10/15/30 fps R: uncompressed	No	75	ACR	30
[53]	2009	SVC	S: CIF T: 3.75/7.5/15/30 fps R: QP 28/36/40/44	No	68	ACR-HR	31
[45]	2009	H.264/AVC	S: QVGA T: 12–15 fps R: QP 26/38	Yes	34	DSCQS	60
[3]	2009	H.264/AVC	S: 576i T: 25 fps R: 2 bit rates	Yes	84	ACR	16
[55]	2010	SVC	S: CIF T: 3.75/7.5/15/30 fps R: QP 28/36/44	No	150	ACR	33
[60]	2010	H.264/AVC SVC	S: VGA/QVGA T: 15/30 fps R: 4 bit rates	No	48	SAMVIQ	15
[61]	2010	H.264/AVC	S: SD/720p/1080p T: 15/30 fps R: 4 bit rates	No	87	ACR	26
[11]	2010	H.264/AVC	S: CIF/4CIF T: 25–30 fps R: 1 bit rate	Yes	144	ACR-HR	44
[69]	2010	MPEG-2 H.264/AVC	S: 768×432 T: 25 fps R: 1 bit rate	Yes	144	SS	38

(continued)

Table 4.4 (continued)

Ref.	Year	Codec[a]	Scalability[b,c]	Transmission error	#Video	Method[d]	#Subject
[62]	2011	H.264/AVC SVC	S: VGA/QVGA T: 15/30 fps R: 2 bit rates	No	390	ACR-HR	28
[42]	2011	SVC WSVC	S: 180p/360p/720p T: 6.25/12.5/25/50 fps R: 4–6 bit rates	No	58	PC	16
[63]	2011	H.264/AVC SVC	S: VGA T:15/30 fps R: 2 bit rates	Yes	131	ACR-HR	29
[85]	2011	MPEG-2 H.264/AVC Dirac	S: 1080p T: 25 fps R: 4 bit rates	Yes	128	ACR	42
[2]	2011	H.264/AVC	S: 1080p T: 25 fps R: 1 bit rate	Yes	120	ACR	22
[54]	2012	SVC	S: 4CIF/CIF/QCIF T: 7.5/15/30 fps R: QP 22/28/36/44	No	224	SS	60

[a] WSVC: Wavelet-based scalable video codec

[b] S: spatial dimension; T: temporal dimension; R: SNR dimension

[c] QCIF: 176 × 144; CIF: 352 × 288; 4CIF: 704 × 576; VGA: 640 × 480; QVGA: 320 × 240; SD: 720 × 576; 576i: 768 × 576 interlacing; 180p: 320 × 180; 360p: 640 × 360; 720p: 1280 × 720; 1080p: 1920 × 1080

[d] SS: Single Stimulus; PC: Paired Comparison

Although there exist many databases for quality assessment of mobile videos, there is a need to develop new databases considering more diverse and realistic conditions for future researches. As mentioned above, currently available databases tend to deal with network artifacts and scalability separately. However, databases considering both network errors and three dimensional scalability conditions will be required. In addition, databases containing subjective ratings collected in real environments instead of controlled laboratory environments are needed for realistic mobile video quality assessment researches. Furthermore, although the existing databases usually consider simulated network environments for inducing transmission errors, databases obtained from real network environments will be valuable.

4.5 Conclusion and Future Challenges

In this chapter, recent advances in the research of subjective and objective quality assessment of mobile videos were reviewed. To understand and model the perceptual mechanism of QoE, various perceptual factors specific in mobile videos were considered in the existing studies, including compression artifacts, packet loss artifacts, delay and freeze artifacts, temporal quality variation, video scalability, etc.

Many subjective studies have been conducted using test methodologies developed for controlled laboratory environments. Further studies are needed to investigate subjective test methodologies and environments appropriate for mobile videos by considering characteristics of mobile devices and viewing behavior.

As shown in this chapter, many objective quality metrics for mobile videos have been developed, but it is also important to evaluate their relative performance via thorough benchmarking studies. One of the latest studies in [32] presents a review of parametric models published by ten different groups of authors [1, 18, 22, 31, 36, 40, 52, 64–66, 82], some of which were explained in Sect. 4.3. The performance of each model is evaluated and contrasted to the other models, using a common video clips set, in different coding and transmission scenarios. It was shown that the model in [31] performs better than the others for the encoding impairments estimation and the models in the Recommendation ITU-T G.1070 [22] performs better for transmission impairments estimation. This study shows the performance comparison only between the parametric models. Therefore, extensive benchmarking encompassing general objective models, standardized models, and mobile-specific models is still desirable in the future.

Due to fast technological development, new types of media are being introduced to consumers such as 3D videos, high dynamic range (HDR) videos, and UHD videos. There will be a high demand to consume these types of media in the mobile environment, for which perceptual quality assessment will also play an important role.

As the 3D video technology becomes popular in the field of industry and research, the mobile 3D technology has also received increasing attention. Generally,3D videos require increased data rates than 2D videos, so understanding the perceptual quality of 3D videos will be more important for efficient mobile video transmission. There exist many researches considering quality assessment of 3D video in general environments (e.g., [17, 29, 41, 49]), but few researches considering mobile 3D videos exist. The recent study in [34] investigated general descriptive characteristics of experienced quality of 3D videos on mobile devices. Experiments including subjective quality assessment were performed by varying contents, levels of depth, compression and transmission parameters, and audio and display factors for 3D. It was concluded that QoE of mobile 3D videos is constructed from four main components: (1) visual quality in terms of depth, spatial, and motion, (2) viewing experience, (3) content, and (4) quality of other modalities and their interactions. For the scalable multiview 3D video coding, objective quality metrics considering each layer were derived based on the subjective quality assessment in [67]. Further researches beyond these will be required.

Although HDR imaging receives more and more attention, the researches dealing with quality of mobile HDR videos are extremely rare. There exist only a few researches about tone mapping operators (TMOs) for visualizing HDR videos on conventional displays and, moreover, researches about HDR videos considering mobile devices are even rarer (e.g., [4, 47]). In the future, it is expected that researches in these direction will be performed actively, where perceptual quality assessment will also play an important role.

As the display technology grows, high resolution displays can be adopted in mobile devices. UHD is one of the future trends and challenges for the industry and researches. It requires more pixels and consequently higher data rates, but it is expected to give more vibrant experience to users. In the context of quality assessment, researches considering UHD videos and their scalabilities can be found recently (e.g., [12, 19]). However, studies considering perceptual quality of UHD videos on mobile devices are extremely rare. However, mobile devices equipped with UHD displays are expected to become available very soon and popular in the near future, so the issue of perceptual quality assessment of UHD videos in such devices, which may be quite different from that of SD or HD videos, will become important accordingly.

Acknowledgements This research was supported by the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the IT Consilience Creative Program (NIPA-2014-H0201-14-1002) supervised by the NIPA (National IT Industry Promotion Agency), and by the Basic Science Research Program through the National Research Foundation of Korea funded by the MSIP (2013R1A1A1007822).

References

1. Argyropoulos, S., Raake, A., Garcia, M.N., List, P.: No-reference bit stream model for video quality assessment of H.264/AVC video based on packet loss visibility. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1169–1172 (2011)

2. Boujut, H., Benois-Pineau, J., Hadar, O., Ahmed, T., Bonnet, P.: Weighted-MSE based on saliency map for assessing video quality of H.264 video streams. In: Proceedings of the IS&T/SPIE Electronic Imaging, pp. 78,670X–78,670X. International Society for Optics and Photonics (2011)
3. Boulos, F., Parrein, B., Le Callet, P., Hands David, S.: Perceptual effects of packet loss on H.264/AVC encoded videos. In: Proceedings of the Fourth International Workshop on Video Processing and Quality Metrics for Consumer Electronics, pp. 1–6 (2009)
4. Castro, T., Chapiro, A., Cicconet, M., Velho, L.: Towards mobile HDR video. In: Proceedings of the Eurographics-Areas Papers, pp. 75–76. The Eurographics Association (2011)
5. Chan, A., Zeng, K., Mohapatra, P., Lee, S.J., Banerjee, S.: Metrics for evaluating video streaming quality in lossy IEEE 802.11 wireless networks. In: Proceedings of the IEEE INFOCOM, pp. 1–9. IEEE (2010)
6. Chan, A.J., Pande, A., Baik, E., Mohapatra, P.: Temporal quality assessment for mobile videos. In: Proceedings of the 18th Annual International Conference on Mobile Computing and Networking, pp. 221–232. ACM (2012)
7. Chandler, D.M., Hemami, S.S.: VSNR: A wavelet-based visual signal-to-noise ratio for natural images. IEEE Transactions on Image Processing **16**(9), 2284–2298 (2007)
8. Chen, J.Y., Thropp, J.E.: Review of low frame rate effects on human performance. IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans **37**(6), 1063–1076 (2007)
9. Coverdale, P., Mollerand, S., Raake, A., Takahashi, A.: Multimedia quality assessment standards in ITU-T SG12. IEEE Signal Processing Magazine **28**(6), 91–97 (2011)
10. Cranley, N., Perry, P., Murphy, L.: User perception of adapting video quality. International Journal of Human-Computer Studies **64**(8), 637–647 (2006)
11. De Simone, F., Tagliasacchi, M., Naccari, M., Tubaro, S., Ebrahimi, T.: A H.264/AVC video database for the evaluation of quality metrics. In: Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), pp. 2430–2433. IEEE (2010)
12. Deshpande, S.: Subjective and objective visual quality evaluation of 4K video using AVC and HEVC compression. In: Proceedings of the SID Symposium Digest of Technical Papers, vol. 43, pp. 481–484. Wiley Online Library (2012)
13. Feghali, R., Speranza, F., Wang, D., Vincent, A.: Video quality metric for bit rate control via joint adjustment of quantization and frame rate. IEEE Transactions on Broadcasting **53**(1), 441–446 (2007)
14. Fiedler, M., Hossfeld, T., Tran-Gia, P.: A generic quantitative relationship between quality of experience and quality of service. IEEE Network **24**(2), 36–41 (2010)
15. Garcia, M.N., Raake, A., List, P.: Towards content-related features for parametric video quality prediction of IPTV services. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 757–760 (2008)
16. Garcia, R., Kalva, H.: Subjective evaluation of HEVC in mobile devices. In: Proceedings of the IS&T/SPIE Electronic Imaging, pp. 86,670L–86,670L. International Society for Optics and Photonics (2013)
17. Goldmann, L., De Simone, F., Ebrahimi, T.: A comprehensive database and subjective evaluation methodology for quality of experience in stereoscopic video. In: Proceedings of the IS&T/SPIE Electronic Imaging, pp. 75,260S–75,260S. International Society for Optics and Photonics (2010)
18. Gustafsson, J., Heikkila, G., Pettersson, M.: Measuring multimedia quality in mobile networks with an objective parametric model. In: Proceedings of the 15th IEEE International Conference on Image Processing (ICIP), pp. 405–408. IEEE (2008)
19. Hanhart, P., Rerabek, M., De Simone, F., Ebrahimi, T.: Subjective quality evaluation of the upcoming HEVC video compression standard. In: Proceedings of the SPIE Optical Engineering+ Applications, pp. 84,990V–84,990V. International Society for Optics and Photonics (2012)

20. Huynh-Thu, Q., Ghanbari, M.: The accuracy of PSNR in predicting video quality for different video scenes and frame rates. Telecommunication Systems **49**(1), 35–48 (2012)
21. ITU-T Recommendation BT.1683: Objective perceptual video quality measurement techniques for standard definition digital broadcast television in the presence of a full reference (2004)
22. ITU-T Recommendation G.1070: Opinion model for video-telephony applications (2004)
23. ITU-T Recommendation J.144: Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference (2004)
24. ITU-T Recommendation J.247: Objective perceptual multimedia video quality measurement in the presence of a full reference (2008)
25. ITU-T Recommendation J.249: Objective perceptual multimedia video quality measurement of HDTV for digital cable television in the presence of a full reference (2011)
26. ITU-T Recommendation J.342,: Objective multimedia video quality measurement of HDTV for digital cable television in the presence of a reduced reference signal (2011)
27. ITU-T Recommendation P.1201: Parametric non-intrusive assessment of audiovisual media streaming quality (2012)
28. ITU-T Recommendation P.1202: Parametric non-intrusive bitstream assessment of video streaming quality (2012)
29. Jin, L., Boev, A., Gotchev, A., Egiazarian, K.: 3D-DCT based perceptual quality assessment of stereo video. In: Proceedings of the 18th IEEE International Conference on Image Processing (ICIP), pp. 2521–2524. IEEE (2011)
30. Jose Joskowicz, R.S.: A model for video quality assessment considering packet loss for broadcast digital television coded in H.264. International Journal of Digital Multimedia Broadcasting **2014**(242531), 1–11 (2014)
31. Joskowicz, J., Ardao, J.: Combining the effects of frame rate, bit rate, display size and video content in a parametric video quality model. In: Proceedings of the 6th Latin America Networking Conference, pp. 4–11. ACM (2011)
32. Joskowicz, J., Sotelo, R., Lopez Ardao, J.: Towards a general parametric model for perceptual video quality estimation. IEEE Transactions on Broadcasting **59**(4), 569–579 (2013)
33. Jumisko-Pyykkö, S., Hannuksela, M.M.: Does context matter in quality evaluation of mobile television? In: Proceedings of the 10th International Conference on Human Computer Interaction with Mobile Devices and Services, pp. 63–72. ACM (2008)
34. Jumisko-Pyykkö, S., Strohmeier, D., Utriainen, T., Kunze, K.: Descriptive quality of experience for mobile 3D video. In: Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries, pp. 266–275. ACM (2010)
35. Khan, A., Sun, L., Fajardo, J., Taboada, I., Liberal, F., Ifeachor, E.: Impact of end devices on subjective video quality assessment for QCIF video sequences. In: Proceedings of the Third International Workshop on Quality of Multimedia Experience (QoMEX), pp. 177–182. IEEE (2011)
36. Khan, A., Sun, L., Ifeachor, E.: Content-based video quality prediction for MPEG4 video streaming over wireless networks. Journal of Multimedia **4**(4) (2009)
37. Kim, C.S., Jin, S.H., Seo, D.J., Ro, Y.M.: Measuring video quality on full scalability of H.264/AVC scalable video coding. IEICE Transactions on Communications **91**(5), 1269–1278 (2008)
38. Kim, S.J., Chae, C.B., Lee, J.S.: Quality perception of coding artifacts and packet loss in networked video communications. In: Proceedings of the IEEE Globecom Workshops (GC Wkshps), pp. 1357–1361. IEEE (2012)
39. Korhonen, J., Reiter, U., Ukhanova, A.: Frame rate versus spatial quality: Which video characteristics do matter? In: Proceedings of the Visual Communications and Image Processing (VCIP), pp. 1–6. IEEE (2013)
40. Koumaras, H., Kourtis, A., Martakos, D., Lauterjung, J.: Quantified PQoS assessment based on fast estimation of the spatial and temporal activity level. Multimedia Tools and Applications **34**(3), 355–374 (2007)

41. Kulyk, V., Tavakoli, S., Folkesson, M., Brunnstrom, K., Wang, K., Garcia, N.: 3D video quality assessment with multi-scale subjective method. In: Proceedings of the Fifth International Workshop on Quality of Multimedia Experience (QoMEX), pp. 106–111. IEEE (2013)
42. Lee, J.S., De Simone, F., Ebrahimi, T.: Subjective quality evaluation via paired comparison: application to scalable video coding. IEEE Transactions on Multimedia **13**(5), 882–893 (2011)
43. Lee, J.S., De Simone, F., Ebrahimi, T., Ramzan, N., Izquierdo, E.: Quality assessment of multidimensional video scalability. IEEE Communications Magazine **50**(4), 38–46 (2012)
44. Liu, T., Wang, Y., Boyce, J.M., Wu, Z., Yang, H.: Subjective quality evaluation of decoded video in the presence of packet losses. In: Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), vol. 1, pp. I–1125–I1128. IEEE (2007)
45. Liu, T., Wang, Y., Boyce, J.M., Yang, H., Wu, Z.: A novel video quality metric for low bit-rate video considering both coding and packet-loss artifacts. IEEE Journal of Selected Topics in Signal Processing **3**(2), 280–293 (2009)
46. Lopez, J., Slanina, M., Arnaiz, L., Menendez, J.: Subjective quality assessment in scalable video for measuring impact over device adaptation. In: Proceedings of the IEEE EUROCON, pp. 162–169. IEEE (2013)
47. Magalhaes, L., Bessa, M., Urbano, C., Melo, M., Peres, E., Chalmers, A.: A survey on HDR visualization on mobile devices. In: Proceedings of the SPIE Photonics Europe, pp. 843,607–843,607. International Society for Optics and Photonics (2012)
48. Minhas, T.N., Lagunas, O.G., Arlos, P., Fiedler, M.: Mobile video sensitivity to packet loss and packet delay variation in terms of QoE. In: Proceedings of the 19th International Packet Video Workshop (PV), pp. 83–88. IEEE (2012)
49. Mittal, A., Moorthy, A.K., Ghosh, J., Bovik, A.C.: Algorithmic assessment of 3D quality of experience for images and videos. In: Proceedings of the IEEE Digital Signal Processing Workshop and IEEE Signal Processing Education Workshop (DSP/SPE), pp. 338–343. IEEE (2011)
50. Moorthy, A.K., Choi, L.K., Bovik, A.C., Veciana, G.D.: Video quality assessment on mobile devices: Subjective, behavioral and objective studies. IEEE Journal of Selected Topics in Signal Processing **6**(6), 652–671 (2012)
51. Ou, Y.F., Liu, T., Zhao, Z., Ma, Z., Wang, Y.: Modeling the impact of frame rate on perceptual quality of video. In: Proceedings of the 15th IEEE International Conference on Image Processing, 2008, pp. 689–692 (2008)
52. Ou, Y.F., Ma, Z., Liu, T., Wang, Y.: Perceptual quality assessment of video considering both frame rate and quantization artifacts. IEEE Transactions on Circuits and Systems for Video Technology **21**(3), 286–298 (2011)
53. Ou, Y.F., Ma, Z., Wang, Y.: A novel quality metric for compressed video considering both frame rate and quantization artifacts. In: Proceedings of the International Workshop Video Processing and Quality Metrics for Consumer (VPQM), pp. 1–5 (2009)
54. Ou, Y.F., Xue, Y., Wang, Y.: Q-STAR: A perceptual video quality model for mobile platforms considering impact of spatial, temporal, and amplitude resolutions. Tech. rep., Polytechnic Institute of NYU (2012)
55. Ou, Y.F., Zhou, Y., Wang, Y.: Perceptual quality of video with frame rate variation: A subjective study. In: Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), pp. 2446–2449 (2010)
56. Park, J., Seshadrinathan, K., Lee, S., Bovik, A.C.: Video quality pooling adaptive to perceptual distortion severity. IEEE Transactions on Image Processing **22**(2), 610–620 (2013)
57. Péchard, S., Pépion, R., Le Callet, P., et al.: Suitable methodology in subjective video quality assessment: a resolution dependent paradigm. In: Proceedings of the Third International Workshop on Image Media Quality and its Applications (IMQA) (2008)
58. Pessemier, T.D., Moor, K.D., Joseph, W., Marez, L.D., Martens, L.: Quantifying subjective quality evaluations for mobile video watching in a semi-living lab context. IEEE Transactions on Broadcasting **58**(4), 580–589 (2012)

59. Pessemier, T.D., Moor, K.D., Joseph, W., Marez, L.D., Martens, L.: Quantifying the influence of rebuffering interruptions on the user's quality of experience during mobile video watching. IEEE Transactions on Broadcasting **59**(1), 47–61 (2013)
60. Pitrey, Y., Barkowsky, M., Le Callet, P., Pepion, R.: Subjective quality assessment of MPEG-4 scalable video coding in a mobile scenario. In: Proceedings of the 2nd European Workshop on Visual Information Processing (EUVIP), pp. 86–91. IEEE (2010)
61. Pitrey, Y., Barkowsky, M., Le Callet, P., Pepion, R., et al.: Subjective quality evaluation of H.264 high-definition video coding versus spatial up-scaling and interlacing. QoE for Multimedia Content Sharing (2010)
62. Pitrey, Y., Engelke, U., Barkowsky, M., Pépion, R., Le Callet, P.: Subjective quality of SVC-coded videos with different error-patterns concealed using spatial scalability. In: Proceedings of the 3rd European Workshop on Visual Information Processing (EUVIP), pp. 180–185. IEEE (2011)
63. Pitrey, Y., Engelke, U., Barkowsky, M., Pépion, R., Le Callet, P., et al.: Aligning subjective tests using a low cost common set. QoE for Multimedia Content Sharing (2011)
64. Quan, H.T., Mohammed, G.: Temporal aspect of perceived quality of mobile video broadcasting. IEEE Transactions on Broadcasting **54**(3), 641–651 (2008)
65. Raake, A., Garcia, M.N., Moller, S., Berger, J., Kling, F., List, P., Johann, J., Heidemann, C.: T-V-model: Parameter-based prediction of IPTV quality. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1149–1152 (2008)
66. Ries, M., Crespi, C., Nemethova, O., Rupp, M.: Content based video quality estimation for H.264/AVC video streaming. In: Proceedings of the IEEE Wireless Communications and Networking Conference, pp. 2668–2673. IEEE (2007)
67. Roodaki, H., Hashemi, M.R., Shirmohammadi, S.: A new methodology to derive objective quality assessment metrics for scalable multiview 3D video coding. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP) **8**(3s), 44:1–44:25 (2012)
68. Seshadrinathan, K., Bovik, A.C.: Motion tuned spatio-temporal quality assessment of natural videos. IEEE Transactions on Image Processing **19**(2), 335–350 (2010)
69. Seshadrinathan, K., Soundararajan, R., Bovik, A.C., Cormack, L.K.: Study of subjective and objective quality assessment of video. IEEE transactions on Image Processing **19**(6), 1427–1441 (2010)
70. Sheikh, H.R., Bovik, A.C.: Image information and visual quality. IEEE Transactions on Image Processing **15**(2), 430–444 (2006)
71. Sohn, H., Yoo, H., De Neve, W., Kim, C.S., Ro, Y.M.: Full-reference video quality metric for fully scalable and mobile SVC content. IEEE Transactions on Broadcasting **56**(3), 269–280 (2010)
72. Stockhammer, T.: Dynamic adaptive streaming over HTTP: standards and design principles. In: Proceedings of the Second Annual ACM Conference on Multimedia Systems, pp. 133–144. ACM (2011)
73. Tominaga, T., Hayashi, T., Okamoto, J., Takahashi, A.: Performance comparisons of subjective quality assessment methods for mobile video. In: Proceedings of the Second International Workshop on Quality of Multimedia Experience (QoMEX), pp. 82–87. IEEE (2010)
74. Wang, D., Speranza, F., Vincent, A., Martin, T., Blanchfield, P.: Toward optimal rate control: a study of the impact of spatial resolution, frame rate, and quantization on subjective video quality and bit rate. In: Proceedings of the Visual Communications and Image Processing 2003, pp. 198–209. International Society for Optics and Photonics (2003)
75. Wang, Y., Schaar, M., Chang, S.F., Loui, A.C.: Classification-based multidimensional adaptation prediction for scalable video coding using subjective quality evaluation. IEEE Transactions on Circuits and Systems for Video Technology **15**(10), 1270–1279 (2005)
76. Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multiscale structural similarity for image quality assessment. In: Proceedings of the Conference Record of the Thirty-Seventh Asilomar Conference on Signals, Systems and Computers, 2004., vol. 2, pp. 1398–1402. IEEE (2003)

77. Winkler, S.: Analysis of public image and video databases for quality assessment. IEEE Journal of Selected Topics in Signal Processing **6**(6), 616–625 (2012)
78. Winkler, S., Mohandas, P.: The evolution of video quality measurement: From PSNR to hybrid metrics. IEEE Transactions on Broadcasting **54**(3), 660–668 (2008)
79. Xue, Y., Ou, Y.F., Ma, Z., Wang, Y.: Perceptual video quality assessment on a mobile platform considering both spatial resolution and quantization artifacts. In: Proceedings of the 18th International Packet Video Workshop (PV), pp. 201–208. IEEE (2010)
80. Yang, F., Song, J., Wan, S., Wu, H.R.: Content-adaptive packet-layer model for quality assessment of networked video services. IEEE Journal of Selected Topics in Signal Processing **6**(6), 672–683 (2012)
81. Yang, F., Wan, S., Xie, Q., Wu, H.: No-reference quality assessment for networked video via primary analysis of bit stream. IEEE Transactions on Circuits and Systems for Video Technology **20**(11), 1544–1554 (2010)
82. You, F., Zhang, W., Xiao, J.: Packet loss pattern and parametric video quality model for IPTV. In: Proceedings of the Eighth IEEE/ACIS International Conference on Computer and Information Science, pp. 824–828. IEEE (2009)
83. Zhai, G., Cai, J., Lin, W., Yang, X., Zhang, W.: Three-dimensional scalable video adaptation via user-end perceptual quality assessment. IEEE Transactions on Broadcasting **54**(3), 719–727 (2008)
84. Zhai, G., Cai, J., Lin, W., Yang, X., Zhang, W., Etoh, M.: Cross-dimensional perceptual quality assessment for low bit-rate videos. IEEE Transactions on Multimedia **10**(7), 1316–1324 (2008)
85. Zhang, F., Li, S., Ma, L., Wong, Y.C., Ngan, K.N.: IVP subjective quality video database. http://ivp.ee.cuhk.edu.hk/research/database/subjective/ (2011)

Chapter 5
High Dynamic Range Visual Quality of Experience Measurement: Challenges and Perspectives

Manish Narwaria, Matthieu Perreira Da Silva, and Patrick Le Callet

5.1 Introduction

Humans perceive the outside visual world through the interaction between light energy (usually measured in candela per square meter cd/m^2) and the eyes. Light energy first passes through the cornea, a transparent membrane. Then it enters the pupil, an aperture that is modified by the iris, a muscular diaphragm. Subsequently, light is refracted by the lend and hits the photoreceptors in the retina. There are two types of photoreceptors: cones and rods. The cones are located mostly in the fovea. They are more sensitive at luminance levels between 10^{-2} and 10^{8} cd/m^2 (referred to as the photopic or daylight vision) [7]. Further, color vision is due to three types of cones: short, middle, and long wavelength cones. The rods, on the other hand, are sensitive at luminance levels between 10^{-6} and 10 cd/m^2 (scotopic or night vision). The rods are more sensitive than cones but do not provide color vision. There is only one type of rod photoreceptors and are located around the fovea. Since there are no rods in the fovea, high frequency patterns cannot be distinguished at low lighting conditions [7].

Pertaining to the luminance levels found in the real world, direct sunlight at noon can be of the order in excess of 10^{7} cd/m^2 while a starlit night in the range of 10^{-1} cd/m^2. This corresponds to more than eight orders of magnitude. It is therefore evident that there is a large range of luminance present in different real-world scenes. The human eye also has the remarkable capability to perceive large dynamic range (about 13 orders of magnitude) especially with sufficient adaptation time [13]. An intuitive example of adaptation is when we arrive in a low lit room on a sunny day. We cannot immediately perceive the visual data in the room and

M. Narwaria • M.P. Da Silva (✉) • P. Le Callet
Lunam University, IRCCyN CNRS UMR 6597, Polytech Nantes, Rue Christian Pauc,
La Chantrerie B.P. 50609 44306, Nantes Cedex 3, France
e-mail: matthieu.perreiradasilva@univ-nantes.fr; dalian94@gmail.com

C. Deng et al. (eds.), *Visual Signal Quality Assessment*,
DOI 10.1007/978-3-319-10368-6_5

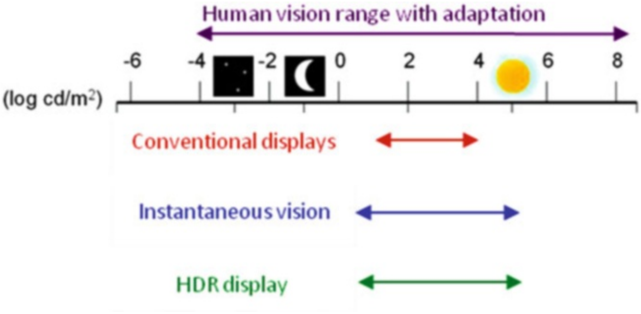


Fig. 5.1 Orders of magnitude of the dynamic range of the eye and displays

it takes a few minutes before one becomes accustomed (adapted) to new luminance levels. However without adaptation, the instantaneous human vision range is smaller and the eyes are capable of dynamically adjusting so that a person can see about five orders of magnitude throughout the entire range. An example comparing the approximate instantaneous orders of magnitude of the human eye, conventional display and the HDR display is shown in Fig. 5.1. Observe how the traditional LDR display covers only a small range (upto three orders of magnitude). On the other hand, HDR displays can better match the instantaneous range of the eye.

As pointed out, the conventional display devices cover only upto three orders of magnitude. Consequently, the scenes viewed on typical low dynamic range (LDR) displays have lower contrast and smaller color gamut than what the eyes can perceive. This leads to loss of visual details and in some cases can even lead to misrepresentation of the scene information. To overcome such limitations, high dynamic range (HDR) has recently gained popularity in both academia and industry. By representing the scene in terms of physical luminance information, HDR can achieve very high contrasts and a wider color gamut, in effect matching the human instantaneous vision range. Due to allowing more scene information representation, HDR helps to capture very fine details which are otherwise difficult to be retained with traditional photography. A visual example to illustrate this is shown in Fig. 5.2. The scene in question has very bright sunlight, shadows, and other details. With single exposure photograph, we can either retain the information in darker areas (longer exposure time) or the ones in brighter areas (shorter exposure time). In both cases, we tend to lose out information either in dark or bright areas. As shown in this figure, the first two are single exposure images with different exposure values (EV). EV basically controls the amount of light allowed while capturing the scene. The third image is of the same scene but HDR processed (tone mapped to 8-bit precision). The reader will notice that this image preserves more details and has a better overall contrast. In other words, HDR helps to retain visual information in very bright and dark areas by minimizing over/under exposure. This leads to better visual experience for the viewers and this is particularly relevant in the context of the recent paradigm shift towards quality of experience (QoE) based multimedia signal processing.

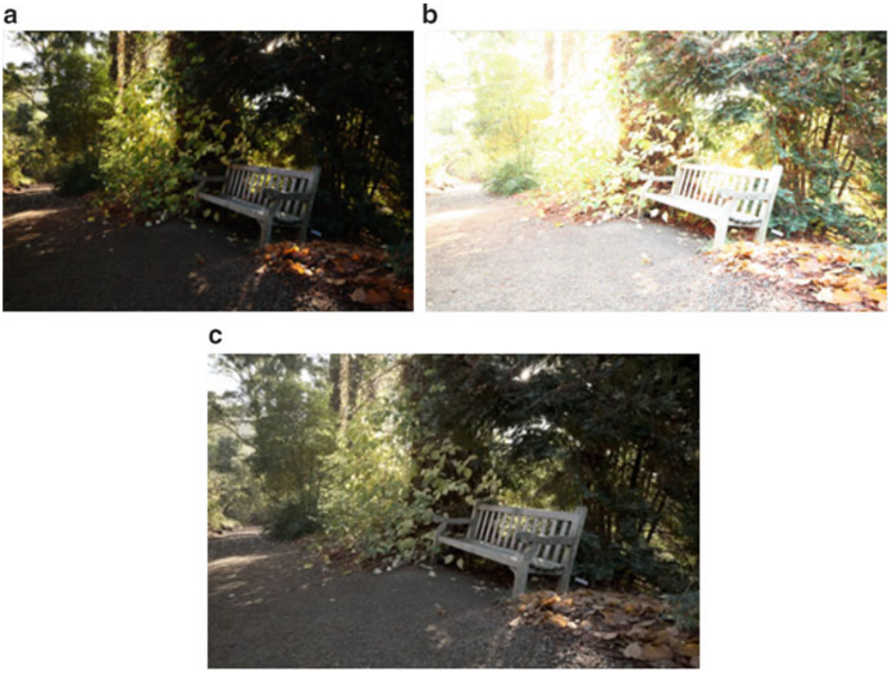


Fig. 5.2 Advantage of HDR over traditional photography. (**a**) Single exposure (−6.89 EV), (**b**) Single exposure (−2.89 EV), and (**c**) HDR processed image (8-bit precision)

Such QoE driven multimedia systems have increasingly come in focus in recent years, both from research and industry perspectives. The aim to capture the end-users' aesthetic expectations rather than simply delivering content based on a technology-centric approach. As discussed, given the specific characteristics of HDR, it is one of the exciting fields towards providing the end users a more immersive and realistic viewing experience and thus improving the QoE. The aim of this chapter is therefore to provide the reader an overview of HDR from the viewpoint of visual experience and in the process outline the challenges that exist.

5.2 The HDR Pipeline

Owing to the characteristics of the HDR signal, its processing right from generation to transmission requires specific tools and different approaches than the traditional LDR processing. A simplified block diagram of a typical HDR processing pipeline is shown in Fig. 5.3. HDR content is first generated by a convenient method. It then needs to be stored appropriately (e.g., after compression). The next step involves suitable processing (e.g., pre-processing) to transmit it to the end user. The end

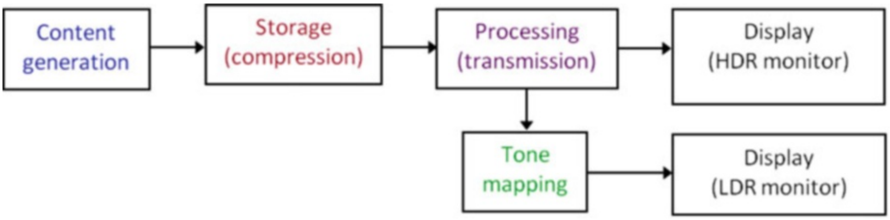


Fig. 5.3 A simplified overview of an HDR processing pipeline

user can view the processed HDR content either directly on an HDR monitor or on an LDR display. In the latter case, a further operation referred to as tone mapping needs to be carried out to fit the dynamic range of the HDR signal to that of the LDR display. An important distinction that should be made at this point is regarding the usage of the terms “content” and “scene.” Throughout the chapter, we will use them in a generic sense in that they can refer to both still images and videos. While capturing HDR still images and videos follows similar approaches, video involves the additional temporal dimension. This introduces more factors that need careful considerations.

5.2.1 Capture

At present, there are several methods to generate HDR content [7]. We will briefly discuss three of them in the following. The first among them employs a weighted fusion of LDR images captured at different exposure levels. Most of the currently available consumer cameras capture 8-bit images (or 14-bit in RAW format). The limited bit-depth is not sufficient to represent the entire dynamic range of a typical real-world scene. Therefore, the authors in [37] proposed the idea of multi-exposure fusion of LDR photographs of the same scene. The underlying goal is to incorporate details from each exposure (from brightest to darkest scene areas) and thus obtain a more realistic appearance of the scene. A visual example is illustrated in Fig. 5.4 from which the reader can notice that the processed HDR images, (g) and (h), incorporate more details than single exposure photographs. While such multi-exposure fusion based HDR capture is reasonably effective, it is certainly not without its challenges. Specifically, there are two major issues that need to be highlighted. First, it is cumbersome since one has to manually obtain several photographs from the same scene by varying the shutter speeds. This is further complicated by the fact that different scenes may require different shutter speeds depending on the dynamic range of the scene in question. The second and more serious issue is with regard to pixel alignment and motion. Thus, utmost care is required while capturing (e.g., using a tripod stand) the LDR photograph. Since the idea is to capture the same scene with varying exposures, motion or instantaneous

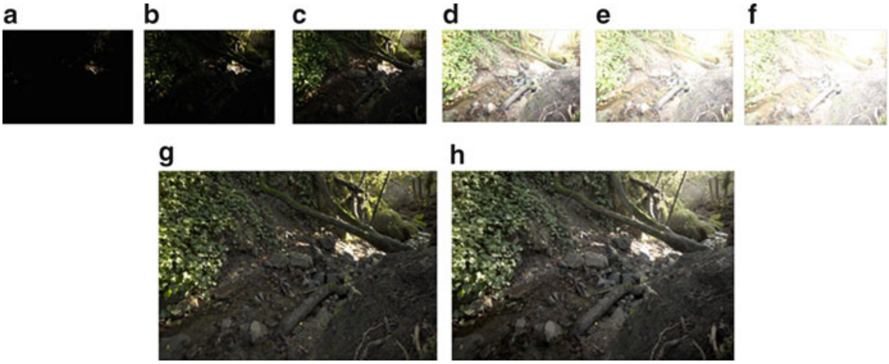


Fig. 5.4 Multi-exposure images and the resulting HDR (one mapped) images, (**a**) −7.21 EV, (**b**) −5.47, (**c**) −3.89 EV, (**d**) −0.19 EV, (**e**) 0.76 EV (**f**) 1.44 EV, HDR image (with luminance range from 10^{-1} cd/m^2 to 10^4 cd/m^2) formed by fusing (**a**)–(**f**). (**g**) and (**h**) Locally rendered tone mapped images obtained by tone mapping the HDR using two different TMOs

change in the scene (e.g., a moving person) will severely hinder the effectiveness of this method. This problem is obviously more pronounced in capturing an HDR video and displaying it in real-time. For example, consider a typical scenario with the rate to play back a video being 25 frames per second. In this case, creating 25 HDR frames per second should be the goal of HDR video. Assuming that an HDR image can be created from fourexposures, a camera would need to capture 100 exposures per second [5]. In addition to this being a challenge for the image sensor and its data interface, there is only one hundredth of a second left for exposing each image. For this to be enough, a lens with a large aperture and possibly an increased gain is required. As already mentioned, it is also likely that the camera or the objects in the scene move while acquiring the sequence of exposures. Therefore, in order to merge them together, the intermediate camera and scene motion must be compensated, otherwise there would be motion blur or ghosting artifacts. Note that motion compensation is a computationally costly step and can introduce significant overhead in the HDR video capture system. Once the images/frames are aligned, they can be merged together into an HDR frame. With regard to the processing time constraints, producing 25 HDR frames per second implies that there are only 40 ms of processing time available for each frame. Capturing the LDR exposures, aligning and merging them and then tone mapping the result for display thus needs to be performed within 40 ms. Given these, it is evident that multi-exposure fusion based HDR video capture is bound to be challenging. Some of these issues, however, can be better handled by the second method for HDR capture through the use of more specialized cameras. There a few commercially available cameras (e.g., SpheronCam HDR by SpheronVR [4]) that have an in-built multi-exposure capturing. Given the recent advances in hardware technologies, it is likely that such cameras will become more common. The third method creates HDR content from virtual environments using physically based renderers. This is more commonly employed in entertainment industries (e.g., digital cinema). At this point, the reader is referred to [7] for further details on the methods for HDR content capture. Given the focus of this chapter, it is assumed that we have a well-captured and realistic HDR scene from either of the mentioned methods and the goal is to process this further keeping in mind the visual appearance to the end user.

5.2.2 Storage

Once the HDR content is generated, it needs to be stored. As pointed out earlier, an HDR pixel is represented using three single precision floating point numbers. This implies that each pixel requires 12 bytes of memory. A simple computation will reveal that this corresponds to approximately 24 MB of data for high definition (HD) resolution of (1,920 by 1,080 pixels). In contrast, an equivalent uncompressed LDR representation (24 bits per pixel) of the same scene will require only a fourth (about 6 MB) of this memory. It is therefore clear that there is need for efficient compression methods to allow for a more compact HDR storage given the high memory demands.

One of the first solutions towards this was proposed in [14] by the introduction of RGBE. This method stores a shared exponent between the three color channels under the assumption that it does not vary much between them. Consequently, RGBE leads to four bytes per pixel (one byte for each color channel plus one byte for the shared exponent). This immediately reduces the memory requirement to one-third of the uncompressed version (which needs 12 bytes per pixel). Another HDR compression approach proposed in [16] is known as the LogLuv encoding. As the name implies, this format stores the luminance in the logarithmic domain and also assigns more bits to it than to the colors. The underlying principle for LogLuv encoding is that human eyes are more sensitive to luminance information than color (so more bits are devoted to the luminance component). In addition, it has been found that the response of human eyes to the absolute luminance levels is approximately logarithmic. Thus, LogLuv encodes logarithm of luminance (as an additional advantage logarithm operation also help in dynamic range compression). Another common HDR format is the half-floating point format, which is a part of the specification of the OpenExr format [2].

Despite the existence of efficient formats for HDR storage (like RGBE, OpenExr), there is a need for developing techniques for further HDR compression. This is because even with the HDR formats there is a huge memory requirement. Consequently, HDR content stored in a standard HDR format should be compressed further to enable more practical deployment and real-time processing. So there is need for research into effective HDR compression schemes and this therefore has been an important research area. A crucial and related issue is that the existing coding architectures have become widely adopted standards supported by almost all software and hardware equipment dealing with digital imaging. As a result, it

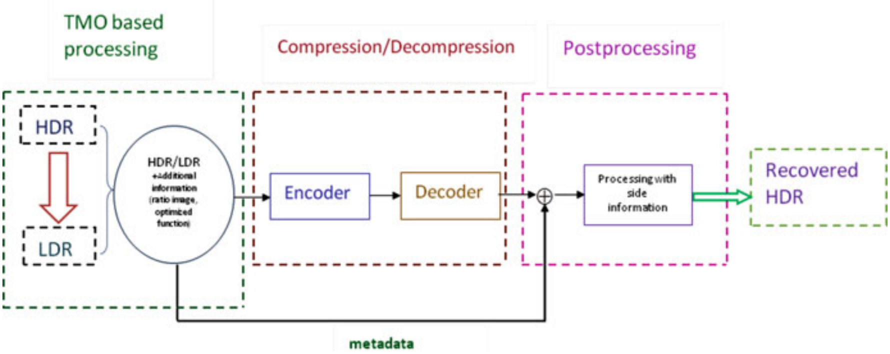

TMO based processing
Compression/Decompression
Postprocessing
HDR
LDR
HDR/LDR +Additional information (ratio image, optimized function)
Encoder
Decoder
Processing with side information
Recovered HDR
metadata

Fig. 5.5 Block diagram of a typical backward compatible HDR compression pipeline

will be of great interest to design HDR compression schemes that are compatible with existing coding architectures. Not surprisingly, substantial research effort has been put into designing HDR compression systems that are backward compatible (for example [17, 30, 33]) with the standard image (e.g., JPEG and JPEG 2000) and video coders (e.g., H.264/AVC). So the aim is to customize the existing codecs so that they can cater to HDR images and videos. Towards that end, dynamic range reduction (or tone mapping) is usually adopted as the first step towards backward compatible HDR compression. In fact, the output of tone mapping is an LDR signal, which requires much smaller memory for storage.

A block diagram showing the steps in a typical backwards compatible HDR compression pipeline is shown in Fig. 5.5. In this figure, we can separate out three main blocks. The first one is the tone mapping operator (TMO) based processing module. Here, a TMO is often used to create an LDR version of the HDR image or video frame. Based on the HDR image, the LDR image and/or the TMO configuration, side information (for instance, a ratio image as in [17] or a non-linear mapping function as in [30]) that will facilitate the decoder's operation is generated. The second block involves the compression of either the tone mapped LDR content or a modified HDR via an existing compression scheme. The encoded bit streams along with metadata are subsequently transmitted to the decoder. The last block, i.e. the post-processing, performs the inverse tone mapping based upon the side information delivered together with the LDR bit stream, re-converting the decoded LDR image into its HDR format. The reader will notice that the HDR compression pipeline shown in Fig. 5.5 is almost entirely compatible with existing coding architectures. The main difference is the extra metadata that needs to be transmitted for enabling the reconstruction of the HDR image. With such architecture, the focus basically shifts to the first and third blocks, namely tone mapping and post processing (or inverse tone mapping). Thus, the problem of HDR compression becomes one of identifying the appropriate tone mapping and inverse tone mapping algorithms. However, this is not straightforward given the fact that it is not easy to convert an HDR image/video to LDR without losing perceivable visual information (related to the difficulty in designing a generic TMO that can retain perceptually useful details). As a result, tone mapping is itself an on-going research topic while inverse tone mapping is an even less investigated topic with few works addressing it (e.g., [8]).

5.2.3 Visualization

Native HDR content visualization via a display is challenging because of physical hardware limitations and not possible with current technologies. With regard to the commercial displays currently available in the consumer market, the average peak luminance (white point) is about 250 cd/m^2 (for LED/LCD) and even lesser (100 cd/m^2) for Plasma displays. As a result, the range of luminance in HDR will almost certainly exceed that of such displays. Additionally, the contrast ratio of LDR displays is not good enough for displaying HDR content. For example, even a good in-plane switching (IPS) LCD panel can achieve a contrast ratio of only about 1000:1 while the required contrast ratio of typical HDR scenes can be more than 10^6:1. More recently, displays with much higher contrast ratio and displayable luminance range have appeared in the market. One such display is the SIM2 Solar47 HDR display [3]. The Solar 47 is a 47-inch, 1080p LCD TV with 2202 white LEDs arrayed behind the imaging panel, and unlike other local-dimming sets, each LED is individually addressable. The core technology in this HDR monitor follows the one proposed in [18]. HDR monitors are typically based on two technologies: (a) modulating a liquid crystal display (LCD) panel using a set of powerful light emitting diodes (LEDs) as the backlight, (b) video projector based. Particularly, the SIM2 HDR display [3] uses an LCD panel, replacing its common back light unit (B.L.U.), typically based on a set of cold cathode fluorescent tube lamps (CCFL), with an array of high-power white LEDs. The idea is to light each small zone of the picture displayed on the LCD, with an LED driven by the specific luminous intensity of that small area of the picture. That means if a scene has black details, the LEDs under those details are turned off to achieve a true black, while where is a high luminous intensity area, LEDs under it are turned on to maximum power. Gray scale areas are then obtained by modulating to intermediate levels the intensity of the LEDs using the HDR picture processing [3].

It should however be made clear that strictly speaking, displays like SIM2 Solar47 cannot be categorized as being entirely HDR since even they cannot display luminance beyond the specified (e.g., in SIM2 the said limit is about 4,000 cd/m^2) limit. Therefore, some kind of range reduction operation (tone mapping) is almost always needed in order to display HDR either on LDR or HDR display. Not surprisingly, an important issue in HDRI is to reduce the dynamic range of the HDR content towards its visualization. This problem has been commonly addressed by employing TMOs and is an important aspect in HDR processing and display. Additionally, tone mapping facilitates the development of backward compatible

HDR compression whereby existing coding standards can be exploited for HDR signal encoding. It is therefore crucial to analyze and understand the impact of TMOs on the overall appearance of the HDR content. This is discussed in the next section.

5.3 Tone Mapping and Its Impact on Visual Experience

Tone mapping is the operation that adapts the dynamic range of HDR content to suit the lower dynamic range available on a given display. The idea is to process the HDR content so that the discrepancy between the tone mapped content and the HDR content is minimal from the viewpoint of two observers, one observing the tone mapped content while the other viewing the actual HDR content. Thus, tone mapping attempts to retain important characteristics of the original HDR content such as local and global contrast, details, naturalness, etc.

5.3.1 Tone Mapping Operators

Several TMOs have been developed over the past years. Some are simple and based on operations such as linear scaling and clipping while the more sophisticated ones exploit several properties of the human visual system (HVS) with the aim of preserving the details. But more often than not, TMOs lead to information loss which can reduce the perceptual quality of the tone mapped contents. This is expected since dynamic range compression invariably tends to destroy important details and textures and can introduce additional artifacts related to changes in contrast and brightness.

TMOs can be broadly classified into two categories, namely local operators and global operators. As the name implies, local operators employ a spatially varying mapping which depends on the local image content. As opposed to this, global operators use the same mapping function for the whole image. Chiu et al. [22] introduced one of the first local TMOs by employing a local intensity function based on a low-pass filter to scale the local pixel values. The method proposed by Fattal et al. [32] is based on compressing the magnitudes of large gradients and solves the Poisson equation on the modified gradient field to obtain tone mapped images. Durand et al. [11] presented a TMO based on the assumption that an HDR image can be decomposed into a base image and a detail image. The contrast of the base layer is reduced using an edge-preserving filter (known as the bilateral filter). The tone mapped image is obtained as a result of multiplication of the contrast reduced base layer with the detail image. Drago et al. [9] adopted logarithmic compression of the luminance values for dynamic range reduction in HDR images. They use adaptively varying logarithmic bases in order to preserve local details and contrast.

The TMO proposed by Ashikimin [23] first estimates the local adaptation luminance at each point which is then compressed using a simple mapping function. In the second stage, the details lost in the first stage are re-introduced to obtain the final tone mapped image. Reinhard et al. [6] applied the dodging and burning technique (traditionally used in photography) for dynamic range compression. A TMO based on a perceptual framework for contrast processing in HDR images was introduced by Mantiuk et al. [35]. This operator involves the transformation of an image from luminance to a pyramid of low-pass contrast images and then to the visual response space. It was claimed that in this framework, dynamic range reduction can be achieved by a simple scaling of the input. Another TMO known as iCAM06 [19] has also been developed. It is based on the sophisticated image color appearance model (iCAM) and incorporates the spatial processing models in the HVS for contrast enhancement, photoreceptor light adaptation functions that enhance local details in highlights and shadows. With regard to global TMOs, the simplest one is the linear operation in which the maximum input luminance is mapped to the maximum output value (the maximum luminance mapping) or the average luminance mapping (i.e., mapping average input luminance to the average output value). Another global TMO is the one proposed by Ward [15] which focuses on the preservation of perceived contrast. In this method, the scaling factor is derived from a psychophysical contrast sensitivity model. Tumblin et al. [21] have reported a TMO based on the assumption that a real-world observer should be the same as a display observer. These are some of the existing TMOs and the list is by no means exhaustive. The interested reader is also referred to survey papers on the topic (e.g., [25]) for a more complete and detailed study of TMOs.

The reader may have noticed that local TMOs seem to have received more attention than the global ones. This is partly due to the fact that as a result of their design local TMOs perform well in preserving the local details (but are less effective in reproducing the overall brightness and contrast). On the other hand, although global TMOs preserve the overall contrast they usually lead to loss of local details. But as an important advantage, global operators are generally computationally more efficient than the local ones. So local and global TMOs have their own advantages and disadvantages. Since tone mapping reduces the dynamic range, it will invariably lead to loss of visual details and as a result affect the perceived appearance of the HDR content. Given that tone mapping is often needed at different stages of HDR pipeline (e.g., for compression and visualization), it is therefore necessary to analyze how they affect the visual experience of the processed HDR content. It should be mentioned that evaluating the overall HDR viewing experience is not an easy task since it is a multi-dimensional phenomenon. Nonetheless, we identify three important attributes that play a significant role in the viewing experience: perceptual visual quality, visual attention, and naturalness. Therefore, we first analyze how TMOs affect perceptual quality and then discuss their impact on visual attention.

5.3.2 Tone Mapping and Visual Quality

There have been several studies related to how TMOs affect visual quality of the tone mapped content. We first briefly describe some of the existing studies related to subjective evaluation of TMOs.

The psychophysical experiments carried out by Drago et al. [10] aimed to evaluate six TMOs with regard to similarity and preference. Three perceptual attributes, namely apparent image contrast, apparent level of detail (visibility of scene features), and apparent naturalness (the degree to which the image resembled a realistic scene) were investigated. It was found that naturalness and details are important attributes for perceptual evaluation of TMOs. The study by Kuang et al. [20] performed a series of three experiments. The first one aimed to test the performance of TMOs with regard to image preference. For this experiment, 12 HDR images were tone mapped using six different TMOs and evaluation was done using the paired comparison methodology. The second experiment dealt with the criteria (or attributes) observers used to scale image preference. The attributes that were investigated included highlight details, shadow details, overall contrast, sharpness, colorfulness, and the appearance of artifacts. The subsequent regression analysis showed that the rating scale of a single image appearance attribute is often capable of predicting the overall preference. The third experiment was designed to evaluate HDR rendering algorithms for their perceptual accuracy of reproducing the appearance of real-world scenes. To that end, a direct comparison between three HDR real-world scenes and their corresponding rendered images displayed on a low dynamic-range LCD monitor was employed. Yoshida et al. [1] conducted psychophysical experiments which involved the comparison between two real-world scenes and their corresponding tone mapped images (obtained by applying seven different TMOs to the HDR images of those scenes).

Similar to other studies, this one was also aimed at assessing the differences in how tone mapped images are perceived by human observers and was based on four attributes: image naturalness, overall contrast, overall brightness, and detail reproduction in dark and bright image regions. In the experiments conducted by Ledda et al. [31], the subjects were presented three images at a time: the reference HDR image displayed on an HDR display and two tone mapped images viewed on LCD monitors. They had to choose the image closest to the reference. Because an HDR display was used, factors such as controlling screen resolution, dimensions, colorimetry, viewing distance, and ambient lighting could be controlled. This is in contrast to using a real-world scene as a reference which might introduce uncontrolled variables. The authors have also reported the statistical analysis of the subjective data with respect to the overall quality and to the reproduction of features and details. Different from the mentioned studies, Cadik et al. [25] adopted both a direct rating (with reference) comparison of the tone mapped images to the real scenes, and a subjective ranking of tone mapped images without a real reference. They further derived an overall image quality estimate by defining a relationship (based on multivariate linear regression) between the attributes: reproduction of

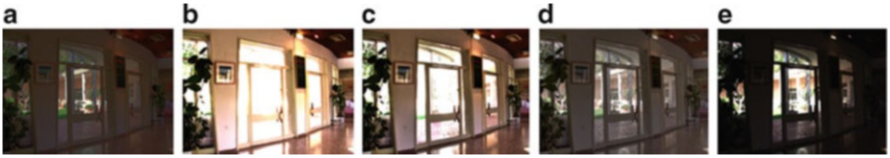


Fig. 5.6 Different visual qualities of LDR images generated by tone mapping an HDR image by different TMOs, (**a**) Ashikmin TMO, (**b**) Drago TMO, (**c**) Drand TMO, (**d**) icam06 TMO, (**e**) linear TMO

brightness, color, contrast, detail and visibility of artifacts. The analysis further revealed that contrast, color, and artifacts are the major contributing factors in the overall judgment of the perceptual quality. However, it was also argued that the effect of attributes such as brightness is indirectly incorporated through other attributes. Another conclusion from this study was that there was agreement between the ranking (of two tone mapped images) and rating (with respect to a real scene) experiments. In contrast to this last observation, Ashikimin et al. [24] found that there were significant differences in subjective opinions depending on whether a real scene is used as a reference or not. A recent survey can be found in [12] that evaluated TMOs for HDR video.

It should be emphasized that most of these studies either ranked the TMOs based on the performance in the respective subjective experiments or outlined the factors affecting visual quality of the tone mapped content. However, it might be misleading to generalize the results from these studies since the number of HDR stimuli was limited. Nevertheless, all of them establish beyond doubt that tone mapping (both in still images and videos) tends to not only reduce the visual quality but also affects the naturalness of the processed HDR content (in addition for video stimuli there could be visible temporal artifacts). Because the underlying philosophy of TMOs concerns with reducing they range, they inevitably saturate visual information leading to loss of details. As a consequence, their use for HDR visualization calls for extreme care. An example to show that TMOs can result in very different visual qualities for the same HDR content is given in Fig. 5.6. Observe how some TMO preserve only indoor details while some do so for outdoor information. Ashikmin and icam06 [19] TMOs seem to provide a better trade-off in maintaining visual details in outdoor and indoor simultaneously.

A related aspect in tone mapping is that of naturalness. While it is quite clear that tone mapping reduces visual quality, how it affects naturalness remains an unanswered question. In fact, all the user studies described previously implicitly account for naturalness. This is because when human observers judge the visual quality of tone mapped content, not only the presence or absence of visual details affects their choice but is also affected implicitly by the naturalness of the tone mapped content. Naturalness is a subjective quality which is difficult to be quantified. In the light of this, it is not surprising that most of the TMOs only focus on retaining details and/or maintaining local and global contrast but do not consider naturalness for processing the HDR content. For instance, an over enhanced tone mapped image might have a very large number of details but can still have poor visual appearance due to being unnatural.

Thus, we have provided a brief overview of the impact of one mapping on visual quality of HDR content. We have also provided several useful references for the reader to explore further. In summary, tone mapping in general, degrades visual quality by destroying scene details, ad-hoc saturation of pixels as well as affecting the natural appearance of the content. In the next section, we discuss the effect of tone mapping on visual attention.

5.3.3 Tone Mapping and Visual Attention

As mentioned in the previous section, the current effort in subjective evaluation has been mainly directed towards assessing the impact of TMOs from quality and aesthetic appeal point of view. From these we may be able to study and analyze people's preference regarding visual appeal of the tone mapped content. However, visual quality is just one of the several aspects that need to be considered to make conclusions on how TMOs affect the overall QoE. One such issue is that of visual attention (VA) which has been well recognized as a crucial aspect in perceptual visual signal processing. It is well known that human eyes tend to focus more on certain areas in an image/video than others. Stated differently, some regions attract more eye attention and these are termed as salient regions. VA is therefore the ability of the HVS to find and focus on relevant information quickly and efficiently [36]. This has several applications since the more important signal information can be extracted and processed accordingly. For example in image/video coding, the visually salient parts can be assigned more bits in order to achieve higher efficiency and better visual quality. Further from an artistic viewpoint, TMOs could possibly change the way a scene is perceived by human eyes. This may lead to changes in the feelings and emotions conveyed by the image. Thus the intention of the artist/content author may not be represented correctly to the viewer. For instance, intricate details (like very fine texture) in some part of an image which attract viewer attention might be lost due to tone mapping and the photographer's intention of producing a compelling picture is jeopardized. It is thus clear that VA plays an important role in human perception and therefore the impact of TMOs should also be analyzed from this viewpoint as well. In the context of VA, eye-tracking is the term commonly used to denote the way of exploring what people look at in any given situation and record their VA strategies with location and duration. The impact of TMOs on visual attention is best explained through VA maps obtained from human observers. To begin the analysis, we recall that TMOs tend to destroy visual details. Moreover, given that some TMOs can preserve details better than others, the VA behavior can change with TMO. To visually exemplify this, we have shown in Fig. 5.7, the tone mapped versions of an HDR image processed by Ashikimin and Tumblin TMOs. We can immediately make two observations from this figure. Firstly, the image

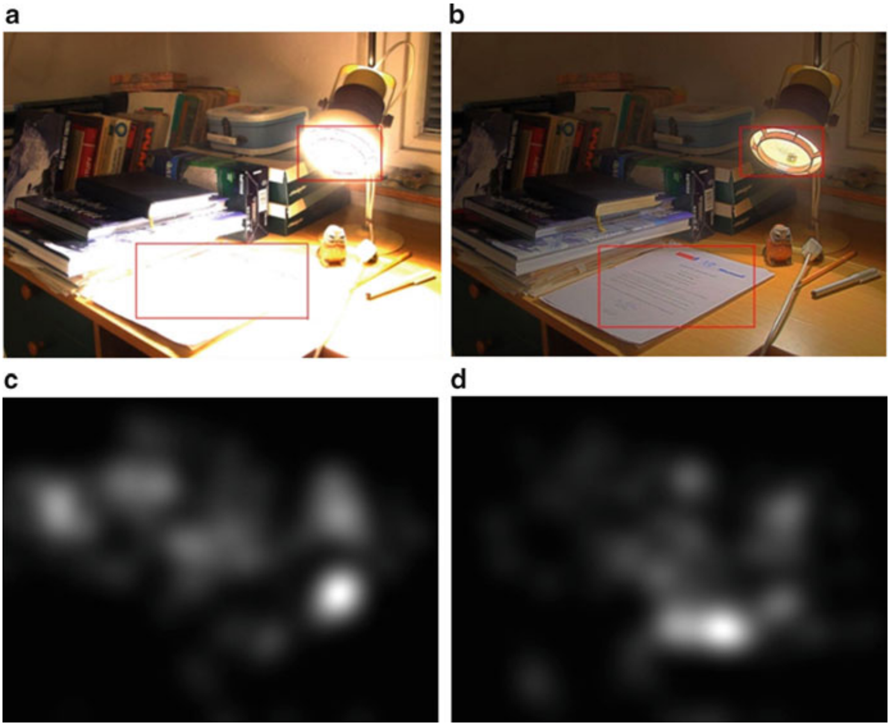


Fig. 5.7 Illustration of the effect of global and local TMOs. (**a**) Image processed by Tumblin TMO (global), (**b**) Image processed by Ashikimin TMO (local), (**c**) VA map for (**a**) and (**d**) VA map for (**b**). The *red boxes* highlight two areas in the images where details are lost and preserved by global and local TMOs, respectively

processed by Ashikimin TMO has more details preserved in the regions highlighted by red boxes. As opposed to this, in the same regions of the image processed by Tumblin TMO, the details are clearly missing. Secondly, the overall contrast of the image in Fig. 5.7a is clearly better than the one in Fig. 5.7b. To visualize and relate this to the impact of these TMOs on VA, we have shown the corresponding VA maps obtained from eye-tracking. The reader can notice that for the image processed by Tumblin, the attention regions are mainly the books in the background while the letter pad (in the foreground) is nearly unnoticed by the observers. We hypothesize that this happens because with Tumblin being a global TMO the overall image contrast is maintained and the details in the dark areas of the image (like the books in background) are well retained. Further the owl below the lamp is also clearly visible and is a salient region. However, as already mentioned, all this comes with the price of losing finer details mainly in the bright areas (like the lamp and the letter pad) as highlighted. As a result these letter pad is nearly non-salient since the useful information (the text inside) has been washed away by the TMO. In contrast to this, the VA map of the image processed by Ashikimin TMO shows that the written text on the letter pad is the most salient portion. Also, the darker background (mainly the books) seems to have become less eye-catching since the contrast in that part is reduced. Another example to illustrate that TMOs can modify the attention regions is shown in Fig. 5.8. Here the images in the first and second rows are the tone mapped versions of rend02_oC95 image processed by Drago and iCAM06 TMOs and the corresponding human priority maps (VA maps), respectively. It can be seen that the two red mats (highlighted by red boxes) are more clearly visible in the image processed by iCAM06 since there is high contrast preserved in and around that region.

Consequently, one can see from the corresponding VA map that these indeed are salient regions for the human observers. On the other hand, in the image processed by Drago [9] there is much lower contrast in the said regions. As a result of these attract much lesser eye attention as seen in the corresponding VA map. A second set of examples is shown in the third and fourth rows of Fig. 5.8. Here the third row shows three tone mapped versions of Oxford_Church image (tone mapped by Tumblin, iCAM06 and Linear TMOs) while the corresponding VA maps are shown in the fourth row below each image. Again, one finds that the orange spot (highlighted by red box) is a salient region only in case on linear TMO (see the VA map in Fig. 5.8j) since this TMO destroys contrast on other regions which makes the orange spot stand out and thus eye catching. As opposed to this, Tumblin and iCAM06 TMOs provide much better contrast in other parts of the image as well. So the orange spot is nearly non-salient in these two images as the observers attention is attracted to other parts. Therefore, based on our experiments and analysis of the VA maps obtained from eye-tracking experiments, we can say that contrast of the resultant tone mapped image plays an important role in VA behavior. As exemplified by visual examples in Figs. 5.7 and 5.8, the areas that attract eye-attention can vary even within the same image depending on whether contrast is preserved or destroyed by the TMO. This therefore suggests that contrast indeed is a vital dimension in HDR content processing from VA viewpoint.

Tone mapping can also be viewed in terms of reduced signal contrast due to tone mapping. With the reduced contrast, regions that may have attracted observers' attention in the HDR content might be reduced. As a direct consequence, the number of salient regions in tone mapped HDR content tend to decrease. To visually exemplify this, consider Fig. 5.9, where image (a) is a tone mapped version (using Reinhard TMO) of an HDR image. In this image, we can easily identify the foreground (mainly comprising of the headlight and front wheel of the bike) and the background (bicycles and the door). The image (b) shows the VA map of image (a) while the corresponding HDR VA map is shown in (c). Observe how the VA map in (b) indicates very few salient points in the background. As opposed to this, the HDR VA map shows that background also had regions which attracted eye attention. The reason is obvious: tone mapping in this case destroys details mainly in the background but the foreground is fairly well preserved in terms of contrast. As a result, the number of salient points in the background reduces drastically. The last visual example is shown in Fig. 5.10. In this figure, the second row shows the corresponding VA maps of the images shown in the first row. Since HDR image

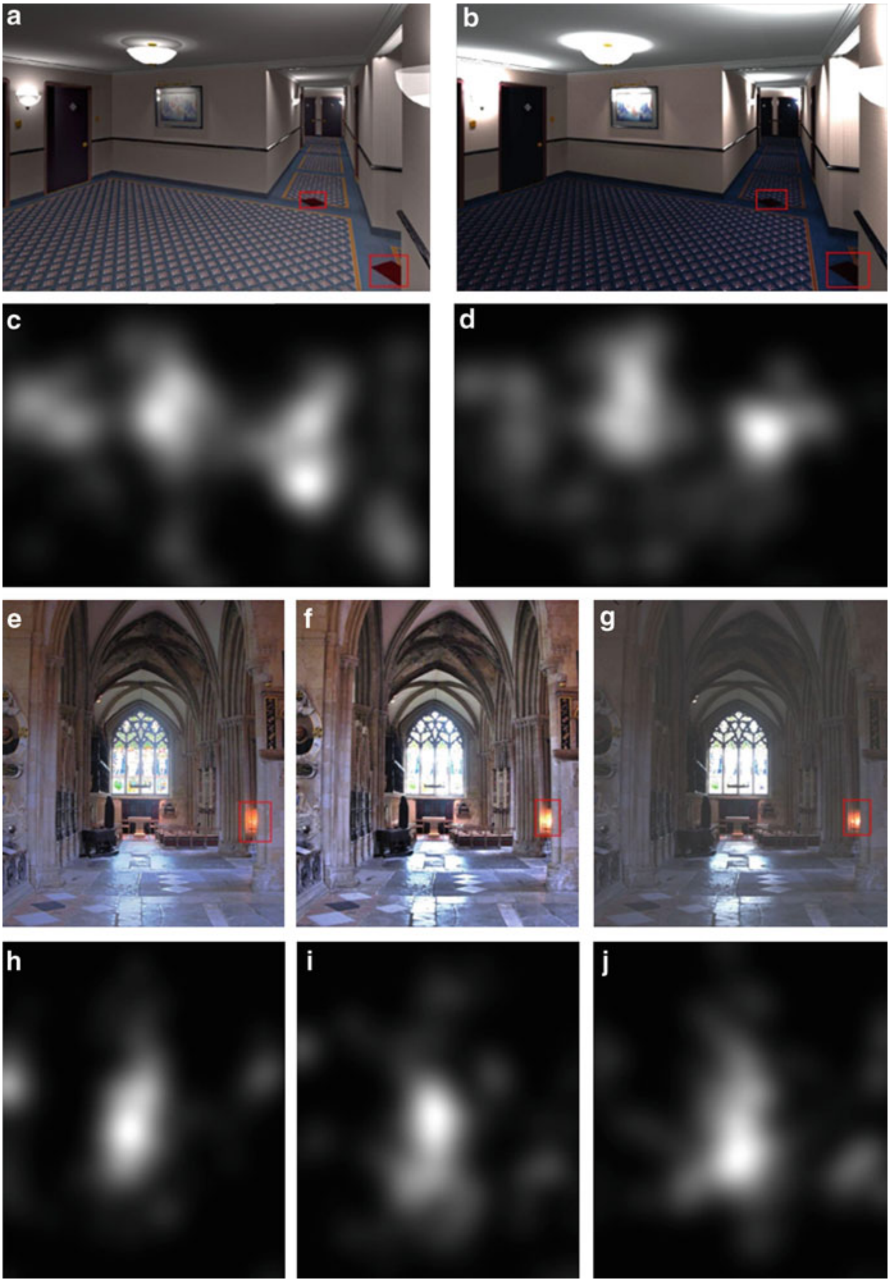


Fig. 5.8 Effect of TMOs on VA. (**a**) rend02_oC95 image processed by iCAM06 TMO, (**b**) rend02_oC95 image processed by Drago TMO, (**c**) VA map for (**a**) and (**d**) VA map for (**b**), (**e**)–(**g**) Oxford_Church image processed by Tumblin, iCAM06 and linear TMOs, respectively, (**h**)–(**j**) VA maps for the images shown in (**e**)–(**g**), respectively. The *red boxes* highlight the area(s) in the images which become salient or non-salient depending on the overall impact of TMO

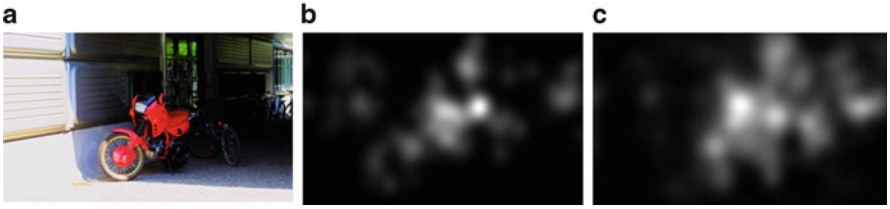


Fig. 5.9 Effect of TMOs on VA. (**a**) tone mapped version of "moto" (Reinhard TMO), (**b**) VA map of image (**a**) obtained from eye-tracking, and (**c**) VA map of "moto" HDR image obtained from eye-tracking

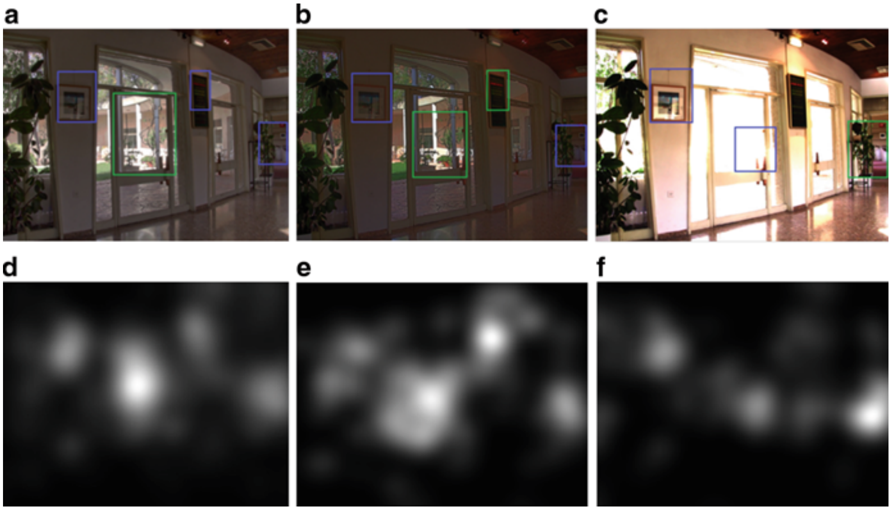


Fig. 5.10 Effect of TMOs on VA. (**a**) "dani_belgium" HDR image (for sake of better visualization tone mapped image processed by icam06 TMO has been displayed), (**b**) tone mapped version of "dani_belgium" (Ashikimin TMO), (**c**) tone mapped version of "dani_belgium" (Drago TMO), (**d**) VA map of HDR image, (**e**)–(**f**) corresponding VA maps of the images shown above

does not display properly, we have shown image processed by icam06 (instead of the original HDR image) for the sake of explanation. Also note that we have used green box to highlight the area(s) that attracted maximum human attention and blue box for area(s) with relatively lower attention.

Considering the HDR VA map in Fig. 5.10d, it shows that there are four main regions which are salient according to human observers. These have been highlighted in image (a) shown just above the HDR VA map and include the outside area seen through the door (highlighted through the green box) and the paintings/board (highlighted through blue boxes). Also, notice that the area highlighted in green attracts more attention as compared to the other three identified regions. Now we observe the effect of tone mapping on these four identified regions. We find that the image processed by Ashikimin TMO (shown in Fig. 5.10b) shows that now there are two regions (highlighted by green boxes) which attract the maximum attention (see the corresponding VA map below this image). Thus, tone mapping modified the visual signal in such a manner that a region which was less salient in the original HDR content has become more salient. Likewise, looking at the VA map in Fig. 5.10b, we find that there is only one region now that attracts maximum attention (this has again been marked in green in the image shown above this VA map) while the attention for other regions reduced considerably. This once again drives home the point: tone mapping can change attentional regions in addition to increasing or decreasing the magnitude of attention. Therefore, as analyzed and explained we note large differences for both intra (i.e., for each image content) and inter (i.e., between different image contents) cases. A theoretical explanation for this is the manner in which TMOs operate. Most of them sacrifice one or the other type of visual information in order to reduce the dynamic range. In the process, additional artifacts (such as additional contours) might be introduced.

The eventual result is that a non-attentional region in the HDR image becomes attentional one in the tone mapped version. The opposite case is that in which structural information is destroyed due to tone mapping. In such cases, an attentional region in the HDR image becomes less important (or less eye catching) in the tone mapped image. For example, a contrast that was visible in the HDR image becomes invisible in the processed image (loss of visible contrast). We have already provided some visual examples to illustrate these points.

Video signals differ from images due to the addition of a temporal dimension in addition to the spatial one. Given that, it will be interesting to analyze the changes in VA behavior due to tone mapping of HDR video sequences. The analysis of the VA maps from different video stimuli leads to similar conclusions as still images. That is TMOs have a large impact on the VA behavior as compared to the HDR video. We present an example in Fig. 5.11. This is from the video sequence[1] "Tunnel1" in which a car is shown to enter into a tunnel (with normal traffic). Inside the

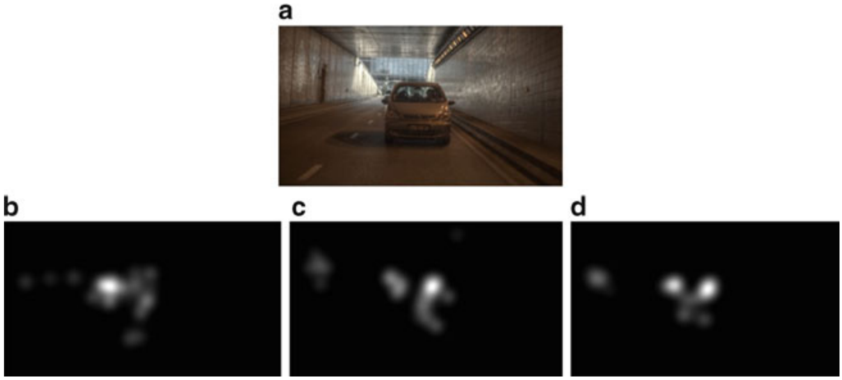


Fig. 5.11 Change in VA behavior in videos, (**a**) video frame, (**b**) HDR VA map, (**c**) VA map of frame tone mapped by Tumblin TMO, and (**d**) VA map of frame tone mapped by Durand TMO

[1] This video sequence was shot as part of the NEVEx project FUI11, related to HDR video chain study.

tunnel, there is relatively lower illumination and so as the car enters into it, there is a large change in scene illumination. Figure 5.11a shows the car inside the tunnel and another car just behind it which also enters the tunnel. Before this time, we found that car was the main region of subject's attention right from the start of the video. However, when the other car enters the frame from behind, it attracts subjects' attention. This is expected since the entry of the new car in the frame is a "new" or a "rare" event (up to this point the subjects' attention is focused on the first car only). That directs the attention to the second car. This is what was observed when the HDR video was viewed on an HDR screen. The corresponding HDR VA map is shown in Fig. 5.11b where one can see that the "second" car is the main region of attention. However, when tone mapped video was shown to the subjects we observed different behavior. In this case, it was found that the "first" car still remains the main focus of attention. This can be clearly observed in the VA map corresponding to Tumblin TMO shown in Fig. 5.11c. That is, despite the occurrence of a new event (the entry of the second car), attention behavior did not change. We can attribute this to the fact that Tumblin TMO could not maintain proper contrast at the tunnel entrance where there is a large change in the intensity (dark inside the tunnel and bright outside it). Due to this, the subjects' attention was not fully diverted towards the "second" car. A different observation was however made for in case of Durand TMO. This TMO could maintain relatively better contrast at the tunnel entrance. Due to this, we have a situation where both the "first" and "second" cars became the regions of attention. This can be seen from the VA map shown in Fig. 5.11d. Thus, depending on the TMO we have different VA behavior for the same scene in the video. This suggests that similar to the case of still images, tone mapping changes VA behavior over time.

Tone mapping is an important HDR processing that enables HDR visualization on traditional display devices. This section was therefore devoted to the study of its impact on the overall QoE. To facilitate discussion, we identified three important dimensions of HDR that tone mapping can affect. These include visual quality, naturalness, and visual attention. The discussion was focused on specifics of how TMO affect these dimensions and several visual examples were provided as illustrations. Regarding visual quality, tone mapping generally leads to loss of contrast (both locally and globally) and result in loss details. This can also have an adverse effect on the naturalness of the tone mapped HDR content. For instance, saturation of color or over-enhancement of details can render the processed content as being unnatural (despite preserving details). In other words, tone mapping can reduce the overall coherency associated with natural signals. We also discussed how tone mapping can impact visual attention behavior. This has the consequence of altering the artistic intention. For example, a photographer might capture a scene with the intention that viewers will pay attention to some areas/aspects of the photograph. However, tone mapping can cause changes that can divert viewers' attention to areas/aspects that are not the same as what the photographer intended. Thus, tone mapping can interfere with the artistic intentions and that can ultimately reduce the visual experience [26,29] of the processed HDR content. Such joint effect of changes in visual quality, naturalness, and/or visual attention have the potential of lowering the enjoyment level of the end-users with regard to HDR viewing.

The reader will notice that the discussions pertaining to tone mapping have been in the context of employing TMOs for HDR visualization. In the next section, we discuss some aspects of quality measurement when viewing an HDR scene.

5.4 HDR QoE

HDR QoE is a rather wide term in that it can include several dimensions including perceptual quality, naturalness, visual attention, aesthetic appeal, and so on. Of course these are not necessarily independent because, for instance, perceptual quality can implicitly account for naturalness. In the following, we first outline the differences between HDR and LDR from the angle of viewing conditions and then introduce the reader to the topic of subjective and objective HDR quality assessment.

5.4.1 Viewing Conditions in HDR

The most important distinction of HDR from LDR is with respect to luminance range (which in turn leads to HDR). Traditional LDR defines a white point (255 for the 8-bit representation). Thus, any intensity more than the defined white needs to be saturated. Moreover, with LDR the pixel values are typically gamma encoded and perceptually uniform. As a result, change in the pixel values can directly be related to the change in visual perception. However, with HDR there is more flexibility to represent the real-world scene luminance without too much saturation. Consequently, there is no fixed white point in HDR that can correspond to the maximum luminance (as it can vary from scene to scene). There is only brighter (or darker) scene intensity. Therefore, HDR viewing involves much higher levels of brightness. Since human vision is sensitive to luminance ratio (rather than absolute luminance), changes in the luminance may not necessarily lead to the same change in visual perception of HDR.

The effect of luminance level on the sensitivity of the HVS is often referred to as luminance masking. Figure 5.12 shows the Campbell–Robson contrast sensitivity chart for two different background luminance levels [38]. For the best viewing, the figure should be viewed on an LCD display of about 200 cd/m^2 and the display function close to the sRGB non-linearity. The solid lines denote the contrast sensitivity of the HVS, which is the contrast level at which the sinusoidal contrast patterns become invisible. Even though the same scales were used for both left and right plots, the CSF is shifted upwards (higher sensitivity) and right (towards higher spatial frequencies) for the brighter pattern. This shows that we are more likely to notice contrast changes, if the stimuli is brighter, as is the case of a brighter display. A further validation of this was done in [38] through a subjective experiment. It was found with statistical evidence that distortions of the same type and with the same

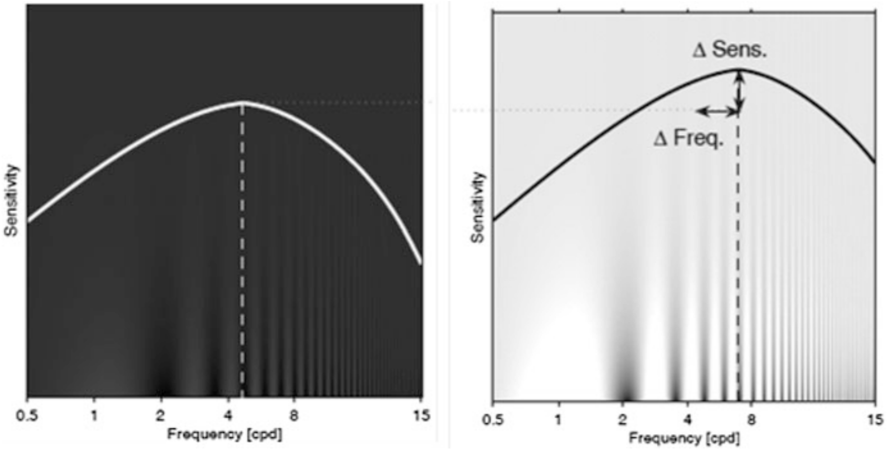


Fig. 5.12 Contrast sensitivity function (CSF) of the human eye in dark (*left*) and bright (*right*) viewing conditions. *Arrows* labeled as $\Delta Sens.$ and $\Delta Freq.$ denote the amount of difference in magnitude and frequency of the peak sensitivity between the dark and bright cases. The figure is reproduced from [38]

magnitude are more annoying when the overall brightness of the image is higher. Thus, with HDR, one needs to take into account the display luminance conditions as it can have a significant impact on the perceived quality of the stimuli.

Another crucial factor with regard to HDR viewing condition is the ambient lighting. Given the high levels of luminance, HDR will require a higher level of ambient lighting as compared to LDR. Obviously with low ambient lighting, HDR viewing can be uncomfortable for viewers. With regard to LDR, the International Telecommunication Union Recommendation (ITU-R) BT500-11 recommendation specifies the room (ambient) illumination to be about 15 % of the perceived screen brightness. It is not clear if this can be applied to HDR viewing. For example, with SIM2 HDR display [3], the maximum luminance is 4,000 cd/m^2 and so the room illumination should be around 600 cd/m^2 according to BT500-11 recommendations. However, it is known that the response of the human eye to luminance is approximately logarithmic and it is likely that a little lower ambient lighting level might be suitable. In any case, it is clear that the current ambient lighting specifications for LDR will not be entirely suitable for HDR viewing.

5.4.2 Subjective Assessment of HDR Quality

Human judgment of visual quality remains the gold standard as far the accuracy of quality prediction is concerned. HDR is no exception. However, as outlined in the previous section, subjective measurement of HDR quality calls for more careful considerations of viewing conditions. Otherwise the results may not reflect the actual perceived quality. Another important factor in HDR subjective test design is the use of TMOs. They will not be used for visualization but for HDR processing (e.g., compression). A problem with TMOs is that they usually require one or more parameters that are left for the user to tune. The issue of best parameter selection is further complicated since it can be HDR content specific. That is, a set of parameters which is suitable for one content may not be optimal for the other (generally speaking the default parameter values might not yield reasonable quality for every HDR content). Therefore, it requires care to find TMO parameters when preparing HDR content for subjective evaluation. Concerning the sources of distortions in HDR, the first is related to tone mapping. Another common distortion is compression related artifacts. Another category of specific artifacts that occur in HDR are those due to inverse tone mapping. Inverse tone mapping is the final step in a typical backwards compatible HDR compression pipeline and can cause saturation, excess or lack of contrast in the HDR scene. Thus, it can be highlighted that HDR processing includes specific distortion sources (that are not typically present in LDR regime) like tone mapping and inverse tone mapping in addition to common artifacts (due to compression, transmission, post-processing etc.). It is also interesting to note that distortions due to tone mapping and inverse tone mapping are not necessarily additive. That is, inverse tone mapping can offset some artifacts from tone mapping. Further, as explained in previous sections, visual attention can be significantly modified due to tone mapping and this can degrade the overall HDR viewing experience. Thus, HDR quality measurement is challenging in that the processed HDR content can suffer from multiple distortions (which need not be independent of each other). This coupled with the fact that the high luminance in HDR can potentially amplify artifacts suggests that extra care needs to be taken for accurate subjective measurement of HDR quality.

Very few research efforts have been reported for subjective HDR quality assessment. The reasons for this are related to the requirement of specialized HDR displays and the unavailability of real HDR content. Nevertheless, two recent studies have employed an HDR display for QoE evaluation. The first one [27] investigated into codec optimization criterion and perceptual quality issues in HDR. The second study [28] analyzed the impact of TMOs in HDR compression. The conclusions from these studies revealed that, indeed, the perceptual quality of the decompressed HDR signal is dependent on the tone mapping method employed and statistical evidences were also presented to support that. The reader will note that both these works are in contrast to most of the existing studies (some of which have been described in previous sections) which focused only on the quality loss in the resultant LDR signal (obtained via tone mapping) and typically employed LDR displays for subjective viewing experiments.

The final point for subjective HDR quality assessment is related to the need of specialized displays. As we have already pointed out, such HDR displays are still not common on a large scale (although this could change within a reasonable time frame). In the light of such constraint, it is natural to ask if HDR quality measurement can be tackled with existing LDR set up (LDR displays, specifications,

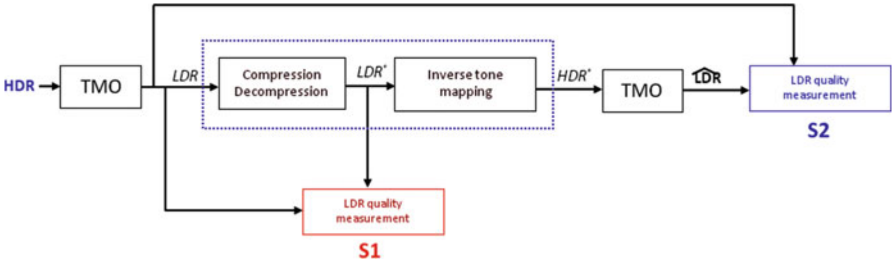


Fig. 5.13 LDR approach to HDR quality assessment

etc.). Indeed, it is not absurd to think of tone mapping the HDR content and estimate its quality subjectively on an LDR monitor. Specifically with regard to quality measurement in a backwards compatible HDR compression system, there can be two possible scenarios to convert HDR quality assessment to an LDR one. These have been illustrated in Fig. 5.13. The first scenario (indicated as S1) is to judge the subjective quality of the decompressed LDR content (before inverse tone mapping) with respect to reference tone mapped LDR. A second possible scenario (indicated as S2 in Fig. 5.13) is to tone map the decompressed HDR content. Then, this can be compared with the tone mapped reference scene to determine its quality. Unfortunately, both the scenarios have their own limitations. The first one is that by viewing HDR on an LDR display, we ignore the distinct aspects of high luminance associated with HDR viewing (explained in the previous section). Another limitation of assessing HDR quality with scenario 1 (S1) is that it ignores the impact of inverse tone mapping completely and takes into account only the compression artifacts. Recall that inverse tone mapping can introduce visible artifacts such as saturation, excess or lack of contrast, etc. Therefore, with this scenario (S1) formulating HDR quality assessment as an LDR one is expected to be less accurate. With scenario 2 (S2) we can note that the tone mapping will operate on different HDR content (pristine HDR and decompressed HDR). Consequently, they will be modified differently and the resulting judgment of visual quality can be erroneous. So it can be concluded that HDR quality can be better judged by simulating proper HDR conditions (use of an HDR display, appropriate ambient lighting, etc.). However, it is fair to reiterate that even HDR displays have their own limitations (in particular due to the dual modulation process) and cannot fully represent HDR content. This is however the best available solution at the time of writing.

5.4.3 *Objective Assessment of HDR Quality*

Objective quality measurement is the use of computational model to predict quality. An objective method for quality prediction is a useful tool in cases where subjective assessment is not feasible (such as real-time applications). Being a mathematical model, an objective method is more convenient to be deployed. However, objective methods cannot be as accurate as the subjective ones. Thus, one line of thinking in the research community has been towards developing more accurate objective methods. While this is reasonable, it is important to understand that the HVS represents a complex visual information processing system. Consequently, all the objective methods (for LDR and HDR) are merely approximations and they cannot be relied upon as a generic solution to quality prediction. Nonetheless, it is also important to highlight that objective methods can achieve a reasonable prediction accuracy in the limited context of an application scenario. For example, the mean squared error (MSE) continues to be deployed extensively in visual data compression. Pertaining to LDR, one finds that a lot of research effort has been spent over the past decade. Most of it is devoted to the development of full-reference methods that require both the reference and processed visual signal for quality computations. In contrast, there exists very few methods for objective HDR quality prediction. The reasons for this are already outlined and related to different viewing conditions as compared to LDR. Thus, mathematical models of HVS's functioning (e.g., contrast sensitivity) used in LDR methods can no longer be effective for HDR. Another reason for slow progress of objective HDR quality measurement can be attributed to the lack of standard databases.

The HDR-VDP-2 (high dynamic range visual difference predictor) [34] is a fairly recent and comprehensive method for objective measurement of HDR quality. It is an extension of the visible differences predictor (VDP) algorithm. The HDR-VDP-2 uses an approximate model of the HVS derived from new contrast sensitivity measurements. Specifically, a customized contrast sensitivity function (CSF) was employed to cover large luminance range as compared to the conventional CSFs. HDR-VDP-2 is essentially a visibility prediction metric. That is, it provides a 2D map with probabilities of detection at each pixel point and this is obviously related to the perceived quality because a higher detection probability implies a higher distortion level at the specific point. Nevertheless, in many cases, it is crucial to know an overall quality score (rather than just the local distortion visibility probability). Pooling is a crucial aspect in converting local error distribution into a single score that denotes the perceptual quality and the HVS can very easily do that accurately. But it is much more difficult to realize that in an objective quality prediction model given the underlying complexities and lack of knowledge of the HVS's pooling mechanisms. It is believed that multiple features jointly affect the HVS's perception of visual quality, and their relationship with the overall quality is possibly nonlinear and difficult to be determined a priori. Therefore, the approach that HDR-VDP-2 takes is that finding the pooling parameters via optimization of correlation with subjective scores.

In its original implementation, the authors of HDR-VDP-2 tried over 20 different combinations of aggregating (or pooling) functions. These included maximum value, percentiles (50, 75, 95) and a range of power means (normalized Minkowski summation) with the exponent ranging from 0.5 to 16. The aim was to maximize the value of Spearman's correlation coefficient in order to find the best pooling function and its parameters. While HDR-VDP-2 is a fairly comprehensive method for HDR quality assessment, there is an issue with regard to pooling in HDR-VDP-2. This is related to parameter optimization. That is, the parameters of the pooling function in HDR-VDP-2 were found by maximizing (optimizing) correlation using existing LDR image databases. Therefore, its effectiveness in predicting the visual quality of HDR images is questionable given the different characteristics LDR and HDR images especially in terms of distortion visibility and overall visual appeal [34]. The reader will notice that objective HDR quality assessment requires much more efforts in terms of both research and implementation. This is more so in the light of the fact that an LDR approach to HDR quality assessment is not as effective and cannot be a substitute to account for the effects that distortions have on HDR viewing.

5.5 Concluding Remarks and Perspectives

HDR imaging is an emerging area within the realm of visual signal processing. It brings to that table two major advantages over the traditional imaging systems. First, it can provide a more immersive and realistic viewing experience to the users. Second, the higher bit-depth required in HDR will allow for more signal manipulation (e.g., pre-processing towards efficient encoding) as compared to the traditional content. However, to exploit HDR technology to its fullest potential, several challenges remain and this chapter has focused on a few of them pertaining to their impact the overall HDR QoE. With regard to HDR processing, tone mapping is often required for HDR viewing on LDR displays, compression, and in many other scenarios where backward compatibility is desired. The aim of this chapter was to throw light on the impact of tone mapping on visual experience. Specifically, we discussed its impact on perceptual quality, visual attention, and naturalness. It is worth highlighting that these play an important role in the QoE in HDR viewing. We reiterate that HDR viewing experience is more immersive than traditional content due to that fact that HDR attempts to reproduce real-world scene information without undue saturation of visual information. In other words, with HDR we directly deal with physical luminance related information and this makes HDR experience more wholesome and enjoyable.

From the objective viewpoint, measurement of HDR QoE remains a challenge primarily due to a larger number of factors involved as compared to traditional video quality. In particular, unlike QoE judgment of traditional visual content, the impact of factors such as naturalness and visual attention modification can be more profound in HDR. Therefore, single measures such as signal fidelity alone cannot be expected to be a reasonable substitute for the overall QoE. On the operational front, HDR poses difficulties because the information is stored in luminance-related format, unlike perceptually scaled pixel values in LDR signals. Finally, native HDR visualization is not possible even with the current HDR display technologies and there is saturation of signal contrast (this is of course due to inherent hardware limitations such as the upper limit on power consumption, heating, etc). Addressing some of the mentioned issues will ultimately be the key to large scale practical deployment of HDR and further interesting applications.

References

1. Yoshida A., Blanz V., Myszkowski K., and Seidel H. Perceptual evaluation of tone mapping operators with real-world scenes. In *Proceedings of SPIE Human Vision & Electronic Imaging X*, pages 192–203, San Jose, CA, USA, 2005.
2. Industrial Light & Magic (2008) OpenEXR. Available at: http://www.openexr.com.
3. SIM2 MULTIMEDIA Available at: http://www.sim2.com/HDR/.
4. Spheron HDR VR. Available at: http://www.spheron.com/home.html.
5. Guthier B. Real-time algorithms for high dynamic range video. *PhD. Thesis*, 2012.
6. Reinhard E., Stark M., Shirley P., and Ferwerda J. Photographic tone reproduction for digital images. *ACM Transactions on Graphics (TOG)*, 21(3):267–276, July 2002.
7. Banterle F., Artusi A., Debattista K., and Chalmers A. *Advanced High Dynamic Range Imaging: Theory and Practice*. AK Peters (CRC Press), Natrick, MA, USA, 2011.
8. Banterle F., Debattista K., Artusi A., Pattanaik S., Myszkowski K., Ledda P., and Chalmers A. High dynamic range imaging and low dynamic range expansion for generating HDR content. *Computer Graphics Forum*, 28(8):3243–2367, December 2009.
9. Drago F., Myszkowski K., Annen T., and Chiba N. Adaptive logarithmic mapping for displaying high contrast scenes. *Computer Graphics Forum*, 22(3):419–426, September 2003.
10. Drago F., Martens W., Myszkowski K., and Seidel H. Perceptual evaluation of tone mapping operators. In *Procedings of ACM SIGGRAPH 2003 Sketches & Applications*, pages 1–1. ACM Press, 2003.
11. Durand F. and Dorsey J. Fast bilateral filtering for the display of high-dynamic-range images. *ACM Transactions on Graphics (TOG)*, 21(3):257–266, July 2002.
12. Eilertsen G., Wanat R., Mantiuk R., and Unger J. Evaluation of tone mapping operators for HDR-video. *Computer Graphics Forum*, 32(7):275–284, October 2013.
13. Mather G. *Foundations of Perception*. Psychology Press, Hove, East Sussex, 2006.
14. Ward G. Real pixels. In *Graphic Gems II*, pages 80–83. Academic Press, 1991.
15. Ward G. A contrast-based scalefactor for luminance display. In *Graphic Gems IV*, pages 415–421. Academic Press, 1994.
16. Ward G. The LogLuv encoding for full gamut, high dynamic range images. *Journal of Graphics Tools*, 3(1):15–31, March 1998.
17. Ward G. and Simmons M. JPEG-HDR: A backwards-compatible high dynamic range extension to jpeg. In *ACM SIGGRAPH 2006 Courses*, 2006.
18. Seetzen H., Heidrich W., Stuerzlinger W., Ward G., Whitehead L., Trentacoste M., Ghosh A., and Vorozcovs A. High dynamic range display systems. *ACM Transactions on Graphics (TOG)*, 23(3):760–768, August 2004.
19. Kuang J., Johnson G., and Fairchild M. iCAM06: A refined image appearance model for hdr image rendering. *J. Visual Communication and Image Representation (JVCI)*, 18(5):406–414, October 2007.
20. Kuang J., Yamaguchi H., Liu C., Johnson G., and Fairchild M. Evaluating hdr rendering algorithms. *ACM Transactions on Applied Perception (TAP)*, 4(2), Article No. 9,July 2007.
21. Tumblin J., Hodgins J., and Guenter B. Two methods for display of high contrast images. *ACM Transactions on Graphics (TOG)*, 18(1):56–94, January 1999.

22. Chiu K., Herf M., Shirley P., Swamy S., Wang C., and Zimmerman K. Spatially nonuniform scaling functions for high contrast images. In *Proceedings of Graphics Interface*, pages 245–253, 1993.
23. Ashikhmin M. A tone mapping algorithm for high contrast images. In *Proceedings of the 13th Eurographics workshop on Rendering (EGRW)*, pages 145–156, 2002.
24. Ashikhmin M. and Goyal J. A reality check for tone-mapping operators. *ACM Transactions on Applied Perception (TAP)*, 3(4):399–411, October 2006.
25. Cadik M., Wimmer M., Neumann L., and Artusi A. Evaluation of HDR tone mapping methods using essential perceptual attributes. *Computers & Graphics*, 32(3):330–349, June 2008.
26. Narwaria M., Silva M., Callet P., and Pepion R. Effect of tone mapping operators on visual attention deployment. In *Proceedings of SPIE 8499, Applications of Digital Image Processing XXXV*, San Diego, California, USA, 2012.
27. Narwaria M., Silva M., Callet P., and Pepion R. Tone mapping-based high-dynamic-range image compression: study of optimization criterion and perceptual quality. *Optical Engineering*, 52(10):102008–102008, 2013.
28. Narwaria M., Silva M., Callet P., and Pepion R. Impact of tone mapping in high dynamic range image compression. In *Proceedings of Eighth International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, Chandler, Arizona, USA, 2014.
29. Narwaria M., Silva M., Callet P., and Pepion R. Tone mapping based hdr compression: Does it affect visual experience? *Signal Processing: Image Communication*, 29(2):257–273, February 2014.
30. Sugiyama N., Kaida H., Xue X., Jinno T., Adami N., and Okuda M. HDR compression using optimized tone mapping model. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1001–1004, 2009.
31. Ledda P., Chalmers A., Troscianko T., and Seetzen H. Evaluation of tone mapping operators using a high dynamic range display. *ACM Transactions on Graphics (TOG)*, 24(3):640–648, July 2005.
32. Fattal R., Lischinski D., and Werman M. Gradient domain high dynamic range compression. *ACM Transactions on Graphics (TOG)*, 21(3):249–256, July 2002.
33. Mantiuk R., Efremov A., Myszkowski K., and Seidel H. Backward compatible high dynamic range MPEG video compression. *ACM Transactions on Graphics (TOG)*, 25(3):713–723, July 2006.
34. Mantiuk R., Jim K., Rempel A., and Heidrich W. HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Transactions on Graphics (TOG)*, 30(4), July 2011.
35. Mantiuk R., Myszkowski K., and Seidel H. A perceptual framework for contrast processing of high dynamic range images. *ACM Transactions on Applied Perception (TAP)*, 3(3):267–276, July 2006.
36. Rensink R. Visual attention. In *Encyclopedia of Cognitive Science*. Nature Publishing Group, London, 2003.
37. Mann S. and Picard R. Being ‘undigitial’ with digital cameras: Extending dynamic range by combining differently exposed pictures. In *Proceedings of IS&T 48th Annual Conference*, pages 422–428. Society for Imaging Science and Technology, 1995.
38. Aydin T., Mantiuk R., and Seidel H. Extending quality metrics to full luminance range images. In *Proceedings of SPIE Human Vision & Electronic Imaging XIII*, pages 68060B–68060B–10, San Jose, CA, USA, 2008.

Chapter 6
Recent Advances of Quality Assessment for Medical Imaging Systems and Medical Images

Du-Yih Tsai and Eri Matsuyama

6.1 Introduction

Medical image quality assessment plays an important role in the design and manufacturing processes of image acquisition and processing systems, including image detectors such as imaging plates or flat panels. It is also critical for comparing and optimizing such as X-ray tube voltage and tube current. Because of this, a variety of research groups have been endeavoring to establish image quality standards and develop quality assessment methods.

In medical imaging, image quality is governed by a variety of factors such as contrast, resolution (sharpness), noise, artifacts, and distortion. Of these factors, resolution and noise are the most commonly used physical characteristics. The resolution and noise properties of imaging systems are described by the modulation transfer function (MTF) [1] and noise power spectrum (NPS) [2], respectively. The MTF is a common metric quantifying the resolution of the reconstructed images. It also describes the ability of an imaging system to reproduce the frequency information contained in the incident X-ray signal. The NPS is a common metric describing the frequency content of the noise of imaging systems. However, one of the dilemmas in medical radiography is the extent to which these metrics affect image quality. In comparison of two imaging systems or image detectors, for example, an imaging system may only be superior in one metric while being inferior to another in the other metric. To deal with this issue, the detective quantum efficiency (DQE) [3] is used as a metric to describe the general quality of the imaging system. The DQE which reflects system efficiency when forming an image using a limited number of X-ray photons can be calculated if the MTF, NPS, and X-ray

D.-Y. Tsai (✉) • E. Matsuyama
Graduate School of Health Sciences, Niigata University,
Niigata Prefecture, Niigata 951-8518, Japan
e-mail: tsai@clg.niigata-u.ac.jp

C. Deng et al. (eds.), *Visual Signal Quality Assessment*,
DOI 10.1007/978-3-319-10368-6_6

photon fluence are known. These metrics are dealt with in the spatial frequency domain. Other than the abovementioned metrics, various quality measures such as signal-to-noise ratio (SNR), mean square error (MSE), peak signal-to-noise ratio (PSNR), contrast-to-noise ratio (CNR), and contrast improvement ratio (CIR) are also commonly used for quality assessment.

Recently, Tsai et al. reported an image quality metric, mutual information (MI) [4], based on Shannon's information theory for assessing overall image quality in medical imaging systems. They used MI to express the amount of information that an output image contains about an input object. The basic idea is that when the amount of the uncertainty associated with an object before and after imaging is reduced, the difference of the uncertainty is equal to the value of MI. The more the MI value provides, the better the image quality is. Therefore, the overall quality of an image can be quantitatively assessed by measuring MI.

In this chapter, we mainly focus on describing recent advances of quality assessment for medical imaging systems and medical images. Section 6.2 generally reviews conventional medical image quality metrics, i.e., MTF, NPS, SNR, DQE, PSNR, CNR, and CIR, followed by describing the recently proposed image quality metrics, MI. Section 6.3 provides two clinical applications of image quality assessment in mammogram enhancement and radiation dose reduction in digital radiography. Section 6.4 ends with a conclusion.

6.2 Representative Quality Metrics for Medical Imaging Systems and Medical Images

6.2.1 Conventional Image Quality Metrics

6.2.1.1 Modulation Transfer Function

The MTF is widely recognized as the most relevant metric of resolution performance in radiographic imaging [1]. The MTF describes the imaging system such as an imaging plate's or an image detector's ability to transfer the input signal modulation of a given spatial frequency to its output. The MTF of a radiographic system has been determined either by evaluating the response of the system to periodic patterns or by measuring the line spread function (LSF) of the system using a narrow slit from which the MTF is calculated by Fourier transformation [5, 6]. The use of a slit requires very precise fabrication and alignment of the device in the radiation beams. An alternative method for determining the MTF of a radiographic system is to measure its edge spread function (ESF) using an opaque object with a straight edge [6]. The edge technique is currently the most widespread approach to measure the MTF [7].

6.2.1.2 Noise Power Spectrum

The NPS is one of the most common metrics describing the noise properties of imaging systems. The NPS is defined as the variance per frequency bin of a stochastic signal in the spatial frequency domain [2]. In other words, the NPS provides a convenient description of the noise amplitude and texture observed in images obtained with a uniform field of radiation having a specific fluence and spectral quality [8]. The NPS is most commonly computed directly from the squared Fourier amplitude of two-dimensional uniform images. The measurement of the NPS is conceptually straightforward but difficult to carry out experimentally, and there has not been universal agreement on the best methods for these measurements. There are two principal difficulties in determining the best method for NPS analysis. The first difficulty is that only a limited amount of data is available for analysis. The second difficulty in making accurate NPS measurements is that practical data contains some static artifactual components in addition to the stochastic noise that one desires to measure [2]. The NPS is often used as an input to the computation of DQE of an imaging system.

6.2.1.3 Signal-to-Noise Ratio

SNR is a generalized, objective image quality metric for X-ray based medical imaging systems. The spatial frequency-dependent SNR describes the ratio between the signal and the noise detected in the X-ray image. Because X-ray follows Poisson statistics, the noise of incoming X-ray at the image detector input (X-ray flux at the entrance window) is described by the standard deviation of the average input quanta per unit area and is equal to the square root of the average input quanta per unit area. Since the signal is described by the average input quanta per unit area, the SNR at the detector input is therefore equal to the square root of the average input quanta per unit area. The input SNR increases with the increase of number of quanta, because the SNR grows as the square root of the number of quanta.

6.2.1.4 Detective Quantum Efficiency

The DQE is a spatial frequency based measurement of the ability of the imaging device to convert the spatial information contained in the incident X-ray fluence into useful image information [3, 9, 10]. It is defined as

$$DQE(u) = \frac{SNR^2(u)_{out}}{SNR^2(u)_{in}} \tag{6.1}$$

where $\mathrm{SNR}(u)_{\mathrm{out}}$ and $\mathrm{SNR}(u)_{\mathrm{in}}$ are the spatial frequency-dependent signal-to-noise ratios of the imaging device at the output and input, respectively. The DQE can be calculated using the following formula [10, 11].

$$DQE(u) = \frac{MTF^2(u)}{q \times NNPS(u)} \tag{6.2}$$

where q is the X-ray photon fluence density (mm^{-2}) used for the uniform exposure image, and NNPS is the normalized NPS. Detailed elaboration of the Eq. (6.2) is described in the literature [12]. For a perfect imaging detector, DQE can reach a maximum value of 1.0.

6.2.1.5 Peak Signal-to-Noise Ratio

PSNR is one of the simplest and widely used metrics in medical image analysis. The PSNR [13] in decibels is adopted for measuring the performance of denoising and is given by

$$PSNR = 10\log_{10} \frac{M \times N \times T^2}{\sum\limits_i \sum\limits_j \left[d\,(i, j) - d'\,(i, j)\right]^2} \tag{6.3}$$

where $M \times N$ is the size of the image, T is the maximum possible value that can be obtained by the image signal, $d(i, j)$ and $d'(i, j)$ are the pixel values of original and processed images, respectively. The higher the PSNR value, the better the performance of denoising is.

6.2.1.6 Contrast-to-Noise Ratio

The CNR [14, 15] is defined as

$$CNR = \frac{|\mu_d - \mu_u|}{\sqrt{0.5\,({\sigma_d}^2 + {\sigma_u}^2)}} \tag{6.4}$$

where μ_d and σ_d are the mean and standard deviation computed in a desired region of interest (ROI) in an image, respectively. μ_u and σ_u are the mean and the standard deviation computed in an undesired ROI such as background, respectively. CNR measurement is proportional to the medical image quality.

6.2.1.7 Contrast Improvement Ratio

The CIR [16] is a quantitative measurement of the contrast improvement and is defined as

$$CIR = \frac{\sum_i \sum_j \left| c\left(i,j\right) - c\prime\left(i,j\right)\right|^2}{\sum_i \sum_j c(i,j)^2} \tag{6.5}$$

where $c(i,j)$ and $c'(i,j)$ are the local contrast values of original and enhanced images, respectively. The local contrast $c(i,j)$ is defined by the difference of mean values in two rectangular windows centered on a pixel at the coordinate (i, j). In detail the $c(i,j)$ is given by

$$c\left(i,j\right) = \frac{\left|p\left(i,j\right) - a\left(i,j\right)\right|}{\left|p\left(i,j\right) + a\left(i,j\right)\right|} \tag{6.6}$$

where p and a are the average values of pixels within a 3×3 region and a 7×7 surrounding neighborhood, respectively. The greater the CIR value, the better the enhancement result is.

6.2.2 A Mutual Information-Based Quality Metric for Medical Imaging Systems

Mutual information (MI) is a concept from information theory. We briefly describe the MI as follows.

Given events $S_1, \ldots S_n$ occurring with probabilities $p(S_1), \ldots p(S_n)$, then the average uncertainty associated with each event is defined by the Shannon entropy as

$$H(S) = -\sum_{i=1}^{n} p\left(S_i\right) \cdot \log_2 p\left(S_i\right) \tag{6.7}$$

Considering x and y as two random variables corresponding to an input variable and an output variable, the entropy for the input and that for the output are denoted as $H(x)$ and $H(y)$, respectively. For this case the joint entropy, $H(x, y)$, is defined as

$$H\left(x,y\right) = H(x) + H_x(y) = H(y) + H_y(x) \tag{6.8}$$

where $H_x(\mathrm{y})$ and $H_y(x)$ are conditional entropies. They are the entropies of the output when the input is known and that of the input when the output is known, respectively. In this situation, we can compute MI as:

$$\begin{aligned} MI\left(x;y\right) &= H(x) - H_y(x) = H(y) - H_x(y) \\ &= H(x) + H(y) - H\left(x,y\right) \end{aligned} \tag{6.9}$$

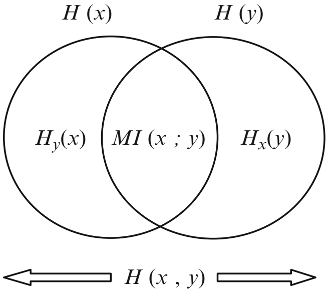


Fig. 6.1 Venn diagram which represents the relationship between input entropy $H(x)$ and output entropy $H(y)$, conditional entropies $H_x(y)$ and $H_y(x)$, joint entropy $H(x, y)$ and the mutual information MI(x; y)

Table 6.1 A data matrix of occurrence frequency for Y outputs to X inputs

	Input x						
Output y	x_1	x_2	…	x_i	…	X	Frequency
y_1	n_{11}	n_{21}	…	n_{i1}	…	n_{X1}	$n_{j=1}$
y_2	n_{12}	n_{22}	…	n_{i2}	…	n_{X2}	$n_{j=2}$
y_3	n_{13}	n_{23}	…	$n_{\mathrm{i}3}$	…	n_{X3}	$n_{j=3}$
…	…	…	…	…	…	…	…
y_j	n_{1j}	n_{2j}	…	n_{ij}	…	n_{Xj}	$n_{j=j}$
…	…	…	…	…	…	…	…
Y	n_{1Y}	n_{2Y}	…	n_{iY}	…	n_{XY}	$n_{j=Y}$
Frequency	$n_{i=1}$	$n_{i=2}$	…	$n_{i=i}$	…	$n_{i=X}$	n

A useful way of visualizing the relationship between these entropies is provided by a Venn diagram as shown in Fig. 6.1 [4].

Consider an experiment in which every input has a unique output belonging to one of various output categories. In this study, for simplicity, the inputs may be considered to be a set of subjects (e.g., phantoms in simplicity) varying in composition, while the outputs may be their corresponding images varying in optical density or gray level. An orderly system is employed in the present study to calculate the entropies of input, output, and their joint entropies. With this orderly system, the amount of MI is easily computed. The frequency with which each output is made to each input is recorded in Table 6.1 [4]. The columns and rows of this table represent various inputs and outputs. The various inputs, x_1, x_2, … x_i, … X, are assumed to take discrete values of input variables x. Likewise, the various outputs, y_1, y_2, … y_j, … Y are discrete values of output variables y. The uppercase X and Y stand for the number of input and output categories, respectively. The subscript i refers to any particular but unspecified input, whereas the subscript j refers to any particular but unspecified output. The number of times input x_i is presented will be symbolized by n_i, the frequency of output, y_j, by n_j, and the frequency, with which the input x_i corresponds to the output y_j, is given by n_{ij}. The total of all frequencies is given by n. It is apparent from Table 6.1 that

$$\sum_j n_{ij} = n_i \tag{6.10}$$

$$\sum_i n_{ij} = n_j \tag{6.11}$$

$$\sum_{ij} n_{ij} = \sum_i n_i = \sum_j n_j = n \tag{6.12}$$

Referring to the definition of information entropy as shown in Eq. (6.7), three informational quantities, namely $H(x)$, $H(y)$, and $H(x, y)$, can be calculated from Table 6.1.

$$H(x) = \sum_i p_i \log_2 \frac{1}{p_i} \tag{6.13}$$

$$H(y) = \sum_j p_j \log_2 \frac{1}{p_j} \tag{6.14}$$

$$H\,(x, y) = \sum_{ij} p_{ij} \log_2 \frac{1}{p_{ij}} \tag{6.15}$$

where $p_i = n_i/n$, $p_j = n_j/n$, and $p_{ij} = n_{ij}/n$. For simplicity, we can rewrite the above equations as follows:

$$H(x) = \log_2 n - \frac{1}{n}\sum_i n_i \log_2 n_i \tag{6.16}$$

$$H(y) = \log_2 n - \frac{1}{n}\sum_j n_j \log_2 n_j \tag{6.17}$$

$$H\,(x, y) = \log_2 n - \frac{1}{n}\sum_{ij} n_{ij} \log_2 n_{ij} \tag{6.18}$$

Then, the MI, $MI(x; y)$, can be obtained from Eq. (6.9) together with Eqs. (6.16)–(6.18). The MI conveys the amount of information that "y" has about "x."

Table 6.2 gives an example of how to calculate MI. Assume that a subject (e.g., a step-wedge) having five steps with different thickness was used for the experiment [4]. The five steps correspond to five inputs present equiprobably. The gray-scale pixel values of 100 pixels in each step after imaging were measured randomly. The distributions of the pixel values are considered as the corresponding

Table 6.2 An example of how to calculate the transmitted information

	Input x					
Output y	1	2	3	4	5	Frequency
1	20					20
2	60	4				64
3	20	88	10			118
4		8	76	14		98
5			12	80	2	94
6			2	6	8	16
7					90	90
Frequency	100	100	100	100	100	500

The frequencies shown in the table are referred to by means of the symbols given in Table 6.1, for example, $n_{23} = 88$, $n_{j=2} = 64$, $n_{i=1} = 100$, $n = 500$, and so on

outputs and their respective frequencies are given in the table. The frequencies will be referred to by means of the symbols given in Table 6.1; for example: $n_{12} = 60$, $n_{j=3} = 118$, $n_{i=2} = 100$, $n = 500$, and so on. Now, there are three information quantities, namely $H(x)$, $H(y)$, and $H(x, y)$, that can be calculated directly from Table 6.2 by using Eqs. (6.16)–(6.18).

For the data given in Table 6.2,

$$H(x) = \log_2 n - \frac{1}{n}\sum_i n_i \log_2 n = \log_2 5 = 2.323$$

(since inputs are equiprobable)

$$H(y) = \log_{2}500 - \frac{1}{500}\,(20\log_2 20 + 64\log_2 64 + 118\log_2 118 \cdots \text{etc}) = 2.575$$

$$H\,(x, y) = \log_2 500 - \frac{1}{500}\,(20\log_2 20 + 60\log_2 60 + 4\log_2 4 \cdots \text{etc} = 3.235$$

Applying Eq. (6.9) to the values calculated above, we have

$$MI\,(x; y) = H(x) + H(y) - H\,(x, y) = 2.323 + 2.575 - 3.235 = 1.663.$$

This is the estimate of the amount of information transmitted by the subject from input to output: 1.633 bits, out of a possible of 2.323 bits.

If the output is identical to the input, then knowing the output provides complete information about the input. In this case, MI is maximized and equal to the input entropy, and the uncertainty of the input is reduced to 0. It means that knowing (or viewing) the image of an object (subject) receives complete information about the object (subject). Thus, the quality of the obtained image arrives at a maximum in terms of the MI. If, on the other hand, the output and the input are independent, then knowing the output does not help making any conclusions about the input. In this case, the MI value is zero, and therefore the uncertainty about the input remains unchanged. This means that the obtained image has the lowest quality from the point of view of the MI.

6.3 Applications of Medical Image Quality Assessment

6.3.1 Improvement of Image Quality in Mammography Using a Wavelet Transform Based Approach

6.3.1.1 Background

Denoising and contrast enhancement operations are two of the most common and important techniques for medical image quality improvement. Because of their importance, there has been an enormous amount of research dedicated to the subject of noise removal and image enhancement [17–20].

With regard to image denoising, some approaches using discrete wavelet transform (DWT) have been proposed [21–23]. The DWT is very efficient from a computational point of view, but it is shift variant. Therefore, its denoising performance can change drastically if the starting position of the signal is shifted. In order to achieve shift-invariance, researchers have proposed the undecimated DWT (UDWT) [24–26]. Mencattini et al. reported a UDWT-based method for the reduction of noise in mammographic images [27]. Zhao et al. proposed an image denoising method based on Gaussian and non-Gaussian distribution assumptions for wavelet coefficients [28]. Huang et al. reported on a denoising method which involves directly selecting the thresholds for denoising by evaluating some statistical properties of the noise [29]. Recently, Matsuyama et al. proposed a modified UDWT approach for mammographic denoising [30]. The results demonstrated that the method could further improve image quality and decrease image processing time.

As regard to the improvement of contrast enhancement, various image enhancement techniques have been proposed [31–35]. These techniques can be divided into several categories, including histogram equalization, region-based, fuzzy, genetic-algorithm, and adaptive methodology. Wavelet-based approaches to enhancement of digital images have been also reported [36–40]. Tsai et al. proposed a method which employs an exponential-type mapping function to the wavelet coefficients of digital chest images and then reconstructs an enhanced image with the mapped wavelet coefficients [37, 38]. Lee et al. used a sigmoid-type mapping function for wavelet coefficient weighting adjustment to improve the contrast of medical images [40]. The method was applied to chest radiographs, mammograms, and chest CT images.

In this study, we expanded upon the previously suggested modified UDWT method [30] and combined it with the sigmoid-type mapping function [40]. By combining the two methods together in sequence, an effective algorithm for both image denoising and enhancement could be obtained. Original images were first

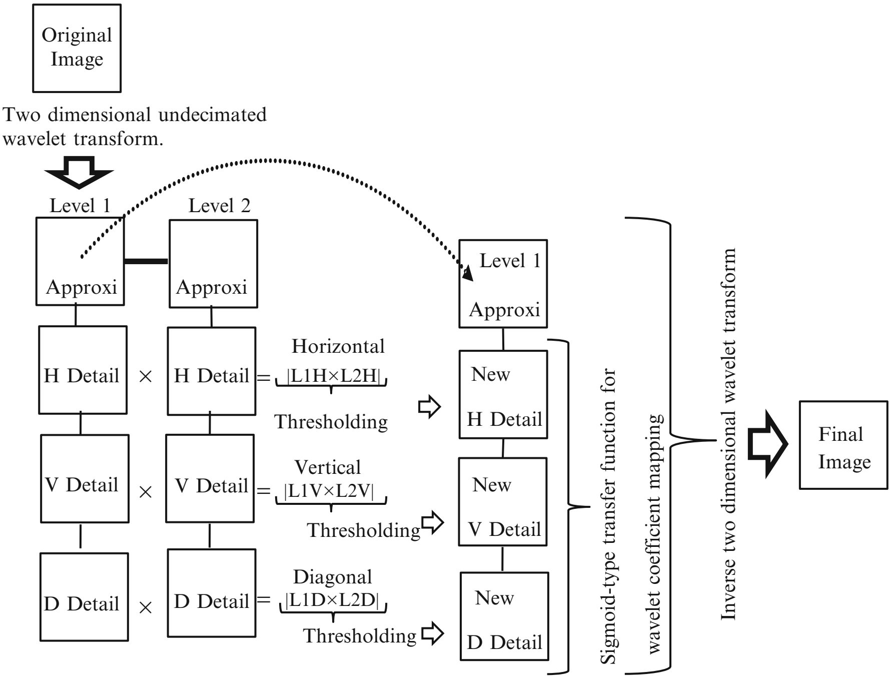


Fig. 6.2 A flow chart summarizing the processing procedure for the proposed algorithm

denoised using the modified UDWT, followed by image enhancement using the wavelet-coefficient mapping function. Finally, a denoised and contrast enhanced image was reconstructed by the inverse wavelet transform.

6.3.1.2 Methods and Materials

Proposed Method

Figure 6.2 shows the flowchart of our proposed method. In the first phase, denoising was applied to original images using our recently reported UDWT [30]. In the second phase, image enhancement was performed using a sigmoid-type transfer function for wavelet coefficient mapping [40].

The UDWT is a wavelet transform algorithm designed to overcome the lack of translation-invariance of the DWT. Unlike the DWT, the UDWT does not incorporate the down sampling operations. Thus, the approximation coefficients (low-frequency coefficients) and detailed coefficients (high-frequency coefficients) at each level are the same length as the original signal. The basic algorithm of the conventional UDWT is that it applies the transform at each point of the image and saves the detailed coefficients and uses the approximation coefficients for the next

Table 6.3 Comparison of three image quality measurements of six different wavelet basis functions for simulated images

	Wavelet basis function					
Image quality measurement	dmey	db2	sym7	coif1	coif5	bior6.8
MI (bit)	0.68	0.81	0.72	0.79	0.69	0.72
MSE	58.43	50.20	55.87	51.01	57.1	55.51
SNR (dB)	27.93	29.10	28.21	28.93	28.04	28.29

MI mutual information, MSE mean square error, SNR signal-to-noise ratio

level. The size of the coefficients array does not diminish from level to level. This decomposition operation is further iterated up to a higher level. There are major differences between the modified UDWT method [30] and the conventional UDWT method. First, the conventional UDWT decomposes the original image (level 0) into one low-frequency band and three high-frequency bands for each resolution level with the same size as the original image. The decompositions are usually conducted up to resolution level 4. In contrast, the modified UDWT method only needs to perform the computation up to resolution level 2 and repeat the computation only one time [30, 41]. Second, the conventional UDWT thresholded the detailed coefficients at all four levels with the same thresholding value, while the modified UDWT method utilizes the hierarchical correlation of the coefficients between level 1 and 2 of the three detailed coefficients for thresholding. In other words, the thresholding values vary and are dependent on the nature of the noise.

The extended UDWT method adopted in the present study was based on the modified UDWT [30].

We evaluated six comparatively popular wavelet basis functions, namely discrete finite impulse response (FIR) approximation of Meyer wavelet (dmey), Daubechies order 2 (db2), Symlets order 7 (sym7), Coiflets order 1 (coif1), Coiflets order 5 (coif5), and biorthogonal 6.8 (bior6.8), as candidates for selection as the most suitable basis function for the UDWT. The evaluation results showed that wavelet-processed images with db2 basis function provided the best results among the six basis functions. Thus, we selected db2 basis function for the proposed method [30, 40, 41]. Table 6.3 shows the results for simulated images processed by the six wavelet basis functions. Quality metrics used for assessment were MI, MSE, and SNR (dB).

A sigmoid-type transfer curve with a one-to-one mapping function was used for enhancement of image contrast [40]. The mapping function was determined based on the following considerations: (a) wavelet coefficients having high values are heavily weighted because they carry more useful information; (b) the coefficients at low levels are heavily weighted because they carry detailed information, such as edge information; and (c) the approximation coefficients are not manipulated to prevent image distortion [38, 40].

Image Data

To evaluate and validate our proposed method, we used a standard mammogram database. The database was from the Mammographic Image Analysis Society (MIAS) [42]. Patient informed consent was not required. A total of 30 mammograms obtained from the database were used for investigation of the effectiveness of the proposed method. The matrix size of each image was 1,024 × 1,024 pixels with 8-bit gray-level resolution.

Quantitative and Perceptual Evaluations

In order to compare objectively the performance of the proposed algorithm against two published algorithms [30, 40], in this study we adopted three image quality metrics. The three metrics are contrast-to-noise ratio (CNR), contrast improvement ratio (CIR), and peak signal-to-noise ratio (PSNR). A visual perceptual evaluation was also designed for performance analysis. We used Scheffe's method of paired comparison to evaluate the preference of overall image quality [43, 44].

The visual perceptual evaluation was made by five experienced radiological technologists (ranging from 20 to 25 years of experience). The obtained 30 mammograms from the database were processed using the proposed method, a modified UDWT method [30], and a sigmoid-type wavelet coefficient (STWC) mapping method [40]. Thus, a total of 90 images were used for image quality evaluation. All images were evaluated on a pair of widely used medical 3M monochrome liquid-crystal display monitors. Each observer reviewed the images independently. The reading time was limited to less than 20 s for each reading. The observers independently evaluated one pair of images, which were shown on the monitors one at a time, using a 5-point grading scale (−2 points to +2 points). If the image shown on the left was much better than that shown on the right in terms of overall image quality, the left image was given +2 points; the left image was given +1 point when it was slightly better than the right one; the left image was given 0 points, when both images were of the same image quality. Conversely, if the image shown on the left was much poorer than that shown on the right in terms of overall image quality, the left image was given −2 points; the left image was given −1 point when it was slightly poorer than the right one. Comparisons were made by use of three possible combinations, that is, modified UDWT/sigmoid mapping, modified UDWT/proposed method, and sigmoid mapping/proposed method combinations. Each pair of images was determined randomly. In addition, the two paired images (left side vs. right side) were arranged on a random basis.

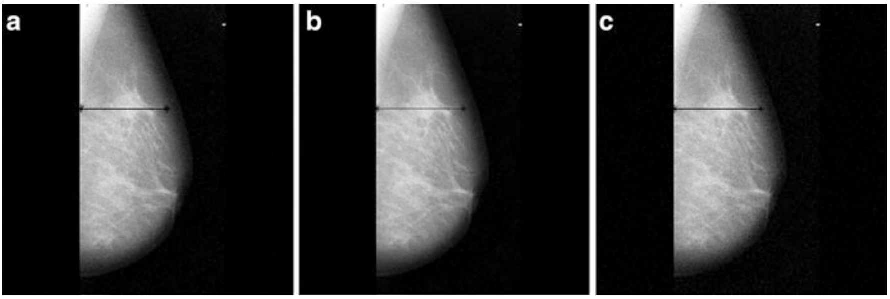


Fig. 6.3 Image processing results for mammograms. (**a**) Image processed by the proposed method, (**b**) image processed by the modified UDWT method, and (**c**) image processed by the sigmoid-type wavelet coefficient mapping method

Table 6.4 Comparison of image processing methods in terms of three quantitative quality metrics for mammograms

Method	CNR	CIR	PSNR
UDWT	8.18	0.28	38.35
Sigmoid	7.64	0.29	36.39
Proposed	8.24	0.67	37.98

6.3.1.3 Results and Discussion

Figure 6.3 illustrates an example of image processing results obtained from the mammograms. Figure 6.3a–c are resulting images processed by using the proposed method, the modified UDWT method, and the STWC mapping method, respectively.

Table 6.4 summarizes the quantitative evaluation results for the proposed method and two published methods in terms of CNR, CIR, and PSNR metrics. As described earlier that the CNR measurement is proportional to the medical image quality. It is obvious from the table that CNR value of the image processed by the proposed method gave the best result as compared to those processed by other two methods. The CIR is a metric used for evaluating the contrast improvement. It is noted from the table that the proposed method shows the greatest value, followed by the sigmoid mapping and modified UDWT. The reason why the proposed method is superior to the sigmoid mapping method is due to the fact that the images processed by the proposed method have been denoised prior to mapping operation. In the case of PSNR measurement, the results listed in Table 6.4 show that the modified UDWT method was slightly better than both the proposed method and sigmoid mapping method from the point of view of denoising performance. The reason might be because some residual (un-removed) noise has also enhanced during enhancement operation. This results in the decrease of PSNR value. However, the images processed by the proposed method showed the best overall image quality in terms of both denoising and contrast enhancement when looking into the values of the PSNR and CNR as shown in Table 6.4.

Table 6.5 Results of mammogram scoring for the three combinations by the five observers

		Observer					
Combination		a	b	c	d	e	Sum
Sigmoid	UDWT	−1.1	−0.87	0	−1.2	−1.2	−4.37
Sigmoid	Proposed	−1.57	−1.4	−1.67	−1.47	−1.6	−7.71
UDWT	Proposed	−1.33	−1.27	−1.47	−1.3	−1.5	−6.87

The results of scoring for the three combinations by the five observers are listed in Table 6.5. As described earlier, if the left image of the paired images (two-image combination) was poorer than the right image in terms of overall image quality, it received a negative score. As indicated by the preference scores shown in the rightmost column of the table, the images processed by the proposed method were judged to have the best quality. The visual evaluation results showed that the images processed by the proposed method were judged to have the best quality as compared to the other two methods.

6.3.1.4 Summary

We proposed a method for improving image quality in mammography by using a wavelet-based approach. The proposed method integrated two components: image denoising and image enhancement. In the first component, a modified UDWT was used to eliminate the noise. In the second component, a wavelet-coefficient mapping function was applied to enhance the contrast of denoised images obtained from the first component. We examined the performance of the proposed method by comparing it with two previously reported methods. The results of quantitative assessment showed that the proposed UDWT method outperformed over the other two methods. The results of visual assessment also indicated that the images processed by the proposed UDWT method showed statistically significantly superior image quality over the other two methods. Our research results demonstrated the superiority and effectiveness of the proposed method. This methodology can be used not only as a means for improving visual quality of medical images but also as a preprocessing module for computer-aided detection/diagnosis systems to improve the performance of screening and detecting regions of interest in images.

6.3.2 The Effect of Radiation Dose Reduction on Image Quality in Digital Radiography

6.3.2.1 Background

The issue of radiation dose exposure to patients from digital radiography is a major public health concern. In particular, it is important to keep radiation dose exposure to a minimum in female patients during their reproductive period, who frequently undergo repeated radiation exposure during the course of diagnostic imaging and treatment follow-up.

It is known that a trade-off exists between noise level and radiation dose. On the one hand, high-dose radiation will lower the noise level, but may expose the patient to excessive radiation. On the other hand, low-dose radiation will lower the SNR of the image and result in reducing the amount of image information. The balancing of radiation dose and image quality should be performed precisely to ensure that patient doses are kept at a reasonable minimum, while maintaining clinically acceptable image quality. To address this issue, much research, including the development of new detectors and image processing methods [45–47], has been carried out. In recent years, several investigators have reported that wavelet-based image processing techniques are effective in the reduction of radiation dose [48–55].

Conventional radiography is widely used for the pelvis and lumbar spine. However, the radiation dose for pelvic and lumbar X-ray examinations using a radiograph is relatively high in order to obtain acceptable image quality. An effort to reduce the exposure dose can have a positive effect on a patient's quality of life.

In this study, we propose an improved wavelet-transform-based method for potentially reducing the radiation dose while maintaining clinically acceptable image quality. The proposed method integrates the advantages of our previously proposed wavelet-coefficient-weighted method [38, 40, 54] and the existing BayesShrink thresholding method [56]. The wavelet-coefficient-weighted method has the advantage of effective edge enhancement accompanied by a slight suppression of noise increase (however, the noise is also enhanced definitely). In contrast, the BayesShrink thresholding method is used for denoising, while the edge is preserved as much as possible (however, the edge also decreases definitely). It is expected that our approach, integrating the two methods, can achieve edge enhancement with almost no significant noise increasing and can be applied to low-dose radiographs to improve image quality. To verify the proposed method's effectiveness in reducing radiation dose in digital radiography, a quantitative and qualitative assessment was performed.

6.3.2.2 Methods and Materials

Proposed Method

The main steps of the proposed method include a previously reported wavelet coefficient adjustment technique for contrast enhancement [38, 40, 54] and a wavelet thresholding technique for noise reduction. Figure 6.4 shows a schematic diagram of the proposed method. As shown in the figure, the proposed method for denoising radiographic images starts by decomposition of the original image by use of the DWT, which results in obtaining different detail wavelet coefficients (horizontal, vertical, diagonal). The three detail coefficients are then processed by use of a sigmoid-type transfer curve for adjustment of wavelet coefficient, followed by BayesShrink thresholding.

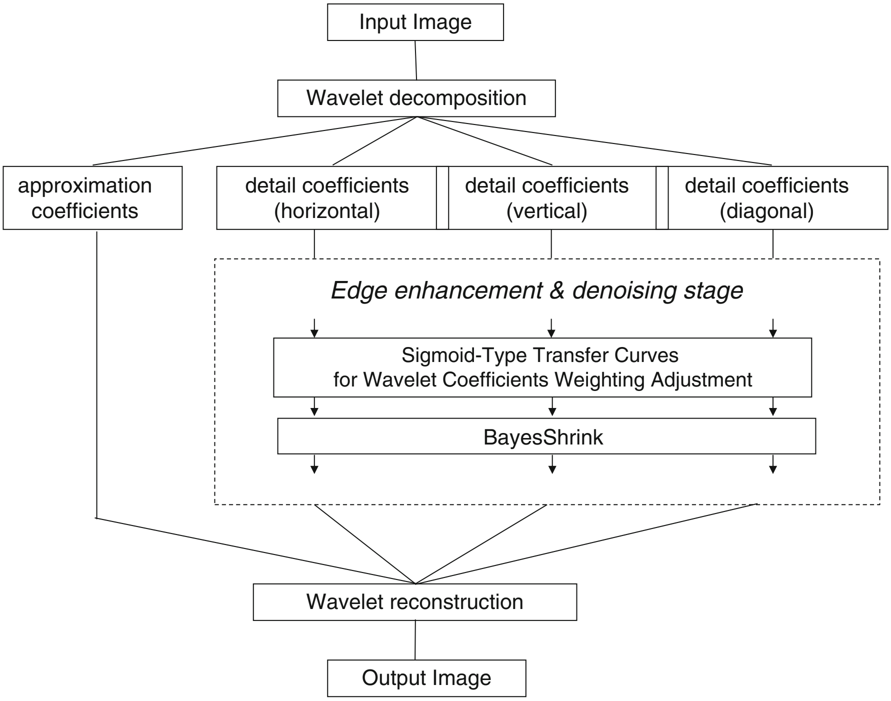

Input Image
Wavelet decomposition
approximation coefficients
detail coefficients (horizontal)
detail coefficients (vertical)
detail coefficients (diagonal)
Edge enhancement & denoising stage
Sigmoid-Type Transfer Curves for Wavelet Coefficients Weighting Adjustment
BayesShrink
Wavelet reconstruction
Output Image

Fig. 6.4 Flow chart of the proposed method

A sigmoid-type transfer curve with a one-to-one mapping function is used for enhancement of image contrast. The mapping function was determined based on the following considerations: (a) in the case of detail components at a specific level, high-value coefficients are weighted because they carry effective information; (b) the coefficients at low levels are heavily weighted, because they carry detailed information, such as edge information; and (c) the approximation coefficients are not manipulated in order to prevent image distortion. A detailed explanation of the sigmoid-type transfer curve for wavelet coefficient weighting adjustment is given in the literature [57].

Wavelet denoising attempts to remove the noise present in an image while preserving the image characteristics. Wavelet thresholding, first proposed by Donoho [48], is a signal-estimation technique that exploits the capabilities of the wavelet transform for signal and image denoising. It removes noise by eliminating coefficients that are insignificant relative to some threshold. Therefore, the selection of the threshold is the most important step in wavelet-based denoising techniques.

Various threshold selection methods have been proposed, such as VisuShrink [58], SureShrink [59], and BayesShrink [56]. In the VisuShrink method, a universal threshold that is a function of the noise variance and the number of samples is developed based on the minimax error measure. The threshold value in the

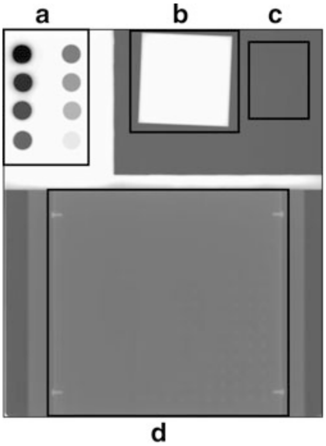


Fig. 6.5 Examples of phantom images used for the measurement of physical characteristics. (**a**) Contrast detail curve, (**b**) MTF, (**c**) NPS, and (**d**) GLC using acrylic disk on Burger phantom

SureShrink method is optimal in terms of the Stein's unbiased risk estimator. The BayesShrink method determines the threshold value in a Bayesian framework, through modeling of the distribution of the wavelet coefficients as Gaussian [60]. Several researchers have compared the three selection methods, and their results have shown that BayesShrink outperforms the other two methods [60–62]. In this study, we employed the BayesShrink method for denoising.

Data Acquisition

Images that were used for measurement of physical characteristics were acquired with use of a multipurpose phantom [63]. Figure 6.5 shows an example of phantom images. A computed radiography (CR) system and an imaging plate were used in this study. A pixel size of 0.1 mm and a quantization level of 10 bits were employed for data acquisition. The system parameter settings for the latitude and sensitivity were fixed at 3 and 200, respectively. Images were taken with a radiation quality of RQA-5 (HVL = 7.1 mm Al, 21 mm Al additional filtration) by using a tungsten target X-ray tube (Hitachi, Tokyo, Japan). The focal spot size of the X-ray tube was 0.6 mm. The source-to-image receptor distance was 190 cm. The amount of exposure was 4.63×10^{-7} C/kg (50 mAs). Twenty phantom images were obtained and used for measuring the presampled MTF, NPS, and gray-level contrast (GLC).

Four different radiation levels were used for investigating the effect of the physical characteristics on the radiation dose. The four radiation level ratios with respect to the reference level, 4.63×10^{-7} C/kg, were 50/100, 64/100, 80/100, and 100/100. A visual evaluation of wavelet-processed images of a human body phantom was performed to confirm the effectiveness of the proposed method in reducing radiation dose. An anterior–posterior (AP) projection of the hip joint and a lateral view of the lumbar spine on the human body phantom were exposed to various dose levels. These two images were also taken at four different radiation level ratios, 50/100, 64/100, 80/100, and 100/100, instead of the reference level that is commonly used in clinical radiology practice. In this study, the hip joint phantom was exposed at 70 kVp and 32 mAs, and the lumbar phantom at 82 kV and 64 mAs.

Measurement of Quality Metrics

The presampled MTFs were measured with an angled-edge method [6]. The edge device was made of a 1-mm-thick sharp-edged tungsten plate whose dimensions were $10 \times 10\ \text{cm}^2$. The direction of the edge was oriented with a small angle (2°–3°). The ESF in the direction perpendicular to the edge was then obtained. To reduce the noise in the edge profile, 20 representations of the sampled ESFs were generated from the ROI. Then the ESFs were differentiated to obtain the LSFs, and the presampled MTFs were deduced by applying Fourier transformation to the LSFs [6, 8]. The resulting MTF was obtained by averaging the 20 MTFs.

NPS measurements were made by exposure of the imaging plate to a uniform beam of radiation. For determination of the NPS, a two-dimensional second-order polynomial was fitted and subtracted to remove background trends. For the calculation, the central portion of each uniform image obtained was divided into four non-overlapping regions, 256×256 in size. A total of 80 regions were used. The NPS was calculated by applying the fast Fourier transform to each of the ROIs and then averaging the resulting spectrum estimates [3, 64].

A commercially available Burger phantom (Kyoto Kagaku, Kyoto, Japan) was employed for measurement of GLC characteristics. In this study, the GLC was used to describe the relative contrast of an image, defined by

$$GLC = \frac{\left| L_{acrylic} - L_{BG} \right|}{L_D - 1} \tag{6.19}$$

where $L_{acrylic}$, L_{BG}, and L_D represent the mean pixel value of an 8.0 mm diameter circle of an acrylic disk 8.0 mm in thickness, the mean pixel value of the background, and the gray level of the CR, respectively. The GLC value ranged from 0 to 1.0. Image contrasts with different gray levels could be compared because the GLC was normalized by $(L_D - 1)$. Low GLC corresponded to low contrast, while high GLC corresponded to high contrast. For clarifying the effect of the radiation dose on the GLC, dose ratios ranging between 100/100 and 50/100, instead of the standard dose, were measured.

Performance Comparison

In order to validate the superiority and effectiveness of the proposed method, we compared the proposed method with three conventional methods, namely the Wiener filter (WF) [65, 66], BayesShrink method [56], and sigmoid-type method [40]. The proposed method and the above-described three methods were applied to the original images for performance comparison.

Visual Evaluation

A visual evaluation was conducted by five experienced radiological technologists. The images were displayed on a liquid-crystal display (1,280 × 1,024 matrix, LCD-1980SXi, Nippon Electric Company, Tokyo, Japan). The parameters of window level, window width, and display image size on the image display apparatus were fixed. Each observer reviewed the images independently. The reading time was not limited. The five radiological technologists independently evaluated the total depiction of each phantom image for diagnostic acceptability by using a 5-point grading scale [(1) no depiction; (2) faint; (3) acceptable; (4) good; (5) excellent]. Statistical analyses were performed with the Friedman test. When a statistically significant difference was found ($p < 0.01$) in the five images (the original and the four image-processed images) at each dose ratio, pairwise comparisons were performed with Scheffe's method. Comparisons were made by use of five possible combinations, namely WF-processed image, BayesShrink-processed image, sigmoid-processed image, proposed filter-processed image, and the original image.

6.3.2.3 Results and Discussion

Figure 6.6a shows the MTF values for the original image and the four processed images. The MTF value for the sigmoid image was the highest, followed by that for the proposed image. Both MTFs were considerably superior to the original image over the entire spatial frequency range. In contrast, the MTF values obtained from the BayesShrink and the WF images were slightly lower than that of the original image. Figure 6.6b shows the NPS values. The NPS values for the sigmoid image were pronouncedly higher than those of the original image. The NPS values for the proposed image were slightly higher or similar to those of the original image. In contrast, the NPS values for the BayesShrink and WF images were lower than those for the original image.

Figure 6.7a shows the GLC as a function of the radiation dose ratio. There were no significant differences in any of the GLCs. Figure 6.7b, c, respectively, show the NPS as a function of the dose ratio at spatial frequencies of 1 and 4 mm^{-1} for the original image and the four processed images. Although the trend of the values measured from the proposed-method-processed image is similar to that measured from the original image in dose ratios ranging from 80/100 to 50/100, the NPS for the proposed method showed improvement in noise level at the dose ratio of 50/100 at the spatial frequency of 4 mm^{-1}.

Figure 6.8 illustrates the mean grades of visual evaluation for the hip joint and the lumbar spine at the four radiation dose ratios. In all cases, significant differences ($p < 0.01$) were found with the Friedman test at various radiation dose ratios. For the hip joint, the mean grade of the proposed method reached three points and was higher than those of the other methods tested at all radiation dose ratios. In the case of the lumbar spine, except for the radiation dose ratio of 50/100, the mean grade for the proposed method was higher than those of the other methods tested.

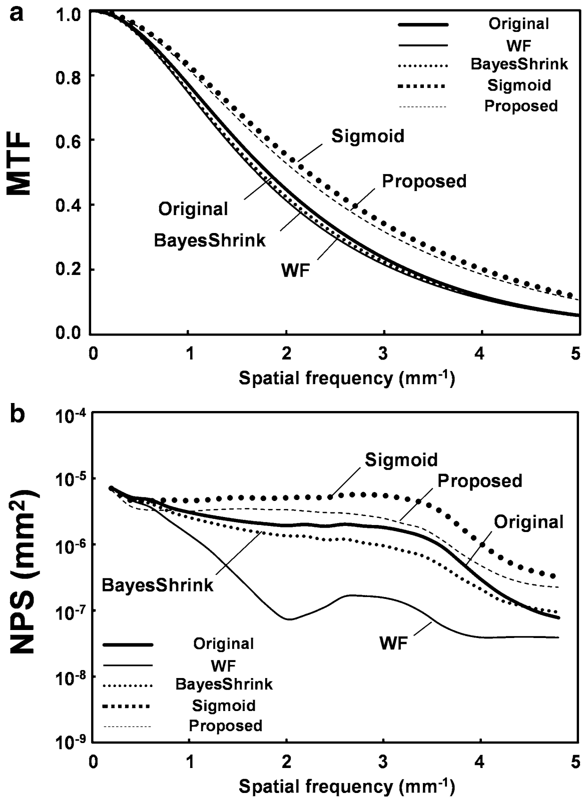


Fig. 6.6 (**a**) Presampled MTFs and (**b**) NPS for the original image and the four processed images

Figure 6.9 illustrates visual evaluation results for the hip joint (AP) and lumbar spine (lateral) at various radiation dose ratios by use of Scheffe's method of paired comparisons. As shown in Fig. 6.9, original image (Org) at 100/100 radiation dose ratio is located at the center (marked as zero) of the straight horizontal bar. The quality of the processed image was superior to that of the Org, if it had a higher score than the Org and there was a statistical significance. The quality of the processed image was inferior to that of the Org, if it had a lower score than the Org and there was a statistical significance. If the processed image had a similar score to that of the Org, the quality was considered to be equivalent to that of the Org. In terms of diagnostic acceptability, the proposed method provided significantly better results than those of the original image up to a 64/100 radiation dose ratio in the hip joint. When the radiation dose ratio was 50/100, no significant difference was found between the image processed by the proposed method and the original image.

In the case of lumbar radiographs, the results obtained from the proposed method were comparable to those of the original image up to a 64/100 radiation dose ratio. However, the proposed method tended to show unsatisfactory results for a 50/100 radiation dose ratio.

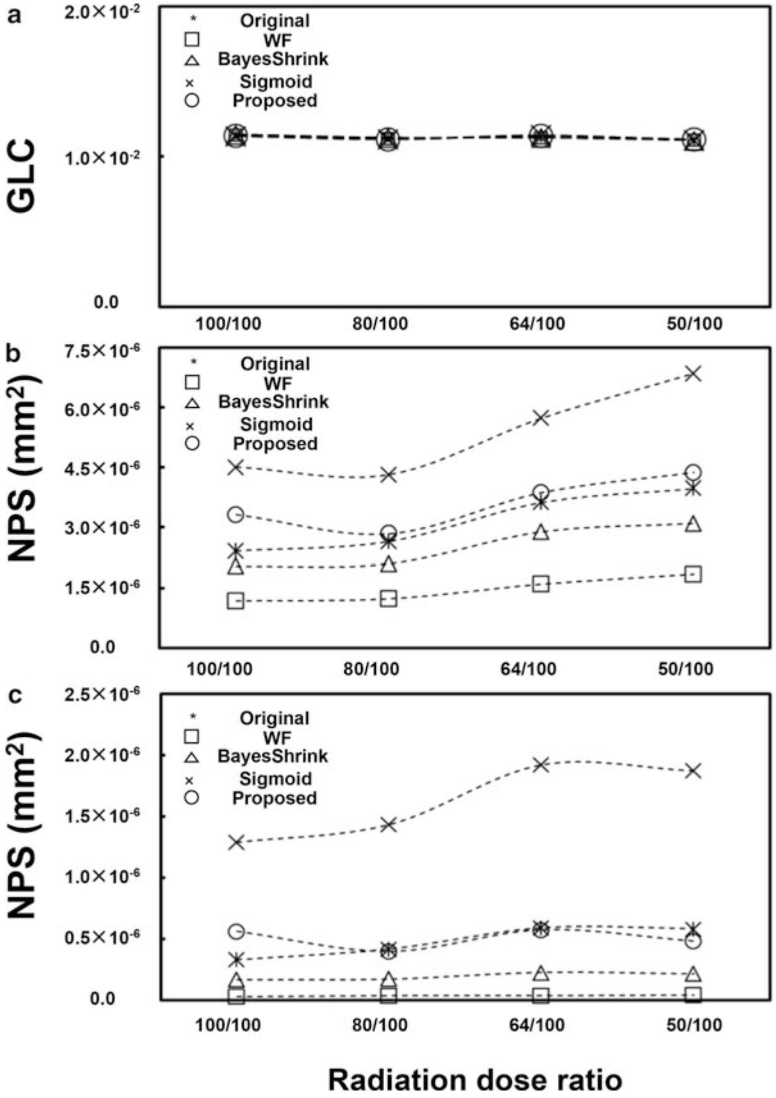


Fig. 6.7 (**a**) Gray-level contrasts as a function of the dose ratio for the original image and the four processed images obtained using a Burger phantom. (**b**) NPSs as a function of the radiation dose ratio, measured from the original image and the various images processed by the WF, BayesShrink, sigmoid-type, and the proposed methods at spatial frequency of 1 mm^{-1}. (**c**) The NPSs at spatial frequency of 4 mm^{-1}

Figure 6.10 illustrates original and processed images of the hip joint and lumbar spine of the human body phantom, which were used for the visual evaluation.

The proposed method provides the benefits of improved resolution and noise suppression. The experimental results demonstrated the method's effectiveness in

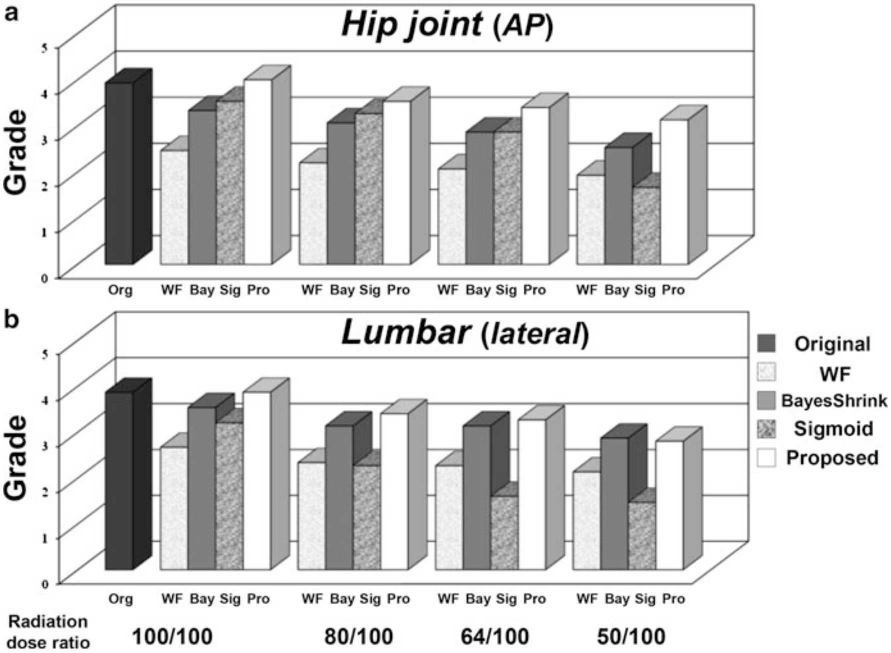


Fig. 6.8 Mean grades of visual evaluation of the original and the four processed images for the hip joint and lumbar spine at the four radiation dose ratios, 100/100, 80/100, 64/100, and 50/100, in comparison with the standard dose. (**a**) Hip joint (AP). (**b**) Lumbar spine (lateral)

dose reduction without degradation in image quality at a lower dose as compared to the standard dose. In the MTF and NPS measurements, the physical properties of the images processed by use of the sigmoid-function and BayesShrink methods show distinct differences. The sigmoid function yields improved spatial resolution characteristics with increasing noise. In contrast, the BayesShrink method provides improved noise properties, but deteriorating spatial resolution. The proposed method incorporates the sigmoid method into the BayesShrink algorithm. As a result, the proposed method shows better spatial resolution and noise properties compared to the original image.

In the GLC measurements, there were almost no differences in any of the GLCs. This implies that contrast was independent of X-ray dose and the proposed method did not contribute to contrast enhancement in the GLC experiment. The NPS values of our proposed method were near to those of the original image at all radiation dose ratios except 100/100. The high NPS value of our method at the 100/100 radiation dose ratio might be due to the fact that some noise enhanced by the sigmoid-function method was not recognized as noise in the BayesShrink method, and thus did not decrease efficiently.

As shown in Figs. 6.8 and 6.9, the results of our study indicate that the proposed strategy significantly improves the quality of low-dose images such that CR images

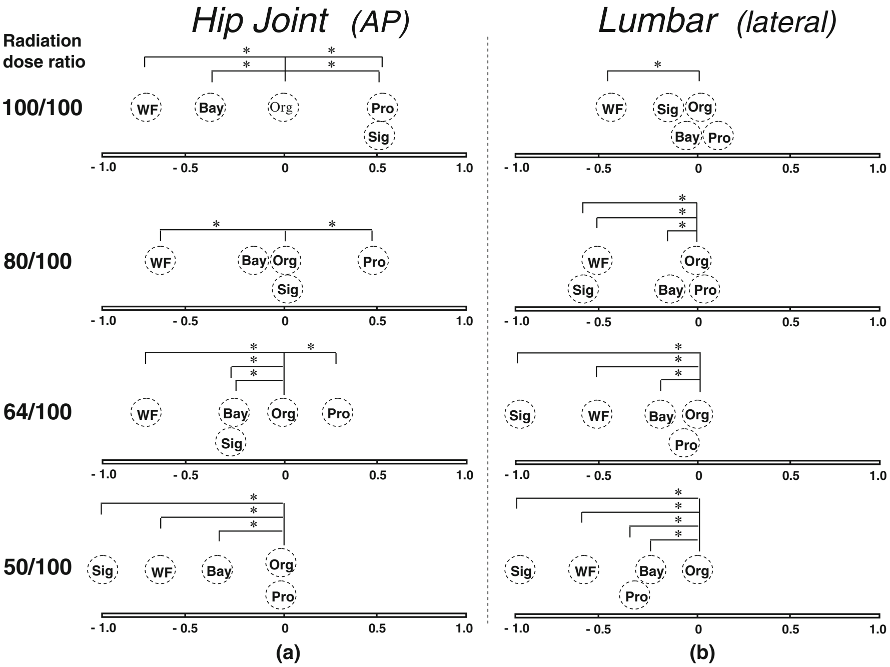


Fig. 6.9 Visual evaluation results using Scheffe's method of paired comparisons of the original image (Org) at the radiation dose ratio of 100/100 and various images processed by the WF, BayesShrink (Bay), sigmoid (Sig), and the proposed (Pro) methods at radiation dose ratios of 100/100, 80/100, 64/100, and 50/100 in comparison with the standard dose. (**a**) Hip joint (AP). (**b**) Lumbar spine (lateral). There was a significant difference ($p < 0.01$) between the original image and the processed image at various dose ratios if the *asterisk* mark is shown

obtained at 50 and 64 % of the standard dose level provide clear depiction in AP views of the hip joint and in lateral views of the lumbar spine, respectively, in terms of visual evaluation. In Fig. 6.10, the visibility of the overall appearance of bones seems to be improved by the proposed method. This may be due to the improvement in resolution and the suppression of noise. Maintaining a well-balanced relationship among contrast, spatial resolution, and noise is important. From this point of view, the proposed method has a well-balanced filter for the AP view of the hip joint and the lateral view of the lumbar spine at a lower dose.

6.3.2.4 Summary

We proposed an improved wavelet-transform-based method for potentially reducing radiation dose while maintaining clinically acceptable image quality. The effectiveness of the proposed method was demonstrated quantitatively and qualitatively. The experimental results showed that the proposed method could

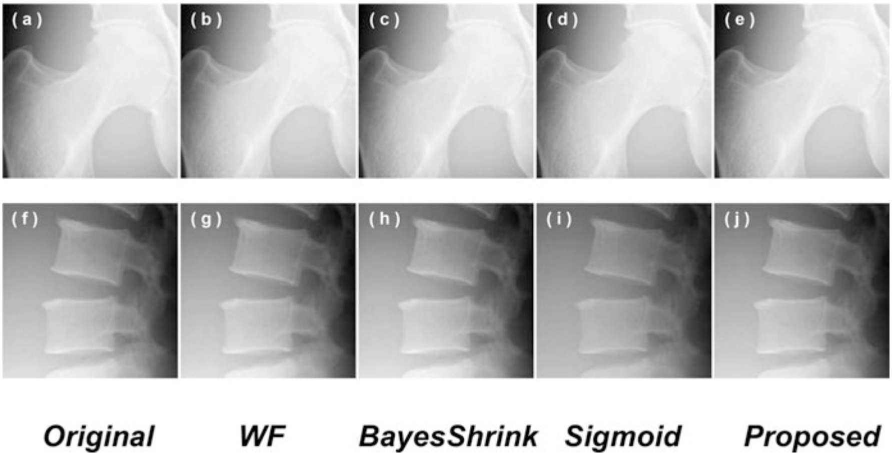


Fig. 6.10 Original and processed phantom images obtained using WF, BayesShrink, sigmoid, and the proposed methods. (**a**–**e**) Hip joint at the standard dose. (**f**–**j**) Lumbar spine at the standard dose

improve resolution while keeping noise level within acceptable limits. Furthermore, the results validated the effectiveness of our proposed method in the reduction of radiation dose. Our visual evaluation showed that an approximately 40–50 % reduction in the exposure dose might be achieved with the proposed method in AP views of hip joint radiographs and lateral views of lumbar spine radiographs. The proposed method has the potential to improve visibility in radiographs when a lower radiation dose is applied.

6.4 Conclusion

In this chapter, we described the importance of medical image quality assessment and major factors used for characterizing physical properties of medical images. We also described the recent trends in assessment of medical imaging systems and medical images. We first briefly reviewed conventional medical image quality metrics and then described a recently proposed image quality metric, mutual information. We also provided two clinical applications, i.e., mammographic image quality enhancement and effect of radiation dose reduction on image quality in digital radiography, to address the importance of quality assessment for medical images.

Acknowledgments The authors would like to thank Dr. Haruyuki Watanabe for his contribution to the data acquisition experiments and cooperation in dose reduction studies. This research was supported in part by a Grant-in-Aid for Scientific Research (23602004) from the Japan Society for the Promotion of Sciences (JSPS).

References

1. Samei E, Ranger NT, Dobbins III JT et al: Intercomparison of methods for image characterization. 1. modulation transfer function. Med Phys 33: 1454–1465, 2006
2. Dobbin III JT, Samei E, Ranger NT et al: Intercomparison of methods for image quality characterization. II. noise power spectrum. Med Phys 33: 1466–1475, 2006
3. Neitzel U, Gunther-Kohfahl S, Borasi E et al: Determination of the detective quantum efficiency of a digital X-ray detector: comparison of three evaluations using a common image data set. Med Phys 31: 2205–2211, 2004
4. Tsai DY, Lee Y, Matsuyama E: Information entropy measure for evaluation of image quality. J Digit Imaging, 21: 338–347, 2008
5. Fujita H, Tsai DY, Itoh K et al: A simple method for determining the modulation transfer function in digital radiography. IEEE Trans Med Imag 11: 34–39, 1992
6. Samei E, Flynn MJ, Reimann DA: A method for measuring the presampling MTF of digital radiographic systems using an edge test device. Med Phys 25: 102–113, 1998
7. Samei E, Buhr E, Granfors P et al: Comparison of edge analysis techniques for the determination of the MTF of digital radiographic systems. Phys Med Biol 50: 3613–3625, 2005
8. Flynn MJ, Samei E: Experimental comparison of noise and resolution for 2k and 4k storage phosphor radiography systems. Med Phys 26: 1612–1623, 1999
9. Spahn M: Flat detectors and their clinical applications. Eur Radiol 15: 1934–1947, 2005
10. Fetterly KA, Hangiandreou NJ: Effect of X-ray spectra on the DQE of a computed radiography system. Med Phys 28: 241–249, 2001
11. Fettery KA, Schueler BA: Performance evaluation of a computed radiography imaging device using a typical front side and novel dual side readout storage phosphors. Med Phys 33: 290–296, 2006
12. Cunningham IA: Applied linear-system theory. In: Beutel J, Kundel HL, VanMetter RL (ed) Handbook of medical imaging, vol 1, Physics and psychophysics: SPIE Press, Bellingham, WA, 2000, pp. 79–159
13. Luisier F, Blu T, Unser M: A new SURE approach to image denoising: interscale orthonormal wavelet thresholding. IEEE Trans Image Process 16: 593–606, 2007.
14. Bao P, Zhang L: Noise reduction for magnetic resonance images via adaptive multiscale products thresholding. IEEE Trans Med Imag 22: 1089–1099, 2003
15. Cincotti G, Loi G, Pappalardo M: Frequency decomposition and compounding of ultrasound medical images with wavelet packets. IEEE Trans Med Imag 20: 764–771, 2001
16. Wang YP, Wu Q, Castleman KR et al: Chromosome image enhancement using multiscale differential operators. IEEE Trans Med Imag 22: 685–693, 2003
17. Mencattini M, Salmeri M, Lojacono R et al: Mammographic images enhancement and denoising for breast cancer detection using dyadic wavelet processing. IEEE Trans Instrum Meas 57: 1422–1430, 2008
18. Scharcanski J, Jung CR: Denoising and enhancing digital mammographic images for visual screening. Comput Med Imag Grap 30: 243–254, 2006
19. Tsai DY, Lee Y, Chiba R: An improved adaptive neighborhood contrast enhancement method for medical images. Proceedings of IASTED International Conference, BioMed, pp. 59–63, 2005
20. Yoon BW, Song WJ: Image contrast enhancement based on the generalized histogram. J Electron Imaging 16: 033005, 1–8, 2007
21. Fodor IK, Kamath C: Denoising through wavelet shrinkage: an empirical study. J Electro Imaging 2: 151–160, 2003
22. Ferreira CBR, Borges DL: Analysis of mammogram classification using a wavelet transform decomposition. Pattern Recogn Lett 24: 973–982, 2003
23. Cho D, Bui TD, Chen G: Image denoising based on wavelet shrinkage using neighbor and level dependency. Int J Wavelets Multiresolut Inf Process 7: 299–311, 2009

24. J. E. Fowler: The redundant discrete wavelet transform and additive noise. IEEE Signal Process Lett 12: 629–632, 2005
25. Starck JL, Fadili J, Murtagh F: The undecimated wavelet decomposition and its reconstruction. IEEE Trans Image Process 16: 297–309, 2007
26. Wang XY, Yang HY, Fu ZK: A new wavelet-based image denoising using undecimated discrete wavelet transform and least square support vector machine. Expert Syst Appl 37: 7040–7049, 2010
27. Mencattini A, Rabottino G, Salmeri M et al: Denoising and enhancement of mammographic images under the assumption of heteroscedastic additive noise by an optimal subband thresholding. Int J Wavelets Multiresolut Inf Process 8: 713–741, 2010
28. Zhao P, Shang Z, Zhao C: Image denoising based on Gaussian and non-gaussian assumption. Int J Wavelets Multiresolut Inf Process 10: 1250014 (11 pages), 2012
29. Huang Z, Fang B, He X et al: Image denoising based on the dyadic wavelet transform and improved threshold. Int J Wavelets Multiresolut Inf Process 7: 269–280, 2009
30. Matsuyama E, Tsai DY, Lee Y et al: A modified undecimated discrete wavelet transform based approach to mammographic image denoising. J Digit Imaging 26: 748–758, 2013
31. Kim W, You J, Jeong J: Contrast enhancement using histogram equalization based on logarithmic mapping. Opt Eng 51: 067002, 2012
32. Papadopoulos A, Fotiadis JI, Costaridou L: Improvement of microcalcification cluster detection in mammography utilizing image enhancement techniques. Comput Biol Med 38: 1045–1055, 2008
33. Rangayyan RM, Shen L, Shen Y: Improvement of sensitivity of breast cancer diagnosis with adaptive neighborhood contrast enhancement of mammograms. IEEE Trans Inf Technol Biomed 1: 161–170, 1997
34. Jiang J, Yao B, Wason AM: Integrating of fuzzy logic and structural tensor towards mammogram contrast enhancement. Comput Med Imag Grap 29: 83–90, 2005
35. Hashemi S, Kiani S, Noroozi N et al: An image contrast enhancement method based on genetic algorithm. Pattern Recogn Lett 31:1816–1824, 2010
36. Strickland RN and Hahn H: Wavelet transforms for detecting microcalcifications in digital mammograms. IEEE Trans on Med Imag 15: 218–229, 1996
37. Tsai DY, Lee Y, Sakaguchi S: A preliminary study of wavelet-coefficient transfer curves for the edge enhancement of medical images. Transactions of the Japanese Society for Medical and Biological Engineering 40: 86–90, 2002
38. Tsai DY, Lee Y: A method of medical image enhancement using wavelet-coefficient mapping functions. Proceedings of IEEE 2003 International Conference on Neural Networks and Signal Processing, vol. 2, pp. 1091–1094, 2003
39. Heinlein P, Drexl J, Schneider W: Integrated wavelets for enhancement of microcalcifications in digital mammography. IEEE Trans Med Imag 22: 402–413, 2003
40. Lee Y, Tsai DY, Suzuki T: Contrast enhancement of medical images using sigmoid-type transfer curves for wavelet coefficient weighting adjustment. Med Imag Inform Sci 25: 48–53, 2008
41. Matsuyama E, Tsai DY, Lee Y et al: Comparison of a discrete wavelet transform method and a modified undecimated discrete wavelet transform method for denoising of mammograms. Proceedings of 34th Annual International Conference of the IEEE EMBS, pp. 3403–3406, 2013
42. Mammographic Image Analysis Society. http://peipa.essex.ac.uk/info/mias.html. Accessed 11 Jan 2012
43. Scheffe H: The analysis of variance. New York: Wiley, 1959
44. Canavos GC, Koutrouvelis JA: An introduction to the design & analysis of experiments. Upper Saddle River: Pearson Prentice Hall, 2008 (eBook)
45. Gruber M, Weber M, Homolka P et al: Feasibility of dose reduction using needle-structured image plates versus powder-structured plates for computed radiography of the knee. Am J Roentgenol 197:318–323, 2011

46. Liu X, Shaw CC, Lai CJ et al: Comparison of scatter rejection and low-contrast performance of scan equalization digital radiography (SEDR), slot-scan digital radiography, and full-field digital radiography systems for chest phantom imaging. Med Phys 38:23–33, 2011
47. Schaefer-Prokop C, Neitzel U, Venema HW et al: Digital chest radiography: an update on modern technology, dose containment and control of image quality. Eur J Radiol 18: 1818–1830, 2008
48. Donoho DL: De-noising by soft-thresholding. IEEE Trans Inf Theory 41:613–627, 1995
49. Ferrari RJ, Winsor R: Digital radiographic image denoising via wavelet-based hidden Markov model estimation. J Digit Imaging 18:154–167, 2005
50. Harpan MD: A computer simulation of wavelet noise reduction in computed tomography. Med Phys 26:1600–1606, 1999
51. Jansen M, Uytterhoeven G, Bultheel A: Image de-noising by integer wavelet transforms and generalized cross validation. Med Phys 26:622–630, 1999
52. Okamoto T, Furui S, Ichiji H et al: Noise reduction in digital radiography using wavelet packet based on noise characteristics. J Signal Processing 8:485–494, 2004
53. Tischenko O, Hoeschen C, Buhr E: Reduction of anatomical noise in medical X-ray images. Radiat Prot Dosim 114:69–74, 2005
54. Watanabe H, Tsai DY, Lee Y et al: An integrated method of wavelet coefficient thresholding for reducing radiation dose while maintaining diagnostic image quality. Med Imag Inform Sci 28:51–56, 2011
55. Yasuda N, Ishikawa Y, Kodera Y: Improvement of image quality in chest MDCT using nonlinear wavelet shrinkage with trimmed-thresholding. Jpn J Radiol Technol 61:1599–1608, 2005
56. Chang SG, Yu B, Vetterli M: Adaptive wavelet thresholding for image denoising and compression. IEEE Trans Image Process 9:1532–1546, 2000
57. Watanabe H, Tsai DY, Lee Y et al: Improvement of image quality and radiation dose reduction in digital radiography using an integrated wavelet-transform-based method. Proceeding of XX IMEKO World Congress, TC13-P-2 (342), pp. 1–4, 2012
58. Donoho DL, Johnstone JM: Ideal spatial adaptation via wavelet shrinkage. Biometrila 81:425–455, 1994
59. Donoho DL, Johnstone JM: Adapting to unknown smoothness via wavelet shrinkage. J Am Stat Assoc 90:1200–1224, 1995
60. Zhang M, Gunturk BK: Multiresolution bilateral filtering for image denoising. IEEE Trans Image Process 17:2324–2333, 2008
61. Karthikeyan K, Chandrasekar C: Speckle noise reduction of medical ultrasound images using Bayesshrink wavelet threshold. Int J of Comput Appl 22:8–14, 2011
62. Sendur L, Selesnick IE: Bivariate shrinkage functions for wavelet-based denoising exploiting interscale dependency. IEEE Trans Signal Process 50:2744–2756, 2002
63. Watanabe H, Tsai DY, Lee Y et al: Evaluation of irreversible compressed images in computed radiography using physical image quality measures. Jpn J Radiol Technol 65:1618–1627, 2009
64. Samei E, Flynn MJ: An experimental comparison of detector performance for computed radiography systems. Med Phys 29:447–459, 2002
65. Bankman IN: Handbook of Medical Imaging. San Diego: Academic Press, pp. 24–26, 2000
66. Lim JS: Two-dimensional signal and image processing. Englewood Cliffs: Prentice Hall, p. 548, 1989

Chapter 7
Visual Quality Assessment of Stereoscopic Image and Video: Challenges, Advances, and Future Trends

Che-Chun Su, Anush Krishna Moorthy, and Alan Conrad Bovik

7.1 Introduction

Along with other digital visual media [8, 10], the amount of stereoscopic/3D content delivered by the cinema, television, and entertainment industries for human consumption has been growing dramatically over the past few years. According to the latest theatrical market statistics gathered by the Motion Picture Association of America (MPAA) [73], the proportion of cinema screens that are 3D has reached 35 % worldwide, and approximately half of all cinema-goers viewed at least one 3D cinema in 2012. As Hollywood director James Cameron, who directed and produced Avatar, one of the most successful 3D presentations of recent times, stated in an interview with BBC news in August 2013 [3]: "All forms of entertainment will eventually be 3D, because that's how we see the world."

In fact, the wave of 3D has not been limited to the entertainment industry. With greatly improved acquisition and display technologies, stereoscopic/3D images and videos provide natural and versatile visual representations in numerous applications, including robot navigation [2], remote education [115], medical body exploration [114], therapeutic treatment [26], and so forth. As these huge volumes of stereoscopic/3D data are making their way to the consumers, efficient compression and transmission of such data, especially over already-stressed wireless networks, becomes important. In every stage of capture, compression, storage

C.-C. Su (✉) • A.C. Bovik
The University of Texas at Austin, Austin, TX, USA
e-mail: ccsu@utexas.edu; bovik@ece.utexas.edu

A.K. Moorthy
Qualcomm Inc., San Diego, CA, USA
e-mail: anushmoorthy@gmail.com

C. Deng et al. (eds.), *Visual Signal Quality Assessment*,
DOI 10.1007/978-3-319-10368-6_7

and transmission, it is desirable to maximize the final visual experience, and incorporating principles of human perception of stereoscopic/3D quality is of importance [11, 84].

The ideal way to assess perceived visual quality is to run a subjective test to gauge human opinion [43]. However, subjective quality assessment has two obvious disadvantages, making it unsuitable for practical applications. First, the procedure of subjective quality assessment is expensive, tedious, and time-consuming as it has to be performed with great care in order to obtain meaningful results. Second, it is impossible to integrate subjective quality assessment operations of any value into real-time systems for processing actual data. Therefore, one develops automated methods or algorithms that attempt to predict the perceptual quality of visual stimuli. Automated evaluation of visual quality with the assistance of an algorithm is referred to as objective quality assessment. In this chapter, we shall focus on the objective quality assessment of stereoscopic/3D images and videos, unless otherwise specified.

We shall first discuss the challenges and difficulties one may face while trying to design and develop an effective objective quality assessment algorithm for stereoscopic/3D content. Next, we shall examine and analyze different types of objective stereoscopic/3D quality assessment algorithms, in terms of design and performance on publicly available databases. Finally, we shall discuss possible future trends in the field of algorithmic stereoscopic/3D quality assessment.

7.2 Challenges in Stereoscopic Quality Assessment

Humans perceive visual stimuli from natural environments via the two horizontally located frontal eyes, and form a three-dimensional percept using the two corresponding responses generated in the primary visual cortex of the brain. The ultimate goal of a stereoscopic/3D content delivery system is to capture, store, transmit, and display the stereoscopic/3D representation such that it recreates the same 3D percept as experienced by a human in natural environments. Since the human visual system undertakes a variety of signal manipulations and operations to convert the visual stimuli to 3D perception, the problem of predicting perceptual quality is not an easy one. This section summarizes some of the possible issues faced when displaying and viewing stereoscopically captured content.

7.2.1 Visual Discomfort

Due to current limits on stereoscopic/3D capture, broadcast, and display technologies, the experience of watching 3D TV or movies is definitely quite different from what humans naturally view in real life [103]. For example, when a stereoscopic signal is presented, it is projected onto a plane that is at a fixed distance from the viewer. This implies that the human brain is forced to recreate the 3D world with multiple depth planes on this fixed-distance plane of projection, resulting in visual discomfort such as eye-strain and fatigue [54].

The major problem while viewing stereoscopic stimuli stems from the vergence-accommodation conflict. Vergence refers to the simultaneous rotational movement of the two eyes around a vertical axis such that the projection of the object falls at the center of the retina. For example, to look at a nearer object, the two eyes need to rotate or verge towards each other, while for an object farther away, they rotate away from each other, i.e., diverge. Accommodation refers to the ability of the eye to automatically change the optical power of the crystalline lens to keep an object of interest focused as its distance varies. Natural stereoscopic viewing requires a synchronization between vergence and accommodation of the two eyes, since these processes work together to create an overall 3D percept, which is often difficult when viewing stereoscopic/3D content on a flat display at a fixed distance. Specifically, when viewing a 3D object projected on the display screen, the accommodation distance remains constant; however, the 3D position of that object may be located behind the screen, on the screen, or in front of the screen, making the vergence distance vary correspondingly. As a result, a conflict between vergence and accommodation is produced in the simulated stereoscopic/3D viewing scenario. Due to capture calibrations and display specifications, there are other issues that lead to visual discomfort and/or dissatisfaction in stereoscopic/3D viewing, including the keystone distortion, puppet theater effect, crosstalk, cardboard effect, shear distortion, and so forth [62].

Measuring the impact of visual discomfort on the perceptual quality of stereoscopic/3D images and videos is complicated, confounded by the fact that the sensibility and tolerance level of people may differ widely from each other [20, 54, 71, 103]. While it may not be possible to reproduce a 3D percept using projections on a 3D screen, appropriate strategies in capture and display may reduce the effects of visual discomfort and fatigue to the minimal degree, so algorithmically predicting the perceptual quality of stereoscopic/3D image and video is possible. Assuming an approximately ideal display setting, our discussion shall focus on "quality" where the stereoscopic/3D stimulus being perceived is afflicted with different types of processing/transmission distortions such as compression, packet-loss, etc.

7.2.2 *Binocular Vision*

The way distortions of stereoscopic/3D images and videos are perceived by the human visual system (HVS) differs significantly from the perception of distortions on 2D content due to the delicate mechanisms in the binocular vision system that handles similarities and dissimilarities between the two different retinal images [39]. The resulting binocular rivalry effect impacts the perceived quality of stereoscopic/3D stimuli [55]. Binocular rivalry is a perceptual effect that occurs when the two eyes view sufficiently different content at the same retinal locations, which causes match failures; thus, preventing the HVS from flawlessly fusing the left- and right-eye images, resulting in an unnatural 3D perception or a bistable alternation between the two images. A special case of binocular rivalry is binocular suppression [6]. When viewing mismatched stereoscopic/3D stimuli, the brain chooses to not integrate these incompatible stimuli into one coherent percept, and this results in the complete suppressions of one of the two stimuli, referred to as binocular suppression [48]. A deeper understanding of these effects in binocular vision is the key to the development of successful perceptual stereoscopic/3D quality algorithms.

7.2.3 Extra Dimensionality

Modeling natural scene statistics (NSS) and understanding how the HVS processes visual stimuli have been regarded as a dual problem [77,78,87,99]. Many successful 2D image and video quality assessment algorithms exploit robust and effective NSS models derived from both pristine and distorted 2D image/video signals. In particular, several NSS-based 2D no-reference quality assessment algorithms, e.g., [64, 69, 89, 104, 119], have been able to deliver competitive performance to full-reference algorithms. However, using a similar approach in 3D quality assessment has not been met with great success [37, 65]. This is because stereoscopic stimuli from natural environments span up to four dimensions (as opposed to three in the 2D case)—two of space, one of depth and one of time—and joint modeling of all of these four dimensions and the relationships between them become extremely difficult. Later, we shall cover recent work that builds and utilizes preliminary statistics related to image and depth/disparity information that could be used to predict the perceptual quality of stereoscopic/3D stimuli.

7.2.4 Quality Assessment Databases

Another major problem that has existed since the beginning of stereoscopic/3D quality assessment research is a dearth of publicly available and comprehensive databases. By comparison to 2D image and video quality assessment, there is no consensus on a database that could be used for fair and efficient evaluation of different stereoscopic/3D quality algorithms' performance. Several issues are involved in creating valuable stereoscopic/3D quality assessment databases: acquisition protocols, image/video formats, distortion types, visual discomfort control, subjective study paradigm, etc. There is currently no commonly accepted protocol for any of these dimensions when it comes to stereoscopic viewing. Since consensual and comprehensive quality assessment databases play a critical role in developing and evaluating different stereoscopic/3D quality metrics, in the recent past, efforts have been made to create useful, publicly available databases with informative subjective studies conducted on them. We summarize existing stereoscopic/3D quality assessment databases in the next section.

7.3 Advances in Stereoscopic Quality Assessment

In general, image and video quality assessment algorithms can be divided into three categories based on the amount of information available to be utilized to compute the quality score: (1) full-reference (FR), (2) reduced-reference (RR), and (3) no-reference (NR) [109]. Full-reference (FR) algorithms require the original reference signal to be able to predict the quality of the distorted signal. Reduced-reference (RR) approaches perform quality assessment on the distorted signal given some small fraction of information about the original reference signal. This fractional information could range from a set of features or parameters extracted from the original pristine signal to extra side-data, e.g., watermark, imposed on the distorted content. However, there is no clear delineation between FR and RR quality models since most FR models work well even when using a greatly reduced amount of reference data, and moreover, there has not been much industry adoption of exclusively RR models. Finally, no-reference (NR) algorithms, which do form a clear separate class of quality assessment models, are able to gauge the quality of the distorted signal without any additional information extracted from the corresponding reference signal. Since the original, pristine versions of visual signals are rarely available to be transmitted over networks, no-reference image/video quality assessment algorithms find great use in practical applications.

In addition to the three categories, i.e., FR, RR, and NR, commonly used to distinguish 2D image/video quality assessment algorithms, we further classify stereoscopic/3D image/video quality assessment algorithms within each category by utilizing computed or measured depth/disparity information from the stereoscopic pairs.

The goal of objective stereoscopic quality assessment (QA) research is to design algorithms that can automatically assess the quality of 3D images or videos in a perceptually consistent manner. Such human opinions of visual quality are generally obtained by conducting large-scale human studies, referred to as subjective quality assessment, where human observers rate a large number of distorted (and possibly reference) signals. When the individual opinions are averaged across the subjects, a mean opinion score (MOS) or differential mean opinion score (DMOS) is obtained for each of the visual signals in the study, where the MOS/DMOS is representative of the perceptual quality of the visual signal [43]. The goal of an objective QA algorithm is to predict quality scores for these signals such that the scores produced by the algorithm correlate well with human opinions of signal quality. Practical application of QA algorithms requires that these algorithms compute perceptual quality efficiently.

In this section, we summarize recent advances in objective and subjective stereoscopic/3D image and video quality assessment.

7.3.1 Stereoscopic Image Quality Assessment

7.3.1.1 Stereoscopic IQA Without Depth/Disparity Information

The most intuitive and direct way of performing quality assessment on stereoscopic image pairs is applying off-the-shelf 2D image quality assessment algorithms to both the left and right images, and aggregating these two quality scores to form a final quality measure of the stereopair. Some of the early work on stereoscopic image quality assessment were algorithms which avoided the computationally intensive computation of disparity maps between the left and right images. Although such an algorithm does not consider any 3D perceptual effects, this type of naive, yet simple stereoscopic quality assessment algorithms can deliver fairly good performance on symmetrically distorted stereopairs, i.e., when there is approximately the same amount of distortion in the left and right images [36, 72, 118]. Amongst the commonly used conventional 2D image quality assessment algorithms, PSNR [110], SSIM [111], and MS-SSIM [112] are full-reference algorithms, RRED [100] is a reduced-reference model, while DIIVINE [69], BLIINDS [89], and BRISQUE [64] belong to the no-reference category.

Yasakethu et al. [118] apply different 2D image quality algorithms to the left and right views independently to obtain corresponding quality scores, and then compute the average of the two quality scores as the final stereoscopic quality measure. They found that some 2D image quality metrics yield fairly good correlation with overall image quality and perceived depth quality.

Gorley et al. [36] designed a full-reference stereoscopic image quality metric by considering computational models of the HVS. In particular, their quality metric accounts for HVS sensitivity to left–right luminance contrast differentials in regions of high spatial frequency. First, they use SIFT [58] and RANSAC [32] algorithm to extract edges, corners, and regions of high spatial frequency within both the left and right images. Then, these feature points are matched between the left and right views, and the average contrasts of the matched regions are calculated. The final quality score is computed as the difference of the average contrast over all matched regions between the pristine and the distorted stereopairs. Their experimental results suggest that the proposed quality metric produces a more useful threshold than the PSNR metric for practical stereoscopic image compression.

7.3.1.2 Stereoscopic IQA with Depth/Disparity Information

In the past few decades, there has been quite a bit of research conducted towards understanding how human vision systems process and encode depth/disparity stimuli from natural environments [22, 29, 33]. Specifically, it has been discovered that there exist neurons in primary visual cortex with specialized receptive fields tuned to particular disparities, i.e., horizontal position shifts. Based on these perceptual findings, several researchers have utilized relevant natural scene statistical models in stereoscopic vision and image processing problems, achieving superior performance to previous solutions [56, 83, 101]. In addition, stereo image compression has also been studied for two decades [9, 21], and recently has been standardized [1].

The use of simplistic stereoscopic image quality assessment algorithms such as the ones summarized above yields insufficient performance except in special cases (such as symmetric distortions). This is expected, since these algorithms do not use any depth information. Modeling depth cues and incorporating binocular rivalry and suppression is essential when developing a stereoscopic quality assessment model. One would conjecture that just as in the case of other stereoscopic vision problems, this approach would lead to significant gains in the performance of quality assessment algorithms as well. Recently researchers have developed stereoscopic image quality assessment algorithms that utilize some form of depth/disparity information to compute the overall stereoscopic quality score.

Benoit et al. [4] estimate the quality of stereoscopic image pairs using disparity information computed by off-the-shelf stereo algorithms [31, 52]. They first utilize two different 2D image quality metrics, SSIM [111] and C4 [12], to compute quality scores between the left reference and left distorted image, and between the right reference and distorted image. Next, they measure the disparity distortion between the reference and distorted disparity maps computed from the reference and distorted stereopairs, respectively. Finally, the overall quality is computed by fusing the 2D image quality score and the disparity distortion measure. Before discussing their findings, it would be prudent to briefly review the two commonly used 2D image quality metrics, SSIM and C4.

SSIM assesses perceptual image quality by evaluating three factors between the reference and distorted images: luminance, contrast, and structural constancy. It assumes that human visual perception is highly adapted for extracting structural information from a scene, and quality evaluation is hence based on the degradation of this structural information in the distorted visual stimulus. C4 uses an elaborate model of the human vision system to extract perceptual representations from the reference and distorted images. It then compares the two perceptual representations to compute the overall quality score.

From their simulation results, Benoit et al. [4] found that the performance of SSIM with added disparity distortion information is enhanced compared to using the simple average SSIM between the left and right reference-distorted image pairs as the overall stereoscopic quality score. This observation suggests that the quality criteria, i.e., luminance, contrast, and structural constancy, used by SSIM are not sufficient to predict the perceptual quality of stereoscopic image pairs, and the addition of disparity information enhances SSIM performance on stereoscopic content. However, the added disparity information does not improve the performance of C4.

In fact, the average of the C4 quality scores computed from the left and right reference-distorted image pairs only correlates as efficiently as the enhanced SSIM metric with the subjective opinion scores of distorted stereoscopic images. The authors hypothesize that since C4 uses a detailed HVS model, depth information does not impact the overall perceptual quality of stereoscopic/3D images [50].

However, the subjective experiment performed in [4] considers only JPEG and JPEG2000 compression distortions symmetrically applied to the left and right images of the stereopair. As we have mentioned before, in this particular setting, a simple average of 2D quality scores does well. There exist other studies which demonstrate the importance of depth/disparity for perceptual quality evaluation of stereo image pairs. For example, Zwicker et al. [121] used a blurring filter, whose intensity depends on the depth of the region where it is applied to enhance the viewing experience. The study by Okada et al. [76] validated this effect by showing that blurring stereoscopic/3D images reduces the discrepancy between responses of accommodation and convergence, resulting in an enhancement of viewers' overall 3D quality of experience (QoE).

Even though depth/disparity information extracted from the both the pristine and distorted left–right image pairs has a substantial effect on the perceptual quality of stereoscopic images, the question of how to exploit this information remains unanswered. Some recent work investigates the important role played by depth/disparity information in stereoscopic quality assessment. You et al. [120] experimented on the idea of quantifying the degradation of disparity information by applying 2D image quality assessment algorithms to the disparity maps computed from both the reference and distorted left–right image pairs [31]. Specifically, they evaluated a large pool of full-reference 2D image quality assessment algorithms, e.g., PSNR, SSIM, MS-SSIM, VSNR [13], VIF [96], UQI [108], etc., on reference-distorted image pairs, as well as on reference-distorted disparity map pairs, and computed an overall stereoscopic/3D quality score, Q_O using the equation:

$$Q_O = c_1 \cdot Q_I^{c_4} + c_2 \cdot Q_D^{c_5} + c_3 \cdot Q_I^{c_4} \cdot Q_D^{c_5} \tag{7.1}$$

where Q_I and Q_D represent the 2D quality scores of the reference-distorted image and disparity map pairs, respectively, and $\mathbf{c} = \{c_i | i = 1, \cdots, 5\}$ is a set of parameters learned by the Levenberg–Marquardt algorithm [60]. They concluded that applying SSIM on the image pair and mean absolute difference (MAD) on the disparity map pair yields fairly good performance in predicting the overall perceptual quality of stereoscopic images. Some 2D image quality assessment algorithms are capable of generating a quality map between the reference and distorted images to depict the quality degradation at each pixel, e.g., PSNR, SSIM, MS-SSIM, UQI, etc. This quality map can also be generated between the reference and distorted disparity maps to capture an approximate distribution of disparity degradation. Thus, a local combination of the 2D image quality score and the disparity quality score can be achieved by first computing the corresponding quality maps, then applying Eq. (7.1) to pool each image-disparity pixel pair, and finally taking the mean over all pixels as the overall stereoscopic/3D quality score. From the experimental results of this local approach, the combination of UQI on the 2D image pair and SSIM on the disparity map pair was found to give the highest correlation with the subjective opinion scores. Although the experiments performed by You et al. [120] only considered asymmetrically distorted stereopairs, where only the right view is distorted and the left view is intact, their work substantiates the fact that depth/disparity information plays an important role in the design of stereoscopic/3D quality assessment algorithms.

Disparity information can also be used indirectly to bolster a stereoscopic/3D image quality assessment algorithm. Sazzad et al. [90] utilized disparity information to design a no-reference image quality assessment algorithm for both symmetrically and asymmetrically JPEG-coded stereo image pairs. Similar to 2D images, they observed that the visibility of blocking and blurring distortions in JPEG-coded stereoscopic images varies with the local texture. This visual effect is referred to as texture/contrast masking [86, 111]. For example, blockiness is more visible in uniform areas, i.e., non-textured surfaces, while blur is more visible in high-textured regions. They first applied a block-based segmentation algorithm to obtain uniform and non-uniform regions in both views, and then extracted features from both regions in the reference and distorted stereo image pairs. The features include a blockiness measure and range strength in images, as well as the average zero-crossing rate at matched block pairs from the left and right views. In particular, each matched block pair is formed by a block from the left image with the corresponding block from the right image at the horizontally displaced location defined by the disparity map. Finally, the overall quality score is computed as a non-linear function of all features with weighting coefficients learned from subjective test data. The performance of this no-reference stereoscopic/3D image quality assessment algorithm is comparable with the full-reference algorithm proposed by Benoit et al. [4] for JPEG stereo image pairs. The above utilization of disparity information closely models a perceptually relevant model—the cyclopean image. We discuss the cyclopean image model in the next section.

7.3.1.3 Binocular Vision

In addition to the two types of quality assessment algorithms discussed above: those which do not use any disparity information, and those which adapt successful 2D metrics and exploit additional depth/disparity information researchers have also attempted to utilize different models of binocular vision for stereoscopic/3D image quality assessment. For example, Bensalma et al. [5] proposed a full-reference stereoscopic/3D image quality assessment algorithm based on the binocular fusion process [33, 39]. In particular, they developed a model that computes the binocular energy by reproducing the neural responses of simple and complex cells in primary cortex [33, 74, 75]. The model tries to mimic the HVS by modelling the simple cells responsible for local spatial frequency analysis and then the complex cells responsible for the generation of the binocular energy. This energy is used as an indicator of the quality. The quality score is computed as the difference of binocular energy between the reference and distorted stereo image pairs. The authors found that the binocular energy difference correlates well with the human opinion scores for both symmetrically and asymmetrically JPEG-coded stereoscopic images.

Another full-reference stereoscopic/3D image quality assessment algorithm based on binocular perception is the one by Ryu et al. [88]. The authors extended the three quality components of SSIM, i.e., luminance, contrast, and structural similarities [111], using a nonlinear binocular perception model presented in [61]. First, the three similarity measures are computed between the left reference-distorted image pair, as well as the right reference-distorted image pair. Then, for each of these measures, the two monocular measures from the left and right pairs are combined into a binocular measure using a weighted summation. Finally, the quality score is computed as a nonlinear aggregate of the three binocular similarity measures. This binocular version of SSIM performs better than the monocular version (which is a simple average between the left and right view), on different types of distorted stereoscopic images including JPEG, JPEG2000, and Gaussian blur.

7.3.1.4 Cyclopean Image

The ultimate goal of a stereoscopic/3D image quality assessment algorithm is to estimate the quality of the true cyclopean image [48] formed in an observer's mind when a left–right image pair is stereoscopically presented. Simulating the true cyclopean image associated with a given stereoscopic image pair is a daunting task, since it would require accounting for a variety of issues, including the display geometry, fixation position, vergence, accommodation, etc. It is still unclear recently how the human visual system forms a cyclopean image based on the two visual stimuli received via the retinas of the two eyes, further complicating the task of cyclopean image modelling. However, one can synthesize an intermediate image that more-or-less agrees with the cyclopean perception of a human. The linear model proposed by Levelt [55] models the formation of the perceived cyclopean image when a stereoscopic stimulus is presented:

$$C = w_L \cdot S_L + w_R \cdot S_R \tag{7.2}$$

where S_L and S_R represent the stimuli to the left and right eyes, respectively; w_L and w_R represent the weighting coefficients of the corresponding stimuli such that $w_L + w_R = 1$; and C is the perceived cyclopean image. Levelt hypothesized that the duration of the dominance period of an eye depends on the stimulus strength in the other eye, making the weighting coefficients positively correlated with the relative stimulus strengths between the two eyes. Therefore, given a stereoscopic image pair with a computed or measured disparity map, a synthesized cyclopean image can be obtained by disparity-compensation and coordinate mapping the stereopair.

Assuming the disparity map is computed using the left image as the reference to match the right image, one can generate the synthesized cyclopean image, I_C, as:

$$I_C(x, y) = w_L(x, y) \cdot I_L(x, y) + w_R(x, y + D(x, y)) \cdot I_R(x, y + D(x, y)) \tag{7.3}$$

where (x, y) represents the pixel coordinate; I_L and I_R are the left and right images, respectively; w_L and w_R represent the corresponding weighting map; and D is the computed disparity map that matches pixels from I_R to those in I_L.

Several recent researchers have attempted to evaluate perceptual quality by the use of this synthesized cyclopean image.

Maalouf et al. [59] proposed a reduced-reference quality metric by comparing the sensitivity coefficients [24] extracted from the two cyclopean images formed by the reference and distorted stereopairs. The cyclopean image is computed using Eq. (7.3) with the weighting map defined as the average of the local disparity-compensated values from the left and right images.

Chen et al. [18] proposed a full-reference quality assessment algorithm exploiting a perceptually synthesized cyclopean image to account for binocular rivalry. First, the authors use a Gabor filter bank to perform a perceptual multi-scale, multi-orientation decomposition on both the reference and distorted stereopairs. Next, they use the energy of the Gabor filter responses to model the stimulus strength, i.e., the weighting map of the left and right images to form the perceptually synthesized cyclopean image using Eq. (7.3). Finally, the task of stereoscopic/3D quality assessment is performed by applying a full-reference 2D image quality assessment algorithm, e.g., PSNR, SSIM, MS-SSIM, VIF, etc., on the reference and distorted synthesized cyclopean images. The performances of the full-reference 2D image quality metrics computed on the reference-distorted synthesized cyclopean image pairs are significantly improved as compared to those computed on the non-cyclopean image pairs, especially for asymmetrical distortions. The authors found that MS-SSIM [112] delivers the best performance, surpassing existing stereoscopic/3D image quality metrics. In [15], the authors extended this full-reference stereoscopic/3D image quality assessment framework to a no-reference version using 2D and 3D NSS features extracted from stereoscopic image pairs. For the 2D NSS features, the authors sought inspiration from a highly competitive no-reference 2D image quality assessment algorithm, BRISQUE [64]. They synthesized the cyclopean image as in the FR case, and modelled the histogram of the coefficients so obtained using the generalized Gaussian distribution (GGD). For the 3D NSS features, since the only accessible 3D information in a no-reference algorithm is the estimated disparity from a stereo matching algorithm, the authors first extracted the same features from the estimated disparity map as those from the synthesized cyclopean image. They modelled the distribution of the uncertainty map produced using an SSIM-based block-matching stereo algorithm using a log-normal distribution. Finally, both the 2D and 3D NSS features were used to train a two-stage support vector machine (SVM) model [69] to predict the perceptual quality of a distorted stereoscopic image pair. The authors demonstrated that when there is only symmetric distortion in a stereoscopic image pair, this cyclopean-image-based no-reference quality metric performs as well as the state-of-the-art 2D no-reference image quality metrics. When a stereopair is asymmetrically distorted, it significantly outperforms both 2D and other 3D no-reference image quality assessment algorithms, and delivers competitive performance relative to effective 3D full-reference image quality assessment algorithms.

Similar to most existing no-reference image quality assessment models, [15] is based on training on recorded human subject scores. However, recent developments in "completely blind" 2D image quality models [66], which require no training on distorted images or on human judgments, suggest that such models might also be developed for the stereoscopic/3D image quality assessment problem.

7.3.2 Stereoscopic Video Quality Assessment

As in the case of stereoscopic images, quality assessment of stereoscopic video faces several challenges owing to the processing chain: acquisition, compression, transmission, decompression, and display [41]. The problem is further exacerbated in stereoscopic video due to the presence of the additional dimension of time. Due to the wide variety of coding techniques and content representations, there may be several format conversions (technically, re-compressions) involved in the processing chain. For example, there are two possible representations for stereoscopic/3D video compression and transmission, left-and-right and 2D-plus-depth formats, where depth-image-based rendering (DIBR) approaches are often used to generate the synthesized views based on the depth map. In addition, due to the huge amount of stereoscopic/3D video data, efficient video coding algorithms are necessary for content transmission and delivery. Since state-of-the-art video coding standards such as H.264/AVC [47] and HEVC [46] employ block-based architectures, it is inevitable that artifacts caused by block-based compression contaminates the stereoscopic/3D video sequences. As a result, different types of distortions can be introduced at different steps in the processing chain, and an open question remains on how these different degradations interact with each other and affect the overall perceptual quality of stereoscopic/3D video. We limit our discussion on algorithms and models developed for objectively assessing the overall perceptual quality of stereoscopic/3D video consisting of left and right 2D video sequences stereoscopically presented to viewers after all processing steps and possible format conversions.

7.3.2.1 Stereoscopic VQA Without Depth/Disparity Information

Similar to stereoscopic/3D image quality assessment, a straightforward strategy of performing stereoscopic/3D video quality assessment is to apply off-the-shelf 2D image/video quality metrics (VQMs) and aggregate the scores from the left and right views to compute an overall perceptual quality prediction. An early effort using this methodology was that by Yasakethu et al. [118]. The authors investigated the relationship between subjective quality measures and several objective quality metrics, e.g., PSNR, SSIM, and VQM [81], for stereoscopic/3D video sequences coded in both left-and-right and 2D-plus-depth formats. VQM was developed by the Institute of Telecommunication Sciences (ITS) to provide an objective measurement for perceived 2D video quality. VQM measures the perceptual effects of video impairments including blurring, jerky/unnatural motion, global noise, block distortion and color distortion, and combines them in to one single metric score. Due to its contemporaneous excellent performance in the Video Quality Experts Group (VQEG) validation tests, VQM was adopted by the American National Standards Institute (ANSI) and International Telecommunication Union (ITU) standardization bodies as a measure for 2D video quality assessment.

In [118], the authors conducted two subjective studies to record human opinion scores on both left-and-right and 2D-plus-depth stereoscopic/3D videos with packet-loss distortion. The stereoscopic/3D video sequences were encoded using the joint scalable video model (JSVM) software, which is the reference software of the scalable video coding (SVC) standard, i.e., Annex G extension of the H.264/AVC, developed by Joint Video Team (JVT) [47]. When encoding left-and-right videos, the base layer of the SVC stream is used to encode the left-view sequence while the right-view sequence is encoded in the enhancement layer. For 2D-plus-depth videos, the base layer is used to encode the 2D video sequence, and the enhancement layer is used to encode the depth map sequence. All stereoscopic/3D video sequences were asymmetrically encoded and corrupted with different packet-loss rates. Experimental results showed that the average VQM score of the decoded left- and right-view video sequences was able to effectively predict the overall perceptual quality of asymmetrically compressed stereoscopic/3D video under packet-loss scenarios.

7.3.2.2 Stereoscopic VQA with Depth/Disparity Information

The HVS utilizes a variety of depth cues available in natural scenes to build a unified perception of depth [40]. Depth cues can be classified into two types: binocular cues and monocular cues. Binocular cues require cooperation from both eyes, e.g., retinal disparity and convergence, while monocular cues can be perceived with a single eye, e.g., object size, occlusion, perspective, motion parallax, etc. In the design of stereoscopic/3D video quality assessment algorithms, one would conjecture that appropriate use of such depth cues would improve performance.

Boev et al. [7] proposed a full-reference quality assessment algorithm for stereoscopic/3D video using both monoscopic and stereoscopic quality measures. The monoscopic quality component measured trivial monoscopic artifacts in 2D images, e.g., blur, noise, blockiness, etc., and the stereoscopic quality component assessed the perceived degradation from binocular depth cues. The authors first apply a Gaussian pyramid to both the left and right views in the reference and distorted video sequences, and compute the corresponding disparity map along with a similarity map generated using SSIM as the perceptual similarity measure at each scale. Next, a monoscopic quality map is constructed by applying SSIM to the reference and distorted cyclopean images formed using the disparity maps, and a stereoscopic quality map is obtained by combining the two absolute difference maps between the reference-distorted disparity and similarity maps. Finally, a monoscopic and a stereoscopic quality measure are computed by aggregating the monoscopic and stereoscopic quality maps across all scales. The authors show that a combination of monoscopic and stereoscopic quality measures correlates well with subjective opinions thereby beating PSNR performance.

Jin et al. [45] proposed a full-reference stereoscopic/3D VQM using 3D-DCT and features from contrast masking. The 3D-DCT is a decorrelating transform that achieves highly sparse representations of 3D visual stimuli [23]. First, the disparity maps of the reference and distorted stereoscopic/3D video sequences are computed. Next, the authors exploit a model of saccades, i.e., the pseudo-random movements the eyes perform while processing spatial information [105], by stacking into a 3D array the current block, its most similar block from the same view, and the two most similar blocks from the other view within a search range based on the disparity. A 3D-DCT transform is applied to the stacked 3D arrays formed by the left and right image pairs from both the reference and distorted stereoscopic/3D video sequences. Finally, the quality score is computed as the mean squared error (MSE) between the reference and distorted frequency-domain coefficients weighted by a perceptual mask which models the human contrast sensitivity function (CSF) [30, 82]. Experimental results show that the proposed measure outperforms other popular full-reference 2D quality metrics, e.g., PSNR, SSIM, MS-SSIM, and UQI, on compressed stereoscopic/3D video sequences using the database from [49].

7.3.2.3 Binocular Vision

The HVS derives correspondences between stereoscopic views predominantly from coarser spatial structures, and then fine-tunes them from finer spatial details [63, 85]. In practice, stereoscopic/3D videos are not compressed to a level so that the global structure of an image is disturbed. In other words, compression artifacts do not affect the global correspondence, resulting in negligible effects in depth perception [95]. Therefore, compression artifacts are perceived as local distortions, mainly affecting the 2D image quality of stereoscopic/3D videos. In [27], Silva et al. conducted a subjective experiment to compare the impact of asymmetrically distorted stereoscopic videos with blurring and blocking artifacts. The results suggest that perceptual quality is dominated conversely by the higher quality view when the lower quality view is degraded by blur. It is the lower quality view that determines the overall quality when blocking artifacts appear.

Based on these findings, Silva et al. [28] proposed an FR quality metric for stereoscopic/3D videos that consists of three types of measurements: structural distortions, blurring artifacts, and content complexity. Specifically, structural distortion is measured by utilizing the correlation coefficient between the reference and compressed views [44], blur artifacts are defined as a loss of edge magnitude in visually significant areas in the compressed view, and content complexity is computed as a combination of the spatial and temporal perceptual information measures defined in [42] from the reference view. Finally, these three measures are non-linearly aggregated to yield the overall perceptual quality score. The authors also conducted two subjective experiments to uncover different patterns of subjective scoring for symmetrically and asymmetrically compressed stereoscopic/3D videos. These subjective results were utilized to train and validate the proposed stereoscopic/3D VQM, which provides good accuracy and consistency in predicting asymmetrically compressed stereoscopic/3D videos.

7.3.3 Databases

Since the human is the ultimate receiver of the visual signal, the performance of any image and video quality assessment algorithm is gauged by its correlation with human subjective judgements of quality.

Hence, the goal of an objective stereoscopic/3D image/video quality assessment algorithm is to take these stimuli as input, and generate the corresponding quality scores that follow human opinion scores, i.e., MOS/DMOS, as close as possible. Practical applications of quality assessment algorithms requires that these algorithms compute perceptual quality scores efficiently and robustly. In this section, we summarize available stereoscopic/3D image and video databases that can serve to help design and validate practical quality assessment algorithms.

As with databases that are widely used to compare 2D IQA and VQA models [67, 93, 94, 97, 98], databases of pristine and distorted 3D content annotated by human subject scores are of great value for evaluating 3D quality models and algorithms.

7.3.3.1 Stereoscopic/3D Image Databases

1. **IRCCyN/IVC 3D Images Database** [4]: One of the first publicly available databases on stereoscopic/3D image quality assessment, the IRCCyN/IVC 3D Images Database developed at the Institut de Recherche en Communications et Cybernétique de Nantes (IRCCyN), France, contains 6 references and 90 distorted stereoscopic images, 15 from each reference pair, at an image resolution of 512×512 pixels. Three different types of distortions are applied symmetrically to the stereoscopic image pairs: JPEG compression, JPEG2000 compression, and Gaussian blur. Human judgments were collected from 19 subjects.
2. **Toyoma/MICT** [91]: This database consists of 490 symmetrically and asymmetrically JPEG-coded stereoscopic image pairs created from 10 reference stereo pairs. The images are at a resolution of 640×480, and the JPEG compression contains seven different quality scales, including the reference. Human judgments were collected from 24 non-expert subjects.
3. **Ningbo Stereoscopic Image Quality Assessment Database (SIQAD)** [107]: The Ningbo SIQAD database consists of ten reference stereoscopic image pairs from the Middlebury Stereo Datasets [92] and a total of 400 asymmetrically distorted pairs at a variety of image resolutions, including 1,390×1,110, 1,342×1,110, 1,330×1,110, 1,276×1,110, and 1,252×1,110. Four different types

of distortions are used in this database: JPEG compression, JPEG2000 compression, Gaussian blur, and white noise. Twenty non-expert subjects were recruited to participate in the subjective study.

4. **MMSPG 3D Image Quality Assessment Database** [34]: This database was developed by researchers at EPFL, Switzerland, to study the impact of acquisition distortions on the quality of stereoscopic images. A subjective study was conduction on (high quality) JPEG-compressed stereoscopic image pairs with six different settings of inter-camera distances. The database contains ten scenes at an image resolution of 1,920×1,080. One scene was used for training, and the other nine scenes were used in the test sessions, resulting in a total of 54 distorted stimuli. Human judgments were collected from 17 subjects.
5. **LIVE 3D Image Quality Database** [68]: Developed at the University of Texas at Austin, USA, the LIVE database is the first publicly available stereoscopic/3D image quality assessment database that provides researchers access to ground-truth depth information. It was constructed in two phases. Phase I [72] contains 20 pristine and 365 distorted stereoscopic image pairs with symmetrical distortions, while phase II [18] contains 8 pristine and 360 distorted stereoscopic image pairs with both symmetrical and asymmetrical distortions. All images are at a resolution of 1,280×720. Both phases include five different types of distortions: JPEG compression, JPEG2000 compression, additive white Gaussian noise, Gaussian blur, and a Rayleigh fast-fading channel distortion. For the subjective studies conducted in both phases, each subject reported normal or corrected normal vision and no acuity or color test was deemed necessary. Stereo Randot Test [113] was used to pre-screen participants for normal stereo vision in phase II. Phase I utilized 32 participators with a male-majority population. In phase II, 6 females and 27 males participated in the experiment, aged between 22 and 42 years. A single stimulus continuous quality evaluation (SSCQE) [43] experiment with hidden reference was conducted in both phases. The two phases together comprise the largest and most comprehensive stereo quality database currently available.
6. **IEEE Standards Association Stereo Image Database** [80]: This database was built by researchers at Yonsei University, Korea, to study the impact of depth distribution and scene content, e.g., outdoor and indoor, on the degree of visual discomfort that is experienced when viewing stereoscopic images. It contains a total of 800 stereo image pairs created from 160 reference scenes using 5 evenly separated convergence points. All images in the database have high resolution 1,920×1,080. A subjective discomfort study was conducted, and human opinion scores were collected from 24 non-expert participants.

7.3.3.2 Stereoscopic/3D Video Databases

1. **MMSPG 3D Video Quality Assessment Database** [35]: Developed by researchers at EPFL, Switzerland, the MMSPG 3D VQA database is one of the first publicly available databases for stereoscopic/3D video quality

assessment. This database was built for a subjective study on the impact of acquisition distortions on the quality of stereoscopic videos. It contains a total of 30 stereoscopic video sequences at a resolution of 1,920×1,080 and a frame rate of 25 fps, captured from six scenes with five different inter-camera distances. All stereoscopic video sequences are encoded and stored in (high quality) H.264/AVC standard format at a bit-rate of 24 Mbps. A single stimulus continuous quality evaluation (SSCQE) [43] method was adopted for collecting human opinion scores from 17 non-expert subjects.

2. **IRCCyN/IVC NAMA3DS1-COSPAD1 3D Video Quality Database** [106]: The IRCCyN/IVC 3D video quality database consists of 110 stereoscopic video sequences created from ten reference videos. Each reference video was degraded with ten different types of distortions, including H.264/AVC video encoding and JPEG2000 image compression, as well as common image processing operations such as downsampling and sharpening. All video sequences were stored at a resolution of 1,920×1,080 and a frame rate of 25 fps, and rated by 29 subjects using the absolute category rating with hidden reference (ACR-HR) [43] scale.
3. **Surrey** [28]: This database contains 116 stereoscopic video sequences created from 14 reference videos at a resolution of 1,920×1,080 and a frame rate of 25 fps. Two video encoding standards, H.264/AVC and HEVC, were adopted to asymmetrically compress the reference video sequences with a wide range of quantization parameter combinations. Human judgements were collected from 16 non-expert subjects using a double stimulus continuous quality scale (DSCQS) [43] method.

A comparison of many of the databases mentioned here, as well as those that are not using a variety of measures (some of which are of questionable relevance or value), appears in [117]. The author of [117] also maintains a comprehensive list of image and video quality assessment databases in [116].

7.3.4 Performance Evaluation

We evaluate and compare the performance of some of the stereoscopic/3D quality metrics we have discussed in this chapter on the LIVE 3D Image Quality Database Phase II, which consists of both symmetrically and asymmetrically distorted stereoscopic image pairs. There are five different types of distortions in the LIVE 3D Image Quality Database Phase II: JPEG and JPEG2000 (JP2K) compression, additive white Gaussian noise (WN), Gaussian blur (Blur), and a Rayleigh fast-fading channel distortion (FF). The degradation of stimuli varies with controlled parameters for each type of distortion, where the ranges of these parameters are decided beforehand to ensure that all types of distortions vary from almost invisible to severely distorted with a good overall perceptual separation between distortion levels throughout. For full-reference algorithms, we use all reference and distorted stereopairs, while for no-reference algorithms, we divide the entire database into

80 % training and 20 % testing such that no overlap occurs between training and testing image content. This train-test procedure is repeated 1,000 times to ensure that there was no bias due to the image content used for training. We report the median performance across all iterations.

We compute both Spearman's rank-order correlation coefficient (SROCC) and Pearson's linear correlation coefficient (LCC) of the quality scores generated by different quality assessment algorithms against the subjective opinion scores (DMOS) to evaluate their performance. SROCC and LCC measure the monotonicity and accuracy, respectively, of the predicted quality score by a quality assessment algorithm against DMOS, where a value of 1 indicates perfect correlation. Since LCC is a linear correlation measure, all algorithm scores are passed through a logistic non-linear function for mapping to the DMOS space before computing LCC. This is a standard procedure used to align the performances of both image and video quality assessment algorithms [97]. The SROCC and LCC scores of the image quality assessment algorithms evaluated on the LIVE 3D Image Quality Database Phase II are summarized and tabulated in Table 7.1. To further analyze the effectiveness of these algorithms, we also report their performance on different types of distorted stereoscopic image pairs, as well as symmetrical and asymmetrical distortions, in Tables 7.2 and 7.3, respectively. Note that no-reference algorithms are italicized in all three tables.

It can be seen from Table 7.1 that the best-performing "simple" algorithm that applies 2D image quality assessment algorithms to stereoscopic/3D image pairs achieves a correlation of 0.8 against subjective opinion scores. On the other hand, the best image quality assessment algorithm utilizing 3D information is able to deliver 3D quality prediction performance that achieves 0.9 correlation against human judgments. In particular, the synthesized cyclopean image boosts the performance of 2D image quality assessment algorithms, e.g., MS-SSIM, by more than 10 % of correlation, which is quite significant.

Moreover, with the aid of synthesized cyclopean image and effective 3D NSS models, the no-reference stereoscopic/3D image quality assessment algorithm proposed by Chen et al. is able to deliver comparable correlation performance comparable to the best full-reference algorithm.

Table 7.1 Comparison of different 2D and 3D image quality assessment algorithms on LIVE 3D Image Quality Database Phase II

Algorithm		LCC	SROCC
2D	PSNR	0.680	0.665
	SSIM [111]	0.802	0.792
	MS-SSIM [112]	0.783	0.777
	BRISQUE [64]	0.782	0.770
3D	Benoit [4]	0.748	0.728
	You [120]	0.800	0.786
	Hewage [38]	0.558	0.501
	Cyclopean MS-SSIM [18]	0.900	0.889
	Sazzad [90]	0.568	0.543
	Chen [15]	0.895	0.880

Table 7.2 Comparison (SROCC) of different 2D and 3D image quality assessment algorithms on different distortion types in LIVE 3D Image Quality Database Phase II

Algorithm		WN	JP2K	JPEG	Blur	FF	Overall
2D	PSNR	0.919	0.597	0.491	0.690	0.730	0.665
	SSIM [111]	0.922	0.704	0.678	0.838	0.834	0.792
	MS-SSIM [112]	0.946	0.798	0.847	0.801	0.833	0.777
	BRISQUE [64]	0.846	0.593	0.769	0.862	0.935	0.770
3D	Benoit [4]	0.923	0.751	0.867	0.455	0.773	0.728
	You [120]	0.909	0.894	0.795	0.813	0.891	0.786
	Hewage [38]	0.880	0.598	0.736	0.028	0.684	0.501
	Cyclopean MS-SSIM [18]	0.940	0.814	0.843	0.908	0.884	0.889
	Sazzad [90]	0.714	0.724	0.649	0.682	0.559	0.543
	Chen [15]	0.950	0.867	0.867	0.900	0.933	0.880

Table 7.3 Comparison (SROCC) of different 2D and 3D image quality assessment algorithms on symmetrically and asymmetrically distorted stimuli in LIVE 3D Image Quality Database Phase II

Algorithm		Symmetric	Asymmetric	Overall
2D	PSNR	0.776	0.587	0.665
	SSIM [111]	0.828	0.733	0.792
	MS-SSIM [112]	0.912	0.684	0.777
	BRISQUE [64]	0.849	0.667	0.770
3D	Benoit [4]	0.860	0.671	0.728
	You [120]	0.914	0.701	0.786
	Hewage [38]	0.656	0.496	0.501
	Cyclopean MS-SSIM [18]	0.923	0.842	0.889
	Sazzad [90]	0.420	0.517	0.543
	Chen [15]	0.918	0.834	0.880

Table 7.2 details the performance of each quality assessment algorithm on different types of distorted stereoscopic/3D image pairs. We can see that almost all 2D and 3D algorithms are able to predict quality scores correlating well with human opinions of stereoscopic image pairs affected by the WN distortion. However, several quality metrics perform poorly when predicting the perceptual quality of stereopairs with JPEG, JP2K, and Blur distortions. These poor performances can be caused by the facilitation effect [14] of distorted stereoscopic/3D image pairs: distortions co-located with high depth variations are more easily found by human subjects, a phenomenon which is not yet well understood or modeled.

Finally, Table 7.3 demonstrates the effectiveness of different quality algorithms on symmetrically and asymmetrically distorted stereoscopic/3D image pairs. It can be clearly seen that the stereopairs contaminated with unequal amount of distortions in the left and right images challenge most of the examined quality assessment algorithms.

In summary, these evaluation results of different 2D and 3D quality assessment algorithms confirm the necessity of utilizing and incorporating accessible 3D information, e.g., measured/estimated depth/disparity maps, when predicting the perceptual quality of stereoscopic/3D image pairs.

7.4 Future Trends

As we have discussed, there has been a considerable amount of work and research conducted in the field of stereoscopic/3D image and video quality assessment. In this section, we discuss the broader topic of predicting the overall QoE when viewing stereoscopic stimuli and speculate on possible research directions for future research in the field of stereoscopic image and video quality assessment.

7.4.1 Quality of Experience

Current research on stereoscopic/3D quality assessment has mainly focused on the discrepancy of image quality between the reference and distorted stereo stimuli utilizing the estimated or measured depth/disparity information. However, due to the extra dimensionality of the stimuli and the wide variety of display technologies, “quality of experience” would be a more appropriate term to define the overall palatability of stereoscopic/3D presentations. Specifically, the additional dimension of depth, along with unwanted side effects induced by imperfect geometry or poor stereography, leading to visual discomfort or fatigue, can affect the experience of viewing stereoscopic/3D stimuli in both positive and negative ways. Hence, a variety of factors need to be considered when creating stereoscopic/3D content, in order to be able to deliver a pleasant stereoscopic/3D viewing experience [25]. Lambooij et al. [53] expressed the stereoscopic/3D QoE as the weighted sum of perceived image quality and depth. More recently, Chen et al. [19] proved that visual comfort becomes the dominant factor (over transmission/compression distortions) in determining stereoscopic/3D viewing experiences when visual discomfort or fatigue occurs. In [20], the authors further proposed to measure the overall visual QoE when viewing stereoscopic/3D stimuli as the weighted sum of image quality, depth quantity, and visual comfort.

In the case of stereoscopic/3D videos, the study conducted by Chen et al. [16, 17] showed that when viewing stereoscopic/3D videos, subjects tend to agree on perceived image quality, but have more diverse opinions on sensation of depth. Since motion parallax, i.e., relative movement between objects, is a strong depth cue, motion serves as an important factor affecting stereoscopic/3D viewing experiences wherein it is able to give rich depth satisfaction, but it also contributes to visual discomfort. López et al. [57] conducted subjective studies to quantify the effects of motion parallax and temporal evolution of depth histograms while viewing stereoscopic/3D videos. They proposed to calculate the overall QoE as a linear combination of the probabilities of detecting window violations, of finding abrupt scene transitions, and of observing excessive negative motion parallax.

These preliminary forays into visual discomfort hint at possible directions for the development of accurate and robust algorithms for overall stereoscopic/3D QoE [79]. For example, Kim et al. [51] proposed a visual fatigue predictor that measures excessive horizontal and vertical disparities using features extracted from estimated disparity maps. In addition to disparity statistics, Park et al. [79] utilized features based on principles of physiological optics and foveation to develop a visual discomfort predictor that accounts for accommodation-vergence mismatches when viewing stereoscopic images.

7.4.2 Content Diversity

So far, all of the discussion in this chapter has ignored the role of the diversity of content. Quality assessment algorithms are almost always content-blind, and only look at low-level image or depth features such as texture and discontinuity to predict the perceptual quality. However, humans judge the overall viewing experience at a much higher-level. For example, when a baby appears in a stereoscopic/3D image or video, subjects may give higher ratings of quality even if the distortion is unacceptable. While current stereoscopic/3D quality assessment databases, similar to the 2D case, attempt to remove this bias by selecting content with no strong feelings of like or dislike, future database creation efforts should include consideration of a diverse range of stereoscopic/3D image and video content. Moreover, even under similar geometric settings, different content may induce different sensation of perceived depth, affecting the overall viewing experience. Therefore, as automatic visual quality assessment aims to replace the human observer, the area of content diversity and bias needs to be explored [70].

7.4.3 Natural Scene Statistical Modeling

Effective and robust NSS models have proved to be an essential ingredient in the development of more successful 2D image/video quality assessment algorithms. Due to high dimensionality of stereoscopic/3D stimuli and an incomplete understanding of human depth perception, stereoscopic/3D quality metrics have not benefited much by statistical modeling. Some early work, however, did attempt to exploit the basic statistics and relationships between image and depth/disparity information in predicting stereoscopic/3D perceptual quality. Ha et al. [37] conducted subjective tests to examine different factors that may affect depth perception and visual comfort while viewing stereoscopic videos. They proposed a no-reference stereoscopic VQM by training a linear regression model with features extracted from motion vector magnitude, intra- and inter-frame disparity variation, and disparity distribution at boundary areas. In [65], Mittal et al. proposed a no-reference model to evaluate the perceptual quality of stereoscopic/3D image and video using sample statistics computed from both the left and right images, the estimated disparity map, and the motion-compensated disparity difference. A simple linear regression model was adopted to map the extracted features to subjective scores.

To explore more advanced and effective statistics embedded in natural image and depth data, Su et al. [102] studied both the joint and conditional distributions of spatially adjacent luminance/chrominance and depth wavelet coefficients in natural scenes, and modeled them using the relevant, versatile, and flexible bivariate GGD. They found that there exist both scale and orientation dependencies in the joint distributions, and both spatially adjacent luminance and chrominance coefficients maintain constant correlation when conditioned on depth wavelet coefficients. When stereoscopic/3D stimuli suffer from different types of distortions, these dependencies and bivariate models are able to provide useful information predictive of perceptual quality. In addition to these early algorithmic and statistical efforts, future development of stereoscopic/3D quality metrics should include a focus on advanced modeling of disparity and motion masking effects, both of which remain poorly understood.

Conclusion

In this chapter, we summarized recent advances in visual quality assessment of stereoscopic/3D image and video. We first outlined practical challenges one may face when attempting to design effective stereoscopic quality metrics. In particular, exploring the high-dimensionality statistics of stereoscopic/3D stimuli and incorporating complicated models of binocular vision are both critical. Our summary demonstrated that while measuring the perceptual quality of stereoscopic/3D content by simply combining off-the-shelf or extending 2D quality assessment algorithms can provide a basic level of performance, the use of depth/disparity statistics, binocular vision models, and so forth result in much more efficient and accurate quality prediction algorithms.

We also discussed possible future directions towards developing successful stereoscopic/3D quality metrics, and described the concept of "quality of experience" which includes not only image quality but also depth sensation and visual comfort.

The field of stereoscopic/3D quality assessment is certainly growing, but is far from mature. There is tremendous scope for research in this area owing to its complex and multidisciplinary nature. We postulate that there remain large gaps between our understanding of human stereo perception and statistical modeling of natural image and depth information. Accurate and

(continued)

robust prediction of stereoscopic/3D perceptual quality will hopefully emerge by combining research findings in both vision science and image/video engineering.

References

1. American Society for Testing and Materials (ASTM): Standard specification for 3D imaging data exchange. Active Standard ASTM E2807-11 (2013)
2. Baltes, J., McCann, S., Anderson, J.: Humanoid robots: Abarenbou and daodan. RoboCup-Humanoid League Team Description (2006)
3. BBC News - Technology: James Cameron: All entertainment 'inevitably 3D'. http://www.bbc.co.uk/news/entertainment-arts-23790877 (2013)
4. Benoit, A., Callet, P.L., Campisi, P., Cousseau, R.: Quality assessment of stereoscopic images. EURASIP Journal on Image and Video Processing **2008**, 1–13 (2009)
5. Bensalma, R., Larabi, M.C.: A perceptual metric for stereoscopic image quality assessment based on the binocular energy. Multidimensional Systems and Signal Processing **24**(2), 281–316 (2013)
6. Blake, R., Westendorf, D.H., Overton, R.: What is suppressed during binocular rivalry? Perception **9**(2), 223–231 (1980)
7. Boev, A., Gotchev, A., Egiazarian, K., Aksay, A., Akar, G.B.: Towards compound stereo-video quality metric: a specific encoder-based framework. In: Proceedings of the IEEE Southwest Symposium on Image Analysis and Interpretation, pp. 218–222 (2006)
8. Bovik, A.: Automatic prediction of perceptual image and video quality. Proceedings of the IEEE **101**(9), 2008–2024 (2013)
9. Bovik, A., Chen, D.: Method and apparatus for processing both still and moving visual pattern images. US Patent 5 282 255 (1994)
10. Bovik, A.C.: The essential guide to video processing. Academic Press (2009)
11. Brown, M.Z., Burschka, D., Hager, G.D.: Advances in computational stereo. IEEE Transactions on Pattern Analysis and Machine Intelligence **25**(8), 993–1008 (2003)
12. Carnec, M., Le Callet, P., Barba, D.: An image quality assessment method based on perception of structural information. In: Proceedings of the IEEE International Conference on Image Processing, vol. 3, pp. 185–188 (2003)
13. Chandler, D., Hemami, S.: VSNR: A wavelet-based visual signal-to-noise ratio for natural images. IEEE Transactions on Image Processing **16**(9), 2284–2298 (2007)
14. Chen, M.J., Bovik, A.C., Cormack, L.K.: Study on distortion conspicuity in stereoscopically viewed 3D images. In: Proceedings of the IEEE IVMSP Workshop, pp. 24–29 (2011)
15. Chen, M.J., Cormack, L.K., Bovik, A.C.: No-reference quality assessment of natural stereopairs. IEEE Transactions on Image Processing **22**(9), 3379–3391 (2013)
16. Chen, M.J., Cormack, L.K., Bovik, A.C.: Distortion conspicuity on stereoscopically viewed 3D images may correlate to scene content and distortion type. Journal of the Society for Information Display, **21**(11) 491–503 (2014)
17. Chen, M.J., Kwon, D.K., Bovik, A.C.: Study of subject agreement on stereoscopic video quality. In: IEEE Southwest Symposium on Image Analysis and Interpretation, pp. 173–176 (2012)
18. Chen, M.J., Su, C.C., Kwon, D.K., Cormack, L.K., Bovik, A.C.: Full-reference quality assessment of stereopairs accounting for rivalry. Signal Processing: Image Communication **28**(9), 1143–1155 (2013)

19. Chen, W., Fournier, J., Barkowsky, M., Callet, P.L.: New stereoscopic video shooting rule based on stereoscopic distortion parameters and comfortable viewing zone. In: Proceedings SPIE, Stereoscopic Displays and Applications XXII, vol. 7863 (2011)
20. Chen, W., Fournier, J., Barkowsky, M., Callet, P.L.: Quality of experience model for 3DTV. In: Proceedings of SPIE, Stereoscopic Displays and Applications XXIII, vol. 8288 (2012)
21. Craievich, D., Bovik, A.: A stereo VPIC system. In: Proceedings of the IEEE Southwest Symposium on Image Analysis and Interpretation, pp. 149–154 (1996)
22. Cumming, B.G.: An unexpected specialization for horizontal disparity in primate primary visual cortex. Nature **418**(6898), 633–636 (2002)
23. Dabov, K., Foi, A., Katkovnik, V., Egiazarian, K.: Image restoration by sparse 3D transform-domain collaborative filtering. In: Proceedings of SPIE, Image Processing: Algorithms and Systems VI, vol. 6812 (2008)
24. Daly, S.J.: Visible differences predictor: an algorithm for the assessment of image fidelity. In: Proceedings of SPIE, Human Vision, Visual Processing, and Digital Display III, vol. 1666, pp. 2–15 (1992)
25. Daly, S.J., Held, R.T., Hoffman, D.M.: Perceptual issues in stereoscopic signal processing. IEEE Transactions on Broadcasting **57**(2), 347–361 (2011)
26. De Kort, Y.A.W., IJsselsteijn, W.A.: Reality check: the role of realism in stress reduction using media technology. Cyberpsychology & Behavior **9**(2), 230–233 (2006)
27. De Silva, V., Arachchi, H.K., Ekmekcioglu, E., Fernando, A., Dogan, S., Kondoz, A., Savas, S.: Psycho-physical limits of interocular blur suppression and its application to asymmetric stereoscopic video delivery. In: Proceedings of the International Packet Video Workshop, pp. 184–189 (2012)
28. De Silva, V., Arachchi, H.K., Ekmekcioglu, E., Kondoz, A.: Toward an impairment metric for stereoscopic video: a full-reference video quality metric to assess compressed stereoscopic video. IEEE Transactions on Image Processing **22**(9), 3392–3404 (2013)
29. DeAngelis, G.C., Ohzawa, I., Freeman, R.D.: Depth is encoded in the visual cortex by a specialized receptive field structure. Nature **352**(6331), 156–159 (1991)
30. Egiazarian, K., Astola, J., Ponomarenko, N., Lukin, V., Battisti, F., Carli, M.: New full-reference quality metrics based on HVS. In: Proceedings of the Second International Workshop on Video Processing and Quality Metrics, vol. 4 (2006)
31. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient belief propagation for early vision. International Journal of Computer Vision **70**(1), 41–54 (2006)
32. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM **24**(6), 381–395 (1981)
33. Fleet, D.J., Wagner, H., Heeger, D.J.: Neural encoding of binocular disparity: energy models, position shifts and phase shifts. Vision Research **36**(12), 1839–1857 (1996)
34. Goldmann, L., De Simone, F., Ebrahimi, T.: Impact of acquisition distortions on the quality of stereoscopic images. In: Proceedings of the International Workshop on Video Processing and Quality Metrics for Consumer Electronics (2010)
35. Goldmann, L., Simone, F.D., Ebrahimi, T.: A comprehensive database and subjective evaluation methodology for quality of experience in stereoscopic video. In: Proceedings of SPIE, Three-Dimensional Image Processing (3DIP) and Applications, vol. 7526 (2010)
36. Gorley, P., Holliman, N.: Stereoscopic image quality metrics and compression. In: Proceedings of SPIE, Stereoscopic Displays and Applications XIX, vol. 6803 (2008)
37. Ha, K., Kim, M.: A perceptual quality assessment metric using temporal complexity and disparity information for stereoscopic video. In: Proceedings of the IEEE International Conference on Image Processing, pp. 2525–2528 (2011)
38. Hewage, C., Martini, M.: Reduced-reference quality metric for 3D depth map transmission. In: Proceedings of the 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video, pp. 1–4 (2010)
39. Howard, I.P., Rogers, B.J.: Binocular vision and stereopsis. Oxford University Press (1995)
40. Howard, I.P., Rogers, B.J.: Perceiving in Depth. Oxford University Press (2012)

41. Huynh-Thu, Q., Le Callet, P., Barkowsky, M.: Video quality assessment: from 2D to 3D – challenges and future trends. In: Proceedings of the IEEE International Conference on Image Processing, pp. 4025–4028 (2010)
42. International Telecommunication Union (ITU): Subjective video quality assessment methods for multimedia applications. ITU-T Rec. P.910 (2008)
43. International Telecommunication Union (ITU): Methodology for the subjective assessment of the quality of television pictures. ITU-R Rec. BT.500-11 (2009)
44. International Telecommunication Union (ITU): Objective perceptual multimedia video quality measurement of HDTV for digital cable television in the presence of a full reference. ITU-T Rec. J.341 (2011)
45. Jin, L., Boev, A., Gotchev, A., Egiazarian, K.: 3D-DCT based perceptual quality assessment of stereo video. In: Proceedings of the IEEE International Conference on Image Processing, pp. 2521–2524 (2011)
46. Joint Collaborative Team on Video Coding (JCT-VC) of ITU-T VCEG and ISO/IEC MPEG: High Efficient Video Coding (HEVC). ITU-T Rec. H.265 | ISO/IEC 23008-2 HEVC (2013)
47. Joint Video Team (JVT) of ITU-T VCEG and ISO/IEC MPEG: Draft ITU-T Recommendation and Final Draft International Standard of Joint Video Specification. ITU-T Rec. H.264 | ISO/IEC 14496-10 AVC (2003)
48. Julesz, B.: Foundations of Cyclopean Perception. The University of Chicago Press (1971)
49. Jumisko-Pyykkö, S., Haustola, T., Boev, A., Gotchev, A.: Subjective evaluation of mobile 3D video content: depth range versus compression artifacts. In: Proceedings of SPIE, Multimedia on Mobile Devices 2011 and Multimedia Content Access: Algorithms and Systems V, vol. 7881 (2011)
50. Kaptein, R.G., Kuijsters, A., Lambooij, M.T.M., IJsselsteijn, W.A., Heynderickx, I.: Performance evaluation of 3D-TV systems. In: Proceedings of SPIE, Image Quality and System Performance V, vol. 6808 (2008)
51. Kim, D., Sohn, K.: Visual fatigue prediction for stereoscopic image. IEEE Transactions on Circuits and Systems for Video Technology **21**(2), 231–236 (2011)
52. Kolmogorov, V., Zabih, R.: Multi-camera scene reconstruction via graph cuts. In: Proceedings of the 7th European Conference on Computer Vision, vol. 2352, pp. 82–96 (2002)
53. Lambooij, M., IJsselsteijn, W., Bouwhuis, D.G., Heynderickx, I.: Evaluation of stereoscopic images: beyond 2d quality. IEEE Transactions on Broadcasting **57**(2), 432–444 (2011)
54. Lambooij, M., IJsselsteijn, W., Fortuin, M., Heynderickx, I.: Visual discomfort and visual fatigue of stereoscopic displays: A review. Journal of Imaging Science and Technology **53**(3), 1–14 (2009)
55. Levelt, W.J.M.: On binocular rivalry, vol. 2. Mouton, The Hague (1968)
56. Liu, Y., Cormack, L.K., Bovik, A.C.: Statistical modeling of 3-D natural scenes with application to bayesian stereopsis. IEEE Transactions on Image Processing **20**(9), 2515–2530 (2011)
57. López, J.P., Rodrigo, J.A., Jiménez, D., Menéndez, J.M.: Stereoscopic 3D video quality assessment based on depth maps and video motion. EURASIP Journal on Image and Video Processing **2013**(1), 1–14 (2013)
58. Lowe, D.: Object recognition from local scale-invariant features. In: Proceedings of the IEEE International Conference on Computer Vision, vol. 2, pp. 1150–1157 (1999)
59. Maalouf, A., Larabi, M.C.: CYCLOP: a stereo color image quality assessment metric. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1161–1164 (2011)
60. Marquardt, D.W.: An algorithm for least-squares estimation of nonlinear parameters. Journal of the Society for Industrial & Applied Mathematics **11**(2), 431–441 (1963)
61. Meegan, D.V., Stelmach, L.B., Tam, W.J.: Unequal weighting of monocular inputs in binocular combination: Implications for the compression of stereoscopic imagery. Journal of Experimental Psychology: Applied **7**(2), 143–153 (2001)

62. Meesters, L.M.J., IJsselsteijn, W.A., Seuntiens, P.J.H.: A survey of perceptual evaluations and requirements of three-dimensional TV. IEEE Transactions on Circuits and Systems for Video Technology **14**(3), 381–391 (2004)
63. Menz, M.D., Freeman, R.D.: Stereoscopic depth processing in the visual cortex: a coarse-to-fine mechanism. Nature Neuroscience **6**(1), 59–65 (2002)
64. Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. IEEE Transactions on Image Processing **21**(12), 4695–4708 (2012)
65. Mittal, A., Moorthy, A.K., Ghosh, J., Bovik, A.C.: Algorithmic assessment of 3D quality of experience for images and videos. In: Proceedings of the IEEE Digital Signal Processing Workshop and IEEE Signal Processing Education Workshop, pp. 338–343 (2011)
66. Mittal, A., Soundararajan, R., Bovik, A.: Making a “completely blind" image quality analyzer. IEEE Signal Processing Letters **20**(3), 209–212 (2013)
67. Moorthy, A., Seshadrinathan, K., Soundararajan, R., Bovik, A.: Wireless video quality assessment: A study of subjective scores and objective algorithms. IEEE Transactions on Circuits and Systems for Video Technology **20**(4), 587–599 (2010)
68. Moorthy, A., Su, C.C., Chen, M.J., Mittal, A., Cormack, L.K., Bovik, A.C.: LIVE 3D Image Quality Database Phase I and Phase II. http://live.ece.utexas.edu/research/quality/live_3dimage.html
69. Moorthy, A.K., Bovik, A.C.: Blind image quality assessment: From natural scene statistics to perceptual quality. IEEE Transactions on Image Processing **20**(12), 3350–3364 (2011)
70. Moorthy, A.K., Bovik, A.C.: Visual quality assessment algorithms: What does the future hold? Multimedia Tools and Applications **51**(2), 675–696 (2011)
71. Moorthy, A.K., Bovik, A.C.: A survey on 3D quality of experience and 3D quality assessment. In: Proceedings of SPIE, Human Vision and Electronic Imaging XVIII, vol. 8651 (2013)
72. Moorthy, A.K., Su, C.C., Mittal, A., Bovik, A.C.: Subjective evaluation of stereoscopic image quality. Signal Processing: Image Communication **28**(8), 870–883 (2013)
73. Motion Picture Association of America (MPAA): Theatrical market statistics. http://www.mpaa.org/policy/industry (2012)
74. Ohzawa, I., Freeman, R.D.: The binocular organization of complex cells in the cat’s visual cortex. Journal of Neurophysiology **56**(1), 243–259 (1986)
75. Ohzawa, I., Freeman, R.D.: The binocular organization of simple cells in the cat’s visual cortex. Journal of Neurophysiology **56**(1), 221–242 (1986)
76. Okada, Y., Ukai, K., Wolffsohn, J.S., Gilmartin, B., Iijima, A., Bando, T.: Target spatial frequency determines the response to conflicting defocus- and convergence-driven accommodative stimuli. Vision Research **46**(4), 475–484 (2006)
77. Olshausen, B., Field, D.: Natural image statistics and efficient coding. Network: Computation in Nerual Systems **7**(2), 333–339 (1996)
78. Olshausen, B.A., Field, D.J.: Vision and the coding of natural images. American Scientist **88**, 238–245 (2000)
79. Park, J., Lee, S., Bovik, A.C.: 3D visual discomfort prediction: vergence, foveation, and the physiological optics of accommodation. IEEE Journal of Selected Topics in Signal Processing, **8**(3), 415–427 (2014)
80. Park, J., Oh, H., Lee, S.: IEEE Standards Association Stereo Image Database. http://grouper.ieee.org/groups/3dhf/
81. Pinson, M.H., Wolf, S.: A new standardized method for objectively measuring video quality. IEEE Transactions on Broadcasting **50**(3), 312–322 (2004)
82. Ponomarenko, N., Silvestri, F., Egiazarian, K., Carli, M., Astola, J., Lukin, V.: On between-coefficient contrast masking of DCT basis functions. In: Proceedings of the Third International Workshop on Video Processing and Quality Metrics, vol. 4 (2007)
83. Potetz, B., Lee, T.S.: Statistical correlations between two-dimensional images and three-dimensional structures in natural scenes. Journal of the Optical Society of America A **20**(7), 1292–1303 (2003)

84. Puri, A., Kollarits, R.V., Haskell, B.G.: Basics of stereoscopic video, new compression results with MPEG-2 and a proposal for MPEG-4. Signal Processing: Image Communication **10**(1), 201–234 (1997)
85. Read, J.: Early computational processing in binocular vision and depth perception. Progress in Biophysics and Molecular Biology **87**(1), 77–108 (2005)
86. Rosenholtz, R., Watson, A.B.: Perceptual adaptive JPEG coding. In: Proceedings of the IEEE International Conference on Image Processing, vol. 1, pp. 901–904 (1996)
87. Ruderman, D.L.: The statistics of natural images. Network: Computation in Neural Systems **5**(4), 517–548 (1994)
88. Ryu, S., Kim, D.H., Sohn, K.: Stereoscopic image quality metric based on binocular perception model. In: Proceedings of the IEEE International Conference on Image Processing, pp. 609–612 (2012)
89. Saad, M., Bovik, A., Charrier, C.: Blind image quality assessment: A natural scene statistics approach in the DCT domain. IEEE Transactions on Image Processing **21**(8), 3339–3352 (2012)
90. Sazzad, Z., Akhter, R., Baltes, J., Horita, Y.: Objective no-reference stereoscopic image quality prediction based on 2D image features and relative disparity. Advances in Multimedia **2012**(8), 1–16 (2012)
91. Sazzad, Z.P., Yamanaka, S., Kawayokeita, Y., Horita, Y.: Stereoscopic image quality prediction. In: Proceedings of the International Workshop on Quality of Multimedia Experience, pp. 180–185 (2009)
92. Scharstein, D.: Middlebury stereo datasets. http://vision.middlebury.edu/stereo/data/
93. Seshadrinathan, K., Soundararajan, R., Bovik, A., Cormack, L.: Study of subjective and objective quality assessment of video. IEEE Transactions on Image Processing **19**(6), 1427–1441 (2010)
94. Seshadrinathan, K., Soundararajan, R., Bovik, A.C., Cormack, L.K.: LIVE Video Quality Database. http://live.ece.utexas.edu/research/quality/live_video.html
95. Seuntiens, P., Meesters, L., Ijsselsteijn, W.: Perceived quality of compressed stereoscopic images: effects of symmetric and asymmetric JPEG coding and camera separation. ACM Transactions on Applied Perception **3**(2), 95–109 (2006)
96. Sheikh, H., Bovik, A.: Image information and visual quality. IEEE Transactions on Image Processing **15**(2), 430–444 (2006)
97. Sheikh, H.R., Sabir, M.F., Bovik, A.C.: A statistical evaluation of recent full reference image quality assessment algorithms. IEEE Transactions on Image Processing **15**(11), 3440–3451 (2006)
98. Sheikh, H.R., Wang, Z., Cormack, L.K., Bovik, A.C.: LIVE Image Quality Assessment Database. http://live.ece.utexas.edu/research/quality/subjective.htm
99. Simoncelli, E.P., Olshausen, B.A.: Natural image statistics and neural representation. Annual Review of Neuroscience **24**(1), 1193–1216 (2001)
100. Soundararajan, R., Bovik, A.: RRED indices: Reduced reference entropic differencing for image quality assessment. IEEE Transactions on Image Processing **21**(2), 517–526 (2012)
101. Su, C.C., Cormack, L.K., Bovik, A.C.: Color and depth priors in natural images. IEEE Transactions on Image Processing **22**(6), 2259 – 2274 (2013)
102. Su, C.C., Cormack, L.K., Bovik, A.C.: Bivariate statistical modeling of color and range in natural scenes. In: Proceedings of SPIE, Human Vision and Electronic Imaging XIX, vol. 9014 (2014)
103. Tam, W.J., Speranza, F., Yano, S., Shimono, K., Ono, H.: Stereoscopic 3D-TV: Visual comfort. IEEE Transactions on Broadcasting **57**(2), 335–346 (2011)
104. Tang, H., Joshi, N., Kapoor, A.: Learning a blind measure of perceptual image quality. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 305–312 (2011)
105. Tovée, M.J.: An introduction to the visual system. Cambridge University Press (1996)

106. Urvoy, M., Barkowsky, M., Cousseau, R., Koudota, Y., Ricorde, V., Le Callet, P., Gutiérrez, J., García, N.: NAMA3DS1-COSPAD1: Subjective video quality assessment database on coding conditions introducing freely available high quality 3D stereoscopic sequences. In: Proceedings of the International Workshop on Quality of Multimedia Experience, pp. 109–114 (2012)
107. Wang, X., Yu, M., Yang, Y., Jiang, G.: Research on subjective stereoscopic image quality assessment. In: Proceedings of SPIE, Multimedia Content Access: Algorithms and Systems III, vol. 7255 (2009)
108. Wang, Z., Bovik, A.C.: A universal image quality index. IEEE Signal Processing Letters **9**(3), 81–84 (2002)
109. Wang, Z., Bovik, A.C.: Modern image quality assessment. Synthesis Lectures on Image, Video, and Multimedia Processing **2**(1), 1–156 (2006)
110. Wang, Z., Bovik, A.C.: Mean squared error: love it or leave it? a new look at signal fidelity measures. IEEE Signal Processing Magazine **26**(1), 98–117 (2009)
111. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004)
112. Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multiscale structural similarity for image quality assessment. In: Conference Record of the Thirty-Seventh Asilomar Conference on Signals, Systems and Computers, 2003, vol. 2, pp. 1398–1402 (2003)
113. Western Ophthalmics Corporation: Stereo Randot Test. http://www.west-op.com/stereorandot.html
114. Westin, C.F.: Extracting brain connectivity from diffusion MRI [life sciences]. IEEE Signal Processing Magazine **24**(6), 124–152 (2007)
115. William, A.M., Bailey, D.L.: Stereoscopic visualization of scientific and medical content. In: ACM SIGGRAPH 2006 Educators Program, 26 (2006)
116. Winkler, S.: Image and video quality resources. http://stefan.winkler.net/resources.html
117. Winkler, S.: Analysis of public image and video databases for quality assessment. IEEE Journal of Selected Topics in Signal Processing **6**(6), 616–625 (2012)
118. Yasakethu, S.L.P., Hewage, C.T.E.R., Fernando, W., Kondoz, A.: Quality analysis for 3D video using 2D video quality models. IEEE Transactions on Consumer Electronics **54**(4), 1969–1976 (2008)
119. Ye, P., Doermann, D.: No-reference image quality assessment using visual codebooks. IEEE Transactions on Image Processing **21**(7), 3129–3138 (2012)
120. You, J., Xing, L., Perkis, A., Wang, X.: Perceptual quality assessment for stereoscopic images based on 2D image quality metrics and disparity analysis. In: Proceedings of the International Workshop on Video Processing and Quality Metrics (2010)
121. Zwicker, M., Yea, S., Vetro, A., Forlines, C., Matusik, W., Pfister, H.: Display pre-filtering for multi-view video compression. In: Proceedings of the 15th International Conference on Multimedia, pp. 1046–1053 (2007)

Chapter 8
Retargeted Image Quality Assessment: Current Progresses and Future Trends

Lin Ma, Chenwei Deng, Weisi Lin, King Ngi Ngan, and Long Xu

8.1 Introduction

Nowadays, the diversity and versatility of the display devices have imposed new demands on digital image processing. The same image needs to be displayed with different resolutions on various devices. The image retargeting approaches [1–15] have been proposed to adapt the source images into arbitrary sizes and simultaneously keep the salient content of the source images. These developed retargeting methods, such as warp [2], seam carving [4–6], and multi-operator [7], try to preserve the salient shape and content information of the source image, and shrink (or expend) the unimportant regions of the image into the given resolution. With such approaches, the images can be displayed on different screens of different resolutions, which will provide better visual experiences for human viewers.

L. Ma (✉)
Huawei Noah's Ark Lab, Hong Kong, China
e-mail: forest.linma@gmail.com

C. Deng
School of Information and Electronics, Beijing Institute of Technology, Beijing, China
e-mail: cwdeng@bit.edu.cn

W. Lin
School of Computer Engineering, Nanyang Technological University, Singapore
e-mail: wslin@ntu.edu.sg

K.N. Ngan
Department of Electronic Engineering, The Chinese University of Hong Kong,
Hong Kong, China
e-mail: knngan@ee.cuhk.edu.hk

L. Xu
National Astronomical Observatories, Chinese Academy of Sciences, Beijing, China
e-mail: lxu@bao.ac.cn

C. Deng et al. (eds.), *Visual Signal Quality Assessment*,
DOI 10.1007/978-3-319-10368-6_8

However, for most of these methods, a simple visual comparison was conducted for the results (comparing the results of different retargeting methods based on a small set of images) to demonstrate the effectiveness of the retargeting methods, which is not suitable for online processing. In order to obtain a retargeted image of good quality, quality assessment of retargeted images is needed for maximize the perceptual quality to guide the retargeting process. Therefore, a new challenge of objectively evaluating the retargeted image perceptual quality is issued, where variant resolutions may be presented, the objective shape may be distorted, and some content information may be discarded.

As human eyes are the final receivers of the retargeted images, the human subjective opinion is the most reliable value to indicate the image perceptual quality. The subjective opinions are obtained through the subjective testing, where a large number of viewers participate in the subjective test and provide their personal opinions of the image quality on some pre-defined scale. By processing the obtained subjective scores, each image will be assigned a score indicating its perceptual quality. The subjective testing method is time-consuming and expensive, which makes it impractical for most image applications. However, the subjective rating obtained can be recognized as the ground truth of the image perceptual quality. Therefore, the subjective rating scores can be employed to validate the objective quality metrics. Subjective studies can also enable the improvement in the performance of the quality metric towards attaining the ultimate goal of matching human perception. Then the developed quality metric can be utilized to guide the corresponding application. Furthermore, the subjective studies can also benefit the image applications for better perceptual quality experience, specifically improving the perceptual quality of the retargeted image. Consequently there is a great demand of image retargeting database for both subjective rating score acquisition and objective quality metric validation.

Objective quality assessment is demanded for not only automatically evaluate the perceptual quality of retargeted images/videos but also can help guiding the retargeting process to make it more efficiently and visually plausible. However, the retargeted image quality assessment is of great challenge, where the source and retargeted images cannot be well matched or aligned for the quality analysis. Specifically, the content information discarding and shape distortion cannot ensure a good pixel matching between the retargeted and source images. Therefore, the perceptual quality analysis requires a high-level understanding of the retargeted image. Evaluating the retargeted image quality makes it more difficult, as the subjective opinions may not be able to characterize the perceptual quality of the images. During the subjective evaluation process [16, 17], it is demonstrated that the participants have difficulties to achieve an agreement. Clearly, the participants present different appreciations of perceptual qualities of the retargeted images. In spite of these difficulties, many research works have been done to develop an automatic quality metric to evaluate retargeted image quality. In this chapter, the authors will briefly review recent progresses of the image retargeting quality assessment, in terms of both subjective and objective measurements. The chapter is organized in the following. Section 8.2 will briefly introduce the retargeting methods for visual signals. And the subjective approaches for evaluating retargeted image are detailed in Sect. 8.3, specifically the CUHK [16] and RetargetMe [17] retargeting databases. Section 8.4 will review recent works on objective retargeting quality metrics. Finally, future trends will be discussed in Sect. 8.5 followed by the conclusions.

8.2 Retargeting Methodologies for Visual Signals

Nowadays, many approaches for retargeting visual signals have been developed [1–15]. The media is retargeted to adapt the displaying resolutions covering images, videos, stereo images, and so on. These methods are employed to generate the corresponding retargeted images/videos for subjective testing, which constitute the subjective quality databases in Sect. 8.3. The algorithms are briefly introduced in the following.

- Cropping: manually cropping the source image to the target size for the best salient information preservation.
- Scaling: simple scaling the source image into the target size.
- Seam carving [4–6]: removing the contiguous chains of pixels that lie in the regions of the smallest gradient magnitude values in the source image. The seams removed are optimized by a dynamic programming approach.
- Optimized seam carving and scale [15]: a measurement named as "seam carving distance" is proposed to measure the similarity of retargeted image and the source one. The method employed the measurement to optimally combine the scaling and seam carving methods.
- Non-homogeneous retargeting [2]: a warping function is optimized to find the optimal squeezed image by reducing the image width. In order to prevent the salient content of the image from shape degradation, the gradient magnitude and face detection results are employed to determine the saliency region.
- Scale and stretch [8]: an objective function is optimized by uniformly scaling the salient regions to preserve the shape information. The saliency map is detected by combining the gradient magnitude and the saliency map detected by Itti et al. [13].
- Shift-map editing [9]: graph cut is used to remove an entire object at a time rather than a seam. The color difference and gradient information is employed to ensure the smoothness.
- Multi-operator process [7]: seam carving, scaling, and cropping are combined together to generate the retargeted image. And a bi-directional warping measurement determines how to combine these operators.
- Energy-based deformation [14]: similar as the scale and stretch method, warping is also used to produce the retargeting image.

- Streaming video [3]: the warping method is also used. The saliency map is generated by combining the visual attention map, the line detection, and important objects.
- Shift-map stereo image retargeting [10]: in the context of 3D image retargeting, the novel viewpoint advocated is that the geometric consistency in the form of preserving disparity values should not be an overpowering objective formulated as hard constraints. Instead, for maximizing viewing experience and comfort, it is desirable to simultaneously retarget the images as well as adjust the disparity values. The method is developed based on the methods of shift-map and importance filtering.
- Stereo image retargeting [11, 12]: the visual distortion in each of the images is minimized as well as the depth distortion. A key property is to take into account the visibility relations between pixels in the image pair (occluded and occluding pixels). As a result, the retargeted pair is geometrically consistent with a feasible 3D scene, similar to the original one.

Referring to these retargeting methods, it can be observed that the cropping, scaling, seam carving, and warping are the basic tools for image retargeting. Many research works are proposed to combine these tools together by optimizing a defined objective measurement, such as [7]. As the foreground objects, including the faces and people, represent the most salient information to the human viewers, the retargeting methods need to prevent the objects from shape distortion by referring to the saliency map.

8.3 Subjective Approaches for Retargeted Image Quality Assessment

Until now, there are two public subjective databases focusing on the quality evaluation of image retargeting, specifically the RetargetMe [17] built by Rubinstein M. et al. and CUHK retargeting database [16] built by Ma L. et al. The two databases are built for different purposes. RetargetMe database concentrates on a comparative study of existing retargeting methods. The authors compared which retargeting method generates the retargeted image with the highest perceptual quality. The subjective test is performed in a pair comparison way, where the participants are shown two retargeted images at a time, side by side, and are asked to simply choose the one is of better quality. The RetargetMe database consists of the retargeted image and its number of times that the retargeted image is favored over another one. CUHK retargeting database targets at perceptual quality evaluation of the retargeted images. Therefore, each retargeted image is presented to the participants against its original form. With the source image as the reference, the perceptual quality of each retargeted image has been subjectively rated on a pre-defined scale. After processing the subjective ratings, the mean opinion score (MOS) value and the corresponding standard deviation are generated for each retargeted image.

For the RetargetMe database [17], the Kendall τ distance [19] is employed to validate different quality metrics by measuring the correlation between two rankings. CUHK retargeting database mainly focuses on evaluating perceptual quality of the retargeted images other than pair-wise comparing the retargeting methods [17]. Therefore, based on CUHK retargeting database, the objective quality metrics can be evaluated in the standardized way [23]. Same as traditional image/video quality assessments, multiple retargeting databases are needed. When constructing different databases, different subjects participated in the subjective testing with different rating scales. Meanwhile, the source image content and image distortions introduced by retargeting methods are quite different. In these respects, the subjective quality databases can be ensured to be of great diversity, which can be employed to evaluate the effectiveness and robustness of the developed objective quality metric. Therefore, CUHK retargeting [16] and RetargetMe [17] databases can be further viewed as complementary to each other.

8.3.1 RetargetMe Database

8.3.1.1 RetargetMe Database Construction

Source Image

Content-aware retargeting methods work best on images where some content can be disposed of. These insensitive contents include either smooth or irregularly textured areas such as sky, water, grass, or trees. As human eyes are insensitive to these contents, most retargeting methods work pretty well. Challenge is posed in images containing either dense information or global and local structures that may be distorted during resizing. To create the RetargetMe database, images generated from various retargeting methods are collected. A set of image attributes are selected by referring to the three major retargeting objectives (preserving content, preserving structure, and preventing artifacts). These attributes are: people and faces, lines and/or clear edges, evident foreground objects, texture elements or repeating patterns, specific geometric structures, and symmetry. The source images for building RetargetMe database are made up of 80 image, each of which has one or more of these attributes. Some of source images are illustrated in Fig. 8.1.

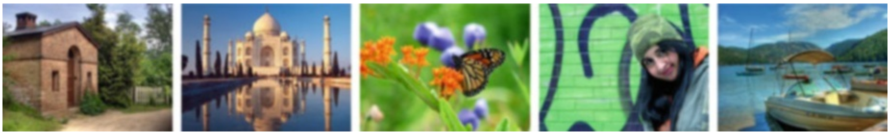


Fig. 8.1 Samples of the images used in RetargetMe database

The resolution change is only restricted to one dimension, either the width or the height of the image. Furthermore, reduction in image size is mostly concentrated during the database construction. In consequence, the authors chose to use considerable resizing (25 % or 50 %). Each method retargets the images into 50 % and 75 % of its original resolution.

Subjective Testing

RetargetMe [17] database focuses on comparing the performances of different retargeting methods. Therefore, during the subjective testing phrase, a pair comparison manner is employed (with and without the source image), where the participants are shown two retargeted images at a time, with the present of the source image, and determine which one they like better. The subjective testing is performed via Internet.[1] By referring to the source image, the participants are asked to choose which one of the two retargeted images are of better quality. Detailed information can be found in [17].

8.3.1.2 Subjective Analysis

For RetargetMe database, a total of 210 participants took part in the test, generating a total of 9,324 pair-wise votes. Half of the participants were volunteers and half workers from Amazon Mechanical Turk. About 40 % were females and 60 % males, average age was around 30, and they had varying degrees of computer graphics knowledge, being naive as to the design and goals of the experiment. To investigate whether the presence of the source image affects the preferred resized result, the authors also conducted a blind version of the exact same test (with 210 new participants), where the source image was not present.

The similarity of choices between participants is studied. A complete agreement means that all the participants voted in the same way. High disagreement, on the other hand, reflects difficulty in making choices, suggesting that the subjective views have different opinions on the perceptual quality of the image. For this purpose, Kendall and Babington–Smith introduced the coefficient of agreement [18].

$$\mu = \frac{2\sum}{\binom{m}{2}\binom{t}{2}} - 1,\ where\ \sum = \sum_{i=1}^{t}\sum_{j=1}^{t}\binom{a_{ij}}{2} \tag{8.1}$$

where a_{ij} is the number of times that method i was chosen over method j, m is the number of participants, and $t = 8$ is the number of retargeting methods tested. If complete agreement is achieved, then $\mu = 1$; the minimum value of μ is attained

[1] http://people.csail.mit.edu/mrub/retargetme/survey/index.php?mode=0.

by an even distribution of answers and is given by $\mu = -1/m$. The coefficient over all images is $\mu = 0.095$, a relatively low value suggesting that the participants in general had difficulty judging the perceptual quality of retargeted image. More detailed information can be found in [17].

8.3.2 *CUHK Retargeting Database*

8.3.2.1 CUHK Retargeting Database Construction

Source Image

The source image composing the CUHK retargeting database contains the frequently encountered attributes, such as the face and people, clear foreground object, natural scenery (containing smooth or texture region), and geometric structure (evident lines or edges). The detailed information of the attributes can be referred to [16]. There are total 57 source images for retargeting by different methods as introduced in Sect. 8.2. The corresponding resolutions of source images are variant, in order to alleviate the influence of the image resolution on the subjective testing. Figure 8.2 illustrates some samples of the source images. The source images are roughly categorized into four classes according to the aforementioned attributes. The attribute information of each source image can be found in [16]. The content-aware retargeting methods make that the perceptual qualities of retargeted results

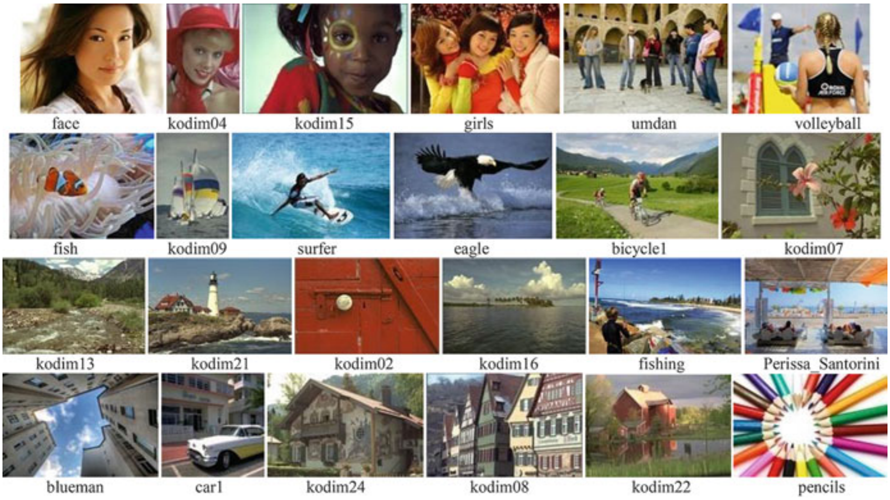


Fig. 8.2 Samples of the source images utilized in the subjective testing for CUHK retargeting database construction. The images in the *top row* mostly contain the attribute of face and people; the images in the *second row* mostly contain the attribute of clear foreground object; the images in the *third row* mostly contain the attribute of natural scenery; the images in the *bottom row* mostly contain the attribute of geometric structure

from different source images are different. The attributes of the images are critical to the perceptual quality of the final retargeted images. Human eyes are very sensitive to the distortion of the faces and geometric structures. However, they can tolerate more distortions on the natural scenery, especially for the texture regions. By including the images with different attributes, we can further analyze how the image content affect the perceptual quality of the retargeted images.

The resolution changes are restricted in only one dimension. The retargeting methods change the resolution of the source images in either the width or height dimension. CUHK retargeting database retargets images in two ratios, shrinking the image to 75 % and 50 %. Three retargeted results of each source image are included. These three retargeted images may be produced by different retargeting methods in different scales. During the subjective testing, the retargeted images with different scales are mixed together to examine its perceptual quality through subjective testing.

Subjective Testing

Both the shape distortion and content information loss of the source image affect the perceptual quality of the retargeted image. Therefore, in order to provide more convincing results, the source image needs to be presented to the subjective viewers as the reference simultaneously during the subjective testing process of CUHK retargeting database. Without the source image as the reference, the viewers are not able to detect the discarded information, which may be the most important part of the source image. Therefore, the simultaneous double stimulus for continuous evaluation (SDSCE) is employed [26, 29] for subjective evaluation.

Two images are juxtaposed on the screen for the human subject. One is the source image for reference and the other is the retargeted image to be evaluated. The participants are aware of which one is the reference image and which one is the retargeted image. The participants are asked to compare the difference between the two images and judge the perceptual quality of the retargeted one. After that, the perceptual quality index of the retargeted image is ranked by the participant. The only difference of the subjective testing in this work was the use of the ITU-R absolute category rating (ACR) scale rather than a continuous scale. The ACR scale employs a five-category discrete quality judgment, as described in [28–30].

The user interface for the subjective testing is designed as shown in Fig. 8.3. The two images, including the source and the retargeted one, are loaded into the memory before displaying. In order to avoid strong visual contrast, the remaining regions of the display area are gray (the pixel values are set equal to 128). The quality scales are labeled to help the human subjects to rate the image quality. The quality scales are labeled as “Bad,” “Poor,” “Fair,” “Good,” and “Excellent,” which range from the lowest to the highest perceptual quality index. During the subjective evaluation, the subjective values are recorded in numerical values. As shown in Fig. 8.3, the “Bad” corresponds to 1 and the “Excellent” corresponds to 5. Therefore, for the obtained subjective ratings, larger values indicate better perceptual qualities of the retargeted

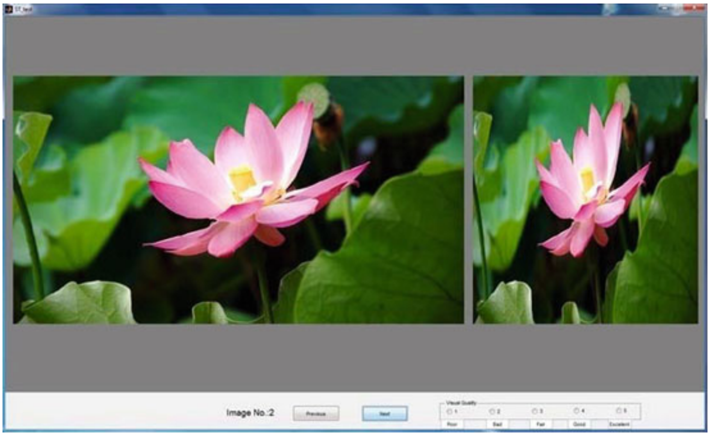


Fig. 8.3 Screenshot of the subjective study interface displaying the images to the human subject

images. The participants select the appropriate quality index based on their own judgement. After choosing the quality of one image, the participant can proceed to the next image.

In order to reduce the effect of the viewer fatigue, the 171 retargeted images are divided into two sessions. The first session contains 69 images, while the second one contains 102 images. For each session, it will take the viewer about 10–20 min to accomplish the subjective testing. The order of the image pairs (the source image and the retargeted image) is randomly arranged, which varies for each participant. Furthermore, in order to avoid the contextual and memory effects on the participants' judgment of the quality, the retargeted images which are generated from the same source image will not be presented consecutively. In order to prevent the scaling effect, which is critical to the image retargeting results, the source image and the retargeted image must be displayed in their native resolution.

All the subjects participating in the subjective testing are the students have normal vision (with or without corrective glasses) and have passed the color blindness test. For the first session, viewers participate in the subjective testing process, with 15 viewers are experts in image processing. And each image in the second session was rated by 34 participants, with 18 viewers are experts in image processing.

8.3.2.2 Processing of Subjective Ratings

Subjective Agreement

For CUHK retargeting database, the quartiles of the subjective scores for each image is employed to analyze the subject agreement, which is illustrated in Fig. 8.4. The lower and higher bound of the red error bar denotes the 25th and 75th

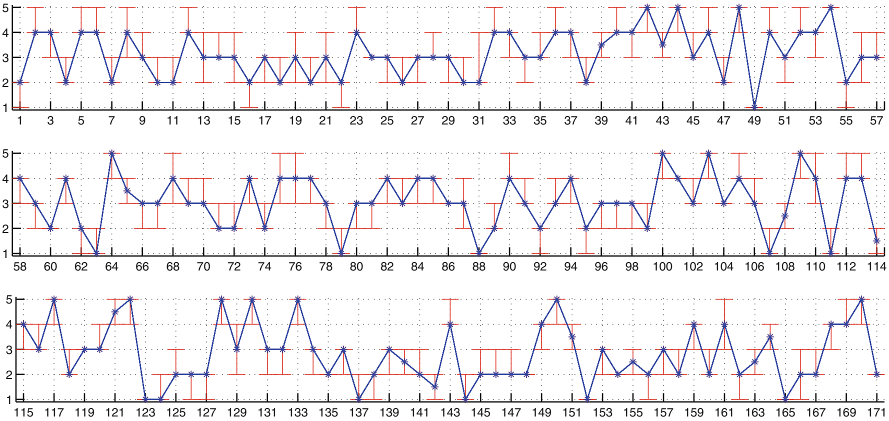


Fig. 8.4 The subjective scores for each image (the horizontal axes correspond to the image number, and the vertical axes correspond to the subjective scores of the viewers. The *blue asterisk* indicates the median value among all the viewers. And the *red error bar* indicates the corresponding 25th and 75th percentiles of the subjective scores)

percentiles of subjective ratings obtained for each image. After sorting the subjective scores, the central 50 % of subject ratings lie within the range. The blue asterisk indicates the median value of the subjective scores. The detailed information of the image number and the corresponding retargeted image name can be found in the [31, 32]. An outlier coefficient (OC) is introduced to evaluate the subjective agreement:

$$OC = \frac{N_{outlier}}{N_{total}} \tag{8.2}$$

where N_{total} indicates the total number of the retargeted images in the database, and $N_{outlier}$ denotes the number of the images, which are regarded as the outlier. If the interval between the higher bound and lower bound error bar in Fig. 8.4 is larger than 1, the image is recognized as outlier. The reason is that viewers may have different opinions on the image quality, but they should at least have the similar judgments. For one image, different viewers may rate "Good" or "Excellent," which are neighboring values. In most cases, the same image will not be scored with greatly differences, such as "Poor" or "Good." Therefore, if the central 50 % subjective ratings are constrained within the interval of 1, the participants have arrived at an agreement of the retargeted image quality. For the constructed database, 15 out of 171 are recognized as the outlier images, which implies $OC = 8.77\,\%$. Therefore, 91.2 % of the images in the database have shown the agreement among participants. Therefore, the images in the database will be rated similarly if subjectively tested by the other viewers. Consequently, these images can be included to construct the database and further employed for validation of the quality metrics.

Screening of the Observers

In the previous section, the subjective ratings of the images have demonstrated high subject agreements. However, in order to obtain the final MOS and standard deviation value for each image, the subject rejection process is suggested by [29]. Let S_{ijk} denote the subjective rating by the subject i to the retargeted image j in session $k = 1, 2$. The S_{ijk} values are firstly converted to Z-scores per session [27]:

$$
\begin{aligned}
\mu_{ik} &= \frac{1}{N_{ik}} \sum_{j=1}^{N_{ik}} S_{ijk} \\
\sigma_{ik} &= \sqrt{\tfrac{1}{N_{ik}-1} \sum_{j=1}^{N_{ik}} \left(S_{ijk} - \mu_{ik}\right)^2} \\
z_{ijk} &= \frac{S_{ijk} - \mu_{ik}}{\sigma_{ik}}
\end{aligned}
\tag{8.3}
$$

where N_{ik} is the number of the test images evaluated by the subject i in session k. It is noted that Z-scores are obtained per session, which account for any differences in subject preferences for the reference images, and different participants between sessions.

After converting the obtained subjective ratings into Z-scores, the subject rejection procedure specified in the ITU-R BT 500.11 [29] is then used to reject the unreliable viewers. The converting process and subject rejection procedure used should be superior to the VQEG studies [20–22]. The mean value μ_{jk} and the variance value σ_{jk} are firstly computed for each image by accounting for the differences of the subjective viewers. Then the kurtosis β_j of the assigned scores is computed to determine whether the scores are normally distributed:

$$
\begin{aligned}
\mu_{jk} &= \frac{1}{N_{jk}} \sum_{i=1}^{N_{jk}} S_{ijk} \\
\sigma_{jk} &= \sqrt{\tfrac{1}{N_{jk}-1} \sum_{j=1}^{N_{jk}} \left(S_{ijk} - \mu_{jk}\right)^2} \\
\beta_j &= \frac{m_4}{(m_2)^2} \;\; with \;\; m_\Delta = \frac{\sum_{j=1}^{N_{ik}} \left(S_{ijk} - \mu_{jk}\right)^\Delta}{N_{jk}}
\end{aligned}
\tag{8.4}
$$

If the kurtosis value β_j falls between 2 and 4, the scores are regarded as normally distribution. The subject rejection procedure is detailed in Fig. 8.5. By performing the procedure, 1 out of 30 participants and 3 out of 34 participants are rejected in subjective session 1 and 2, respectively.

Fig. 8.5 Detailed algorithm of the subject rejection process

```
For each subject $i$, find the $P_{ik}$ and $Q_{ik}$
if $2 \leq \beta_j \leq 4$ (normally distributed)
    if $S_{ijk} \geq \mu_{jk} + 2\sigma_{jk}$, then $P_{ik} = P_{ik} + 1$;
    if $S_{ijk} \leq \mu_{jk} - 2\sigma_{jk}$, then $Q_{ik} = Q_{ik} + 1$;
else
    if $S_{ijk} \geq \mu_{jk} + \sqrt{20}\sigma_{jk}$, then $P_{ik} = P_{ik} + 1$;
    if $S_{ijk} \leq \mu_{jk} - \sqrt{20}\sigma_{jk}$, then $Q_{ik} = Q_{ik} + 1$;
if $\frac{P_{ik}+Q_{ik}}{N_{jk}} > 0.05$ and $\frac{P_{ik}-Q_{ik}}{P_{ik}+Q_{ik}} < 0.3$, then **REJECT** the subject $i$.
```

After subject rejection, Z-scores are then linearly rescaled to lie in the range of $[0, 100]$. Assuming that the Z-scores assigned by a subject are distributed as a standard Gaussian [24, 25], 99 % of the scores will lie in the range $[-3, +3]$. Re-scaling is accomplished by linearly mapping the range $[-3, +3]$ to $[0, 100]$ by:

$$\tilde{Z}_{ijk} = \frac{100(z_{ijk}+3)}{6} \tag{8.5}$$

Finally, the MOS value of each retargeted image is computed as the mean of the rescaled Z-scores, together with the standard deviation:

$$MOS_{jk} = \frac{1}{M_k} \sum_{i=1}^{M_k} \tilde{Z}_{ijk} \tag{8.6}$$

$$std_{jk} = \sqrt{\frac{1}{M_k - 1} \sum_{i=1}^{M_k} \left(\tilde{Z}_{ijk} - MOS_{jk}\right)^2}$$

where M_k is the number of remaining subjects of session k after the subject rejection. The MOS value together with the standard deviation is recorded for each retargeted image as the ground truth indicating the retargeted image perceptual quality. They can be further analyzed and used for evaluating the performances of the quality metrics. The final subjective scores after conversion, with the standard deviation indicating the error bar, are illustrated in Fig. 8.6.

As aforementioned, the perceptual qualities of the retargeted images in the database should span the entire range of visual quality and exhibit good perceptual quality separation [25]. The histogram of the MOS values is shown in Fig. 8.7. It can be observed that the image perceptual qualities range from low to high values. Also it demonstrates that the subjective study samples a range of perceptual quality in an approximately uniform fashion, which results in a good separation.

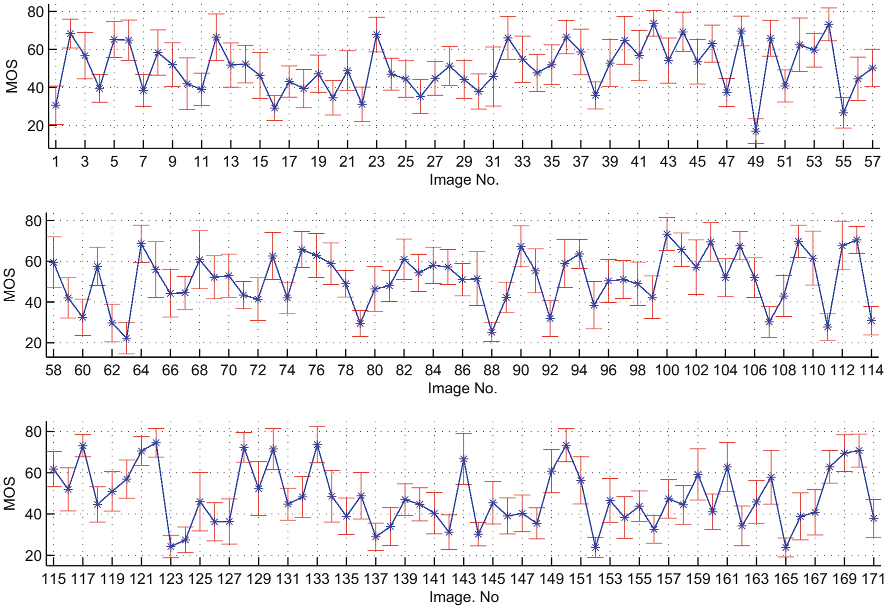


Fig. 8.6 The obtained MOS value of each retargeted image after processing (the horizontal axes correspond to the image number, and the vertical axes correspond to the MOS value. The *blue asterisk* indicates the obtained MOS value. And the *red error bar* indicates the standard deviation of the subjective scores)

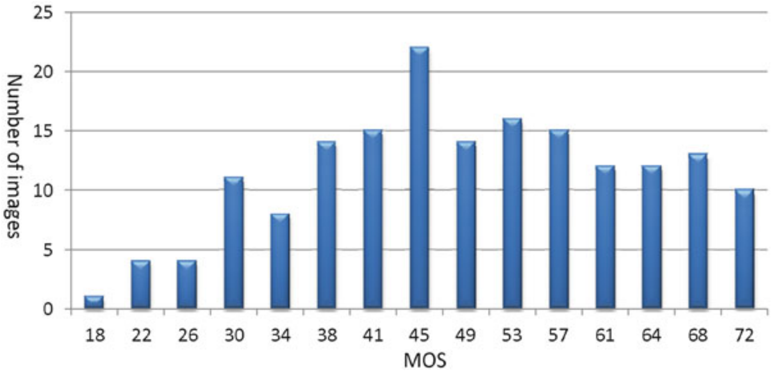


Fig. 8.7 Histogram of the MOS values in 15 equally spaced bins between the minimum and maximum MOS values of the image retargeting database

8.4 Objective Approaches for Retargeted Image Quality Assessment

8.4.1 Objective Quality Metrics

Image retargeting quality metric has been recently developed [33–39], in order to not only evaluate the retargeted image quality automatically and reliably in lieu of the subjective testing but also help to improve the performance of the retargeting methods. Nowadays, there are many quality metrics, which have been developed, such as the bidirectional warping (BDW) in [7], the quality metric scale space matching (SSM) in [38], the MPEG-7 descriptors in [39], the earth mover's distance (EMD) [33, 34], the bidirectional similarity (BDS) [35, 36], SIFTflow [37], and reduced-reference retarget metric [44]. The information about the metrics is detailed in the following.

8.4.1.1 Scale Space Matching

SSM is designed to facilitate extraction of global geometric structures from retargeted images. The proposed method is based on the scale invariant feature transform (SIFT) [45]. Given an original image I_{ori} and a retargeted image I_{ret}, two scale spaces $SP(I_{ori}) = \{I^0_{ori}, I^1_{ori}, \ldots, I^n_{ori}\}$, $SP(I_{ret}) = \{I^0_{ret}, I^1_{ret}, \ldots, I^n_{ret}\}$ of I_{ori} and I_{ret} are constructed, respectively, with the same Gaussian convolution kernel. Then distinctive invariant feature points (DIFPs) are detected in both $SP(I_{ori})$ and $SP(I_{ret})$ using the local extrema detection method in [45]. The attributes of each DIFP include location, scale, and orientation.

Assume that a correspondence from pixels of D^{i+1}_{ori} to pixels of D^{i+1}_{ret} has been established. The correspondence at scale i is established in both intra- and inter-scale manners. First, if DIFPs exist in both D^i_{ori} and D^i_{ret}, each DIFP pair (p^{DIFP}, q^{DIFP}), $p^{DIFP} \in D^i_{ori}$ and $q^{DIFP} \in D^i_{ret}$, is matched and evaluated using the local image descriptor (LID) [45]. This offers the intra-scale constraints. The inter-scale constraints are achieved by propagating pixel pair matching from coarse scale $(D^{i+1}_{ori}, D^{i+1}_{ret})$ to fine scale (D^i_{ori}, D^i_{ret}) in [38]. At the end of the hierarchical constraint-matching propagation process, a many-to-many mapping between pixels in I^0_{ori} and I^0_{ret} is established at the finest scale 0. This mapping can again be interpreted as a bipartite graph $G_{geostruct}$ that serves as the correspondence of two geometric structures in I^0_{ori} and I^0_{ret}. The similarity of two images I^0_{ori} and I^0_{ret} is defined as the similarity of two geometric structures measured as a weighted summation of edge-matching costs in $G_{geostruct}$. A simplified, non-weighted similarity metric is given by:

$$Sim(I^0_{ori}, I^0_{ret}) = \frac{\#_{ver}}{pn(I^0_{ori})+pn(I^0_{ret})} \cdot \frac{1}{\#_{edge}} \cdot \sum_{i=1}^{\#_{edge}} SSIM(v_0(e_i), v_1(e_i)) \qquad (8.7)$$

where $pn(I)$ is the number of pixels in image I, $e_i \in G_{geostruct}$, $\#_{ver}$ and $\#_{edge}$ is the number of vertices and edges in $G_{geostruct}$, respectively, $v_0(e_i)$, $v_1(e_i)$ are two vertices of e_i, and $SSIM(\cdot)$ is the structural similarity (SSIM) metric in [46] using a local 8×8 square window. The more similar I^0_{ori} and I^0_{ret} are, the more correspondences between pixels of I^0_{ori} and I^0_{ret} and the weight $\frac{\#_{ver}}{pn(I^0_{ori})+pn(I^0_{ret})}$ is closer to 1. The value of $Sim(\cdot)$ ranges between [0, 1]. Given two identical images, their similarity is maximized to be 1.

By considering that human visual system seems to selectively focus on salient regions, the distortions within salient regions should be more dominant. For each pixel in salient regions, if there is no corresponding pixels in the other image, it is linked to a dummy vertex dv of that image in $G_{geostruct}$ and set $SSIM(\cdot, dv) = 0$. This gives rise to a modified, saliency-based graph $SG_{geostruct}$. For each edge in $SG_{geostruct}$, if one of its vertices is in a salient map, its weight is set to be:

$$w_s = \frac{pn(I^0_{ori})+pn(I^0_{ret})+C}{pn(I^{salience}_{ori})+pn(I^{salience}_{ret})+C} \tag{8.8}$$

where $pn(I^{salience}_{ori})$ and $pn(I^{salience}_{ret})$ are the salient regions in I^0_{ori} and I^0_{ret}, respectively, and C is a small constant that prevents denominator very close to zero. The image size is in the magnitude of 10^5, and the scale 10^{-4} of the image size is used, *i.e.*, $C = 10$. If the area of salience regions is small, the weight w_s is large. If all pixels in images are salient, the weight is minimized to be one. For the remaining edges in $SG_{geostruct}$, the weight is set to be one. The saliency-based similarity metric is given by:

$$SalSim(I^0_{ori}, I^0_{ret}) = \frac{\#_{ver}}{pn(I^0_{ori})+pn(I^0_{ret})} \cdot \frac{1}{\sum_{i=1}^{\#_{edge}} w_i} \cdot \sum_{i=1}^{\#_{edge}} w_i \cdot SSIM(v_0(e_i), v_1(e_i)) \tag{8.9}$$

where w_i is the weight of edges is $SG_{geostruct}$.

8.4.1.2 MPEG-7 Descriptors

MPEG-7 [39] considered many descriptors from the color and texture perspectives, such as scalable color (SC) descriptor, color layout (CL) descriptor, color structure (CS) descriptor, homogeneous texture (HT) descriptor, and edge histogram (EH) descriptor. Detailed information is introduced in the following.

- CL [41] specifies the spatial distribution of colors. The extraction for the descriptor consists of four stages; image partitioning, dominant color selection, DCT transform, and non-linear quantization of the zigzag-scanned DCT coefficients. In the first stage, an input picture is partitioned into 64 blocks. The size of the each block is $W/8 \times H/8$, where W and H denote the width and height of an input picture, respectively. In the second stage, a single dominant color is selected in each block to build a tiny image whose size is 8 × 8. Any method for dominant color selection can be applied. Simple average colors is calculated as the dominant colors. In the third stage, each of the three components (Y, Cb, Cr) is transformed by 8×8 DCT, and we obtain three sets of DCT coefficients. A few low frequency coefficients are extracted using zigzag scanning and quantized to form the CL for a still picture. Image-to-image or sketch-to-image search can be implemented by calculating a distance of the descriptors. The distance between two CL descriptors is calculated as follows:

$$D = \sqrt{\sum_{i \in Y} w_i^1 (Y_i - \acute{Y}_i)^2} + \sqrt{\sum_{i \in Cb} w_i^2 (Cb_i - C\acute{b}_i)^2} + \sqrt{\sum_{i \in Cr} w_i^3 (Cr_i - C\acute{r}_i)^2} \tag{8.10}$$

Here, Y_i, Cb_i, and Cr_i denote the i-th coefficients of Y, Cb, Cr color component and w_i^l, w_i^2, and w_i^3 do the weighting values for the i-th coefficient, respectively. The weighting values should be decreased according to the zigzag-scan order.

- SC is defined in the hue-saturation-value (HSV) color space with fixed color space quantization, and uses a novel Haar transform encoding. The Haar transform based encoding facilitates a scalable representation of the description, as well as complexity scalability for feature extraction and matching procedures.
- CS expresses local color structure in an image using an 8×8-structuring element. It counts the number of times a particular color is contained within the structuring element as the structuring element scans the image. Suppose $c_0, c_1, c_2, \ldots, c_{M-1}$ denote the M quantized colors. A color structure histogram can then be denoted by $h(m), m = 0, 1, \ldots, M-1$, where the value in each bin represents the number of structuring elements in the image containing one or more pixels with color c_m. The hue-min-max-difference (HMMD) color is used for CS extraction.
- HT is computed by first filtering the image with a bank of orientation and scale sensitive filters, and computing the mean and standard deviation of the filtered outputs in the frequency domain. Specifically, the frequency space is partitioned into 30 channels with equal divisions in the angular direction and octave division in the radial direction. The individual feature channel is modeled using 2-D Gabor function. Then the image texture in each filtered channel is computed. The HT is extracted by concatenating mean intensity, the standard deviation of the image texture, the energy, and energy deviation.
- EH captures the edge distribution in spatial domain. For local edge distribution description, the image is divided into 4×4 sub-images, each of which is examined by five different orientations: vertical, horizontal, two diagonals, and isotropic (non-directional). For each sub-image, a 5-bin histogram is built by classifying edges to these five categories. Histogram concatenation generates the feature, which results in $4 \times 4 \times 5 = 80$ length description. Only the intensity component is employed for edge detection. And the L_1-norm distance is employed to measure the feature distance between two images, which is defined as $EH(S, T) = \parallel EHF(S) - EHF(T) \parallel_1$, where EHF is the edge histogram feature.

8.4.1.3 Earth Mover's Distance

EMD is based on the minimal cost that must be paid to transform one distribution into the other. The signature $\{S_j = (m_j, w_j)\}$, which represents a set of feature clusters, is viewed as the histogram distribution. The point m_j is the central value in bin j of the histogram, and w_j is to indicate the corresponding proportion. The definition of cluster is open. The color, position, and texture information can be employed to obtain the feature clusters. Only the size of the clusters in the feature space needs to be limited. Let $P = \{(p_1, w_{p_1}), \cdots, (p_m, w_{p_m})\}$ be the first signature with m clusters; $Q = \{(q_1, w_{q_1}), \cdots, (q_n, w_{q_n})\}$ is the second signature with n clusters. And $D = [d_{ij}]$ is the ground distance matrix, where d_{ij} is the ground distance between clusters p_i and q_j. d_{ij} can be any distance and will be chosen according to the problem at hand. The purpose is to find a flow $F = [f_{ij}]$, with f_{ij} as the flow between p_i and q_j, that minimizes the overall cost:

$$work(P, Q, F) = \textstyle\sum_i^m \sum_j^n d_{ij} f_{ij} \tag{8.11}$$

After obtaining the optimal flow F, EMD is defined as the work normalized by the total flow:

$$EMD(P, Q) = \frac{\sum_i^m \sum_j^n d_{ij} f_{ij}}{\sum_i^m \sum_j^n f_{ij}} \tag{8.12}$$

8.4.1.4 Bidirectional Similarity

Two signals S (original image) and T (retargeted image) are considered to be "visually similar" if as many as possible patches of S (at multiple scales) are contained in T, and vice versa. The dissimilarity can be formulated as:

$$(S, T) = \underbrace{\frac{1}{N_S} \sum_{P \subset S} min_{Q \subset T} D(P, Q)}_{d_{complete}(S,T)} + \underbrace{\frac{1}{N_T} \sum_{Q \subset T} min_{P \subset S} D(Q, P)}_{d_{cohere}(S,T)} \tag{8.13}$$

P and Q denote patches in S and T, respectively. And let N_S and N_T denote the number of patches in S and T. For each patch $Q \subset T$ we search for the most similar patch $P \subset S$, and measure their distance $D(P, Q)$, and vice-versa. The patches are taken around every pixel at multiple scales, resulting in significant patch overlap. $D(P, Q)$ can be any distance measurements between two patches, such as sum squared distances (SSD) or SSIM [46]. The two terms have important commentary roles. The first term, $d_{complete}(S, T)$ measures the deviation of the target T from "completeness" w.r.t. S. Namely, it measures if all patches of S have been preserved in T. The second term $d_{cohere}(S, T)$ measures if there are any "newborn" patches in T which have not originated from S. Therefore, the $d_{complete}(S, T)$ tries to represent the input image well (be complete), and the $d_{cohere}(S, T)$ makes sure the retargeted image is visually pleasing (coherent). The dissimilarity measurement is minimized in order to generate a retargeted image [35, 36].

8.4.1.5 SIFTflow

The SIFTflow descriptor characterizes view-invariant and brightness-independent image structures. Matching SIFT descriptors [42] allows establishing meaningful correspondences across image with significantly different image content. Furthermore, the pixel displacement (indicating by the SIFT correspondence matching) should be spatial coherent, which means that close-by pixels should have similar displacement. The cost function is defined as:

$$E(w) = \sum\nolimits_p \parallel s_1(p) - s_2(p+w) \parallel_1 + \tfrac{1}{\sigma^2} \sum\nolimits_p (\mu^2(p) + \nu^2(p)) + \\ \sum_{(p,q)\in\epsilon} \left(\min(\alpha|\mu(p) - \mu(q)|, d) + \min\left(\alpha|\nu(p) - \nu(q)|, d\right)\right) \tag{8.14}$$

where $w(p) = \left(\mu(p), \nu(p)\right)$ is the displacement vector at pixel location $p = (x, y)$, $s_i(p)$ is the SIFT descriptor extracted at location p in image i, and ϵ is the spatial neighborhood of a pixel. SIFTflow employs the SIFT for feature matching. And the local smoothness is preserved by the vector difference constraint.

8.4.1.6 Pyramid Histogram of Visual Words (PHOW)

First, SIFT [42] are computed at points on a regular grid with spacing M pixels. At each grid point the descriptors are computed over four circular support patches with different radii. Consequently, each point is represented by four SIFT descriptors. Multiple descriptors are computed to allow for scale variation between images. The dense features are vector quantized into N visual words using k-means clustering. Based on the SIFT descriptor and image spatial layout, we can obtain the pyramid histogram of visual words (PHOW) representation. In forming the pyramid the grid at level l has 2^l cells along each dimension. Consequently, level 0 is represented by N-vector corresponding to the N bins of the histogram, level 1 by a $4N$-vector, etc. PHOW is a vector with the dimensionality of $N \sum_{l=0}^{L} 4^l$.

8.4.1.7 GIST

GIST descriptor is extracted based on a very low dimensional representation of the scene, which is termed as the *Spatial Envelope* in [43]. A set of perceptual dimensions, such as naturalness, openness, roughness, expansion, ruggedness, is employed to represent the dominant spatial structure of a scene. For naturalness, the structure of a scene strongly differs between man-made and natural environments. Straight horizontal and vertical lines dominate man-made structures whereas most natural landscapes have textured zones and undulating contours. Therefore, scenes with edges biased toward vertical and horizontal orientation would have a low degree of naturalness. For openness, a scene can have a closed spatial envelop full of visual references or it can be vast and open to infinity. The existence of a horizon line and the lack of visual reference confer to the scene a high degree of openness. For roughness, it depends on the size of elements at each spatial scale. Roughness is correlated with the fractal dimension of the scene and thus, its complexity. For expansion, the convergence of parallel lines gives the perception of depth gradient of the space. A flat view of a building would have a low degree of expansion. On the contrary, a street with long vanishing lines would have a high degree of expansion. For ruggedness, it refers to the deviation of the ground with respect to the horizon. A rugged environment produces oblique contours in the picture and hides the horizon line. Therefore, rugged environments are mostly natural.

8.4.2 Performances on RetargetMe Database

For RetargeMe database, the authors tried to estimate how well the objective metrics agree with the users' subjective preferences. The agreement can be evaluated by the correlation between the rankings induced by the subjective and objective measures. For every image I, the subjective similarity vector $s =< s_i, \cdots, s_n >$ for $n = 8$ objective quality methods can be obtained, where s_i is the number of times the retargeting result T_i using method i was favored over another result. We also define $o =< o_1, \cdots, o_n >$ as the respective objective distance vector for the same image I calculated by one of the objective measures. For a given objective measure D, the entry $o_i = D(I; T_i)$ is the distance between I and T_i with respect to measure D (in this case, the lower o_i the better the method i is). The analysis results for all images and measures can be found in [17].

8.4.2.1 Correlation

To compare between s and o, the authors first sort them and then rank the retargeting measures according to the sorted order. The subjective vector s is sorted in descending order, while the objective vector o is sorted in ascending order. We only need to compare s and o to statistically determine the correlation between two rankings, $rank_{desc}(s)$ and $rank_{asc}(o)$ induced by these vectors. Kendall distance [19] is employed to measure the degree of correlation between the two rankings:

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)} \tag{8.15}$$

where n is the length of the rankings, n_c is the number of concordant pairs, and n_d is the number of discordant pairs over all pairs of entries in the ranking. It is easy to see that $-1 \leq \tau \leq 1$ with increasing value indicates increasing rate of agreement. Note that $\tau = 1$ in case of perfect agreement (equal rankings), and $\tau = -1$ is case of perfect disagreement. In case $\tau = 0$, the rankings are considered independent.

8.4.2.2 Significance Test

To measure the significance of a correlation estimate, we need to consider the distribution of the τ coefficient. It turns out that the distribution of τ tends to normality for large n [19]. For RetargetMe database, we can easily estimate the distribution of τ for $n = 8$ by considering the rank correlation of all possible permutations of 8 elements with regard to an objective order $1, 2, \cdots, 8$. The distribution has normal characteristics, with zero-mean and $\sigma = 0.2887$. For a given set of observed τ coefficients, χ^2 test against the null hypothesis that the observed τ coefficients are randomly sampled from the τ distribution is employed.

8.4.2.3 Analysis

The distribution of τ scores over all images is gathered for each measure. The mean and variance of this distribution is calculated to represent the score of the metric in this experiment. Table 8.1 presents these scores, with breakdown according to image attribute, and the total score over the entire database. The results are shown for the full rank-vectors, and also with respect to the $k = 3$ results ranked highest by each measure. For the latter, Eq. (8.15) is modified such that only pairs (i, j) for which $(rank_1(i) \leq k \vee rank_1(j) \leq k) \wedge (rank_2(i) \leq k \vee rank_2(j) \leq k)$ are considered, and the denominator is modified to be the total number of such pairs. For reference, the results for a random metric is added in Table 8.1. For a given pair of images, the measurement simply returns a uniformly random number in (0, 1).

It can be observed that, low-level metrics show smaller agreements with the users, although EH achieves higher scores for images containing apparent geometric structures or symmetries. However, both BDS and BDW show low agreements with the user data as well. The near-zero correlation for nearly all image classes suggests they cannot well match the viewers' preferences for retargeted images. Their unsatisfying performance is attribute to both the way they construct correspondence between the images, and with the image features they use for measuring the distance.

The measurements employed the patch differences to indicate the image dissimilarity. Those are strict measures which assign high penalties to patch variations caused by small local deformations (e.g., small scale or rotation). However, such deformations might be acceptable by human viewers, and so may be reasonable to use for retargeting purposes. As for the correspondence, since BDS uses global patch comparison, a deformed region in the result might be matched to a different part of the original image that has similar appearance. Thus, record of specific

Table 8.1 Correlation of objective and subjective measures for the complete rank ($k = \infty$)

	Metric						Total		
Metric	Lines edges	Faces people	Texture Texture	Foreground objects	Geometric structures	Symmetry	Mean	std	ρ-value
BDS	0.040	0.190	0.060	0.167	−0.004	−0.012	0.083	0.268	0.017
BDW	0.031	0.048	−0.048	0.060	0.004	0.119	0.046	0.181	0.869
EH	0.043	−0.076	−0.060	−0.079	0.103	0.298	0.004	0.334	0.641
CL	−0.023	−0.181	−0.071	−0.183	−0.009	0.214	−0.068	0.301	0.384
SIFTflow	0.097	0.252	0.119	0.218	0.085	0.071	0.145	0.262	0.031
EMD	0.220	0.262	0.107	0.226	0.237	0.500	0.251	0.272	10^{-5}

In each column the mean τ correlation coefficient is shown ($-1 \leq \tau \leq 1$), calculated over all images in the database with the corresponding attribute. The last three columns show the mean score, standard deviation, and respective ρ-value over all image types changes in content might not be reflected in the distance. BDW does constrain the correspondence such that regions in the result will be matched to approximately the same regions in the original image. However, due to information insufficiency (it is of one-dimensional), it has difficulty in dealing with the results produced by some operations. SIFTflow and EMD on the other hand, use a dense SIFT descriptor which can robustly capture structural properties of an image. EMD even uses a state-of-art color descriptor.

The results using these measures show good evidence in Tables 8.1 and 8.2. It is evident that EMD and SIFTflow produce rankings which better agree with the user labels compared with the other objective measures. EMD shows somewhat better results for the full ranking, while the two are on par with respect to the top-ranked results. In general, the measures have stronger correlation with the subjective results on images with faces or people, and evident foreground objects. Tables 8.1 and 8.2 also show the calculated ρ-values for this analysis. BDS, SIFTflow, and EMD show significant results for $\rho < 0.01$, and so we can claim the fact that EMD and SIFTflow have better correlation with the users with high statistics confidence. For $k = 3$, the calculated correlations for all metrics are significant at $\alpha = 0.01$ confidence level.

Image descriptors such as SIFT are demonstrated to be more suitable than patch-based distances to describe local permissible content changes. Moreover, the constrained alignment produced by these methods also appears to better model the deformations introduced by retargeting operators, and thereby provides more reliable content matching for retargeting measures.

8.4.3 Performances on CUHK Retargeting Database

For the MPEG-7 descriptors, the public MPEG-7 low level feature extraction tools[2] are employed to extract the corresponding descriptors, such as SC, CS, CL, HT, and EH. According to the default settings, the lengths of the MPEG-7 descriptors are different. The lengths of SC, CS, CL, HT, and EH are 128, 64, 120, 62, and 80, respectively. For the EMD, the code[3] is employed to depict the perceptual quality of the retargeted image. For PHOW, SIFT descriptors are first computed at points on a regular grid with spacing M pixels, here $M = 10$. At each grid point the descriptors are computed over circular support patches with radii $r = 4, 8, 12$, and 16 pixels. The patches with radii 4 do not overlap and the other radii do. As all the images are color, this process will provide a 128×3 D-SIFT descriptor for each point. These descriptors are rotation invariant. The k-means clustering is performed over 2,000 training images. Finally the vocabulary consisting of 2,000 visual words is used here. Then the vocabulary is used to extract PHOW features,

[2] http://www.cs.bilkent.edu.tr/~bilmdg/bilvideo-7/Software.html.

[3] http://www.seas.upenn.edu/~ofirpele/FastEMD/.

Table 8.2 Correlation of objective and subjective measures for the three highest ranked results ($k = 3$)

	Metric						Total		
Metric	Lines edges	Faces people	Texture Texture	Foreground objects	Geometric structures	Symmetry	Mean	std	ρ-value
BDS	0.062	0.280	0.134	0.249	−0.025	−0.247	0.108	0.532	0.005
BDW	0.213	0.141	0.123	0.115	0.212	0.439	0.200	0.395	0.002
EH	−0.036	−0.207	−0.331	−0.177	0.111	0.294	−0.071	0.593	0.013
CL	−0.307	−0.336	−0.433	−0.519	−0.366	0.088	−0.320	0.543	10^{-6}
SIFTflow	0.241	0.428	0.312	0.442	0.303	0.002	0.298	0.483	10^{-6}
EMD	0.301	0.416	0.216	0.295	0.226	0.534	0.326	0.496	10^{-6}

In each column the mean τ correlation coefficient is shown ($-1 \leq \tau \leq 1$), calculated over all images in the database with the corresponding attribute. The last three columns show the mean score, standard deviation, and respective ρ-value over all image types

Table 8.3 Performances of different shape descriptors on CUHK retargeting database

	LCC	SROCC	RMSE	OR
SC	0.1508	0.1792	13.347	0.2164
CS	0.1520	0.1688	32.731	0.5322
CL	0.1033	0.0850	13.429	0.2398
HT	0.0829	0.0890	35.151	0.5673
EH	0.3031	0.2729	12.866	0.2047
EMD	0.2760	0.2904	12.977	0.1696
PHOW	0.3706	0.2308	12.540	0.2222
BDS	0.2896	0.2887	12.922	0.2161
SIFTflow	0.3141	0.2899	12.817	0.1462
GIST	**0.5443**	**0.5114**	**11.326**	**0.1579**
PHOW+GIST	0.5440	0.5090	11.329	0.1579
MPEG-7	0.1164	0.1502	24.357	0.4094
MPEG-7+PHOW+GIST	0.1168	0.1504	24.257	0.4094
Combination	**0.5999**	**0.5609**	**10.801**	**0.1228**

which generates a vector with the dimensionality as 2,000. The authors employed VLFeat [40] to extract the PHOW descriptors. For GIST, the code provided by the authors[4] is employed. As a result, the dimension of the GIST feature is 960.

The performances of different metrics are illustrated in Table 8.3. The linear correlation coefficient (LCC) measures the prediction accuracy. The Spearman rank-order correlation coefficient (SROCC) provides an evaluation of the prediction monotonicity. The root mean square error (RMSE) is introduced for evaluating the error during the fitting process. The outlier ratio (OR) evaluates the consistency attributes of the objective metric, which represents the ratio of "outlier-points" to the total points. Firstly, we compared the performances of each shape descriptors. It can be observed the GIST can achieve the best performances, which significantly outperforms the other shape descriptors. The reason is that GIST considers the image shape from several perspectives, such as naturalness, openness, roughness, expansion, and ruggedness. By considering the shape information from these perceptual dimensions, the object shape can be accurately depicted. Therefore, as some retargeting methods significantly degrade the shape information, such as seam-carving [4], the distortions introduced can be more precisely captured. Moreover, GIST is regarded as a global descriptor, which is believed to be able to capture more shape information from the global viewpoint compared with other shape descriptors. For EMD, the composed histogram only represents the edge distribution of the image, which cannot accurately represent the object shape and the content information of the image. BDS tries to capture how much information one image conveys of the other image in a bidirectional way. However, although it

[4] http://people.csail.mit.edu/torralba/code/spatialenvelope/.

is claimed that the spatial geometric relationship is considered by a multiple scale approach, the order-relationship can still not be preserved, such as the local-order of each pixel or patch. Therefore, the dissimilarity metric of BDS does not accurately depict the object shape distortion either. SIFTflow employs the SIFT descriptor to detect the correspondence between two images. It is claimed that the order-relationship of the pixels or patches is captured. However, the content information loss during the retargeting process is not considered. PHOW can somewhat extract some shape information. However, a visual vocabulary is introduced to compose the corresponding histogram at each pyramid scale. Therefore, the shape information is mostly extracted from the local perspective, although a pyramid structure is employed for PHOW. The global shape information cannot be accurately described. For the descriptors of MPEG-7, EH performs the best. The reason is that the local shape information is depicted by the edge histogram in local region. The global shape information is somewhat captured by concatenating the local edge histogram. For the other shape descriptors, CS, SC, and CL mostly focus color part. Although color can somewhat represent the shape information, the accuracy cannot be ensured by the color features. HT concatenates the energy of each frequency channel, which does not pay much attention on the shape description. These are the main reasons why the other shape descriptors cannot depict the perceptual quality of the retargeted image, compared with GIST.

Secondly, we test the performance by combining these shape descriptors together. MPEG-7 combines CS, SC, CL, EH, and HT together. PHOW+GIST concatenates PHOW and GIST together to generate a vector feature with the dimensionality as 2,960. MPEG-7+GIST+PHOW combines the shape features (MPEG-7 features, GIST, PHOW) together. From the experimental results in Table 8.3, the performances of these combinations cannot outperform its best component. Therefore, we cannot expect better performance by simply combining as many as shape descriptors. The shape descriptors may conflict with each other for evaluating the retargeted image perceptual quality.

Finally, as discussed in [31], the distortion introduced in retargeting process can be categorized into shape distortion and content information loss. And the measurements for these two distortions are complementary for each other. By combining these measurements together, a better performance is expected. We combine shape descriptors (GIST, and EMD) together with content information loss measurements (BDS and SIFTflow). The combination process is a simple summation process. The quality score is obtained by:

$$Q = \alpha \times log_2(GIST) + log_2(EM) + log_2(BDS) + log_2(SIFTflow) \quad (8.16)$$

where α is simply set as 10. With the evaluation on the database, the best performance is observed, which greatly outperforms GIST.

8.5 Future Trends

As demonstrated in previous subsections, the performances of the objective quality metrics for retargeted images are still not good enough. The statistical correlations between the subjective MOS values and the metric outputs are not close. Even fused together, the LCC and SROCC values are smaller than 0.6, indicating that performance of objective metrics are bad. Here we discuss how to design an effective objective quality metric for evaluating the perceptual quality of the retargeted image for which, we consider the source image content, retargeting scale, the shape distortion and content information loss measurement, the HVS properties, and so on.

- Shape distortion description. Shape distortion is a closely related factor. Therefore, the recently developed metrics, such as EH, EMD, SIFTflow, GIST, and PHOW are designed to capture the object shape of the image and measure the corresponding differences between the source and retargeted image. However, its performance is not good enough, and we can find in Table 8.3 that the LCC and SROCC values are around 0.35. Therefore, shape distortions introduced during retargeting need to be more accurately captured. Recently, D'Angelo A. [47, 48] proposed a full-reference quality metric to evaluate the geometrical distortions of the images. Their approach is based on the assumption that the HVS is sensitive to the image structures, such as edges and bars, calculated from the results of Gabor filter. By considering this descriptor for evaluating the geometrical distortion, the shape distortion introduced during the retargeting process is believed to be more accurately described. Therefore, it can help to improve the performance of the objective quality metric.
- Fusion of the shape distortion and content information loss. As illustrated in Table 8.3, the content information loss alone does not much influence the final perceptual quality of the retargeted image. But combined with the shape distortion and content information loss we can improve the performance, as illustrated in Table 8.3. The combinations of the four objective quality metrics can beat the other metrics. Therefore, if we develop accurate metrics to capture the shape distortion and content information loss, how to fuse them together needs to be further considered. The fusion strategy should consider the contribution of the two factors to the final retargeted image quality.
- Source image quality and retargeting scale. The source images that we employed to build our database are of different resolutions and different qualities. This may affect the subjective viewers' judgment of the retargeted image perceptual quality. Moreover, the retargeting scale will also affect the retargeted image quality. Given one source image, the larger the retargeting ratio, the better the perceptual quality of the retargeted image is. Therefore, the final perceptual quality index of the retargeted image needs to account for the quality of the source image as well as the retargeting scale.
- Image content. The image content relates closely to the crop margin of the source image (the maximum region that can be cropped without losing an object of

interest). If the source image contains the “clear foreground object” or “natural scenery” attribute, the crop margin will be very large. Therefore, retargeting the source image into 75 % and 50 % ratios will not significantly affect the perceptual quality. Otherwise, if the source image contains the “face and people” or “geometric structure” attribute, the crop margin will be very small. In this case, no existing retargeting methods can preserve good perceptual quality during retargeting. In this respect, the image content and the crop margin of each source image need to be included to depict the perceptual quality of the retargeted image.

- HVS saliency. Additionally, the HVS demonstrates different conspicuities over different regions of image, indicating that the shape distortions and content information loss in the salient regions are more sensitively perceived by the viewers than those in the non-salient regions. That is also the reason why several retargeting methods consider the saliency or visual attention map during the retargeting process [2, 3, 8]. The viewers’ assessment on the quality of the retargeted image is prejudiced during the subjective testing process. Therefore, the effect of the HVS saliency needs to be considered to model the subjective viewer’s behavior, which will lead to a more effective quality metric for retargeted images. The simplest way of incorporating the HVS saliency is to weight the corresponding shape distortion and content information loss by the saliency map detected from the source image, which has been demonstrated to be effective in evaluating the perceptual quality of the traditional distorted image.

Moreover it should be noted that more shape descriptors are not able to ensure a better performance. Some shape descriptors, such as HT and CL, cannot effectively evaluate the retargeted image quality. Therefore, shape descriptors should be carefully selected for metric design which needs to represent the shape distortions introduced by retargeting. Secondly, some shape descriptors, such CL, HT, CS, and SC, focus on color/energy distribution or layout. Although they can somewhat capture the shape information, the structure of the retargeted image is preferred to be beneficial for retargeted image quality measurement. GIST is a global shape descriptor, which significantly outperforms the other descriptors. The other descriptors, such as PHOW and EH, tried to depict the global information in a bottom-up manner, where the local information are grouped or concatenated together to represent global information. The experimental results demonstrate that the quality evaluation of retargeted image should concentrate on the global information. It truly matches the HVS property. During the subjective test, the viewer’s preference is highly affected by the global shape information. If the shape information appears to be very annoying globally, the subjective score will be absolutely very low, no matter how well the local shape information is preserved. While the global shape information is well preserved, the viewer will then clearly check the local shape information. Therefore, during the quality metric design, global shape information should be of the highest priority. In the following, the local shape information should be considered to be complementary.

Conclusion
This chapter reviews current progresses of the retargeted image quality assessment from both subjective and objective assessment perspective. Nowadays, there are two retargeted quality image databases for public use, which are built through the subjective testing process. Based on these two databases, many quality metrics are compared, such as GIST, PHOW, SIFTflow, EH, and so on. However, new challenges have been issued for better quality metrics to evaluate retargeted images, which need to consider the shape distortion descriptor, the loss information descriptor, the HVS property, and so on.

Acknowledgements The work described in this chapter was partially supported by a grant from the Research Grants Council of the Hong Kong SAR, China (Project CUHK 415913); the National Nature Science Foundation of China under Grant No. 61301090; the ROSE Lab grant from the Singapore National Research Foundation; the Supporting Program for Beijing Excellent Talents under Grant No. 2013D009011000001, and the National Natural Science Foundation of China under Grant No. 61202242.

References

1. Shamir A., and Sorkine O.: Visual media retargeting. ACM SIGGRAPH Asia Courses, (2009).
2. Wolf L., Guttmann M., and Cohen-Or D.: Non-homogeneous content-driven video-retargeting. Proceedings of International Conference on Computer Vision, (2007).
3. Krahenbuhl P., Lang M., Hornung A., and Gross M.: A system for retargeting of streaming video. Proceedings of SIGGRAPH Asia, (2009).
4. Avidan S., and Shamir A.: Seam carving for content-aware image resizing. Proceedings of SIGGRAPH, (2007).
5. Rubinstein M., Shamir A., and Avidan A.: Improved seam carving for video retargeting. Proceedings of SIGGRAPH, (2008).
6. Shamir A., and Avidan S.: Seam-carving for media retargeting. Communications of the ACM, 52(1), 77–85, Jan. (2009).
7. Rubinstein M., Shamir A., and Avidan S.: Multi-operator media retargeting. Proceedings of SIGGRAPH, (2009).
8. Wang Y., Tai C., Sorkine O., and Lee T.: Optimized scale-and-stretch for image resizing. Proceedings of SIGGRAPH Asia, (2008).
9. Pritch Y., Kav-Venaki E., and Peleg S.: Shift-map image editing. Proceedings of International Conference on Computer Vision, (2009).
10. Qi A., and Ho J.: Shift-map based stereo image retargeting with disparity adjustment. Proceedings of Asian Conference on Computer Vision, (2013).
11. Dekel T., Moses Y., and Avidan S.: Stereo seam carving: a geometrically consistent approach. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(10), 2513–2525, Oct. (2013).
12. Dekel T., Moses Y., and Avidan S.: Geometrically consistent stereo seam carving. Proceedings of International Conference on Computer Vision, (2011).
13. Itti L., Koch C., and Niebur E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(11), 1254–1259, Nov. (1998).

14. Karni Z., Freedman D., and Gotsman C.: Energy-based image deformation. Proceedings of Symposium on Geometry Processing, (2009).
15. Dong W., Zhou N., Paul J. C., and Zhang X.: Optimized image resizing using seam carving and scaling. Proceedings of SIGGRAPH, (2009).
16. Ma L., Deng C., Lin W., and Ngan K. N.: Image retargeting subjective quality database. Available. http://ivp.ee.cuhk.edu.hk/projects/demo/retargeting/index.html.
17. Rubinstein M., Gutierrez D., Sorkine O., and Shamir A.: A comparative study of image retargeting. Proceedings of SIGGRAPH Asia (2010). Available. http://people.csail.mit.edu/mrub/retargetme/.
18. Kendall M. G., and Babington Smith B.: On the method of paired comparisons. Biometrika, 31, 324–345, (1940).
19. Kendall M. G.: A new measure of rank correlation. Biometrika, 30, 81–93, (1938).
20. VQEG.: Final report from the video quality experts group on the validation of objective models of video quality assessment II. (2009). Available. http://www.its.bldrdoc.gov/vqeg/projects/frtv_phaseII/downloads/VQEGII_Final_Report.pdf.
21. VQEG. Final report from the video quality experts group on the validation of objective models of video quality assessment. (2000). Available. http://www.its.bldrdoc.gov/vqeg/projects/frtv_phaseI/.
22. VQEG. Final report from the video quality experts from group on the validation of objective models of multimedia quality assessment Phase 1. Available. ftp://vqeg.its.bldrdoc.gov/Documents/Projects/multimedia/MM_Final_Report/.
23. Sheikh H. R., Sabir M. F., and Bovik A. C.: A statistical evaluation of recent full reference image quality assessment algorithms. IEEE Transactions on Image Processing, 15(11), 3440–3451, Nov. (2006).
24. Soundararajan K., Soundararajan R., Bovik A. C., and Cormack L. K.: Study of subjective and objective quality assessment of video. IEEE Transaction on Image Processing, 19(6), 1427–1441, Jun. (2010). Available. http://live.ece.utexas.edu/research/quality/live_video.html.
25. Soundararajan K., Soundararajan R., Bovik A. C., and Cormack L. K.: A subjective study to evaluate video quality assessment algorithms. Proceedings of SPIE, Human Vision and Electronic Imaging, Jan. (2010).
26. Pinson M. H., and Wolf S.: Comparing subjective video quality testing methodologies. Proceedings of SPIE, 5150(3), 573–582, (2003).
27. Van Dijk A. M., Martens J. B., and Watson A. B.: Quality assessment of coded images using numerical category scaling. Proceedings of SPIE Advanced Image and Video Communications and Storage Technologies, (1995).
28. VQEG: Final report from the video quality experts from group on the validation of objective models of multimedia quality assessment Phase 1. Available. ftp://vqeg.its.bldrdoc.gov/Documents/Projects/multimedia/MM_Final_Report/
29. ITU-R Recommendation BT.500–11.: Methodology for the subjective assessment of the quality of television pictures", ITU, Geneva, Switzerland, (2002).
30. ITU-T Recommendation P.910.: Subjective video quality assessment methods for multimedia applications. ITU, Geneva, Switzerland, (2008).
31. Ma L., Lin W., Deng C., and Ngan K. N.: Image retargeting quality assessment: a study of subjective scores and objective metrics. IEEE Journal of Selected Topics in Signal Processing, 6(6), 626–639, Oct. (2012).
32. Ma L., Lin W., Deng C., and Ngan K. N.: Study of subjective and objective quality assessment of retargeted images. Proceedings of International Symposium on Circuits and Systems, (2012).
33. Pele O., and Werman M.: Fast and robust earth mover's distances. Proceedings of International Conference on Computer Vision, (2009).
34. Rubner Y., Tomasi C., and Guibas l. J.: The earth mover's distance as a metric for image retrieval. International Journal of Computer Vision, 40(2), 99–121, Nov. (2000).

35. Simakov Yaron Caspi D., Shechtman E., and Irani M.: Summarizing visual data using bidirectional similarity. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, (2008).
36. Barnes C., Shechtman E., Finkelstein A., and Goldman D. B.: Patchmatch: a randomized correspondence algorithm for structural image editing. Proceedings of SIGGRAPH, (2009).
37. Liu C., Yuen J., Torralba A., Sivic J., and Freeman W. T.: SIFT flow: dense correspondence across different scenes. Proceedings of European Conference on Computer Vision, (2008).
38. Liu Y., Luo X., Xuan Y., Chen W., and Fu X.: Image retargeting quality assessment. Proceedings of EUROGRAPHICS, (2011).
39. Manjunath B. S., Ohm J. R., Vasudevan V. V., and Yamada A.: Color and texture descriptors. IEEE Transaction on Circuits and System for Video Technology, 11(6), 703–715, Jun. (2001).
40. Vedaldi A., and Fulkerson B.: VLFeat: An open and portable library of computer vision algorithms. Available. http://www.vlfeat.org/, (2008).
41. Kasutani E., and Yamada A.: The MPEG-7 color layout descriptor: a compact image feature description for high-speed image/video segement retrieval. Proceedings of International Conference on Image Processing, 674–677, (2001).
42. Lowe D.:Object recognition from local scale-invariant features. Proceedings of International Conferene on Conmputer Vision, (1999).
43. Oliva A., and Torralba A.: Modeling the shape of the scene: a holistic representation of the spatial envelope. Internaltional Journal of Computer Vision, 42(3), 145–175, (2001).
44. Lu W., and Wu M.: Reduced-reference quality assessment for retargeted images. Proceedings of International Conference on Image Processing, 1497–1500, (2012).
45. Lowe D.: Dictinctive image features from scale invariant keypoints. International Journal of Conputer Vision, 60(2), 91–110,(2004).
46. Wang Z., Bovik A., Sheikh H., Simoncelli E.: Image quality assessment: from error visibility to structureal similarity. IEEE Transactions on Image Processing, 13(4), 600–612, Apr. (2004).
47. D'Angelo A., Menegaz G., and Barni M.: Perceptual quality evaluation of geometric distortions in images. Proceedings of SPIE Human Vision and Electronic Imaging, 6492, (2007).
48. D'Angelo A., Zhao Z., and Barni M.: A full-reference quality metric for geometrically distorted images. IEEE Transactions on Image Processing, 19(4), 867–881, Apr. (2010).

Chapter 9
Quality Assessment in Computer Graphics

Guillaume Lavoué and Rafał Mantiuk

9.1 Introduction

The realm of computer graphics is an intensive producer of visual content. Depending on the concerned sub-areas (e.g., geometric modeling, animation, rendering, simulation, high dynamic range (HDR) imaging, and so on) it generates and manipulates images, videos, or 3D data. There is an obvious need to control and evaluate the quality of these graphical data regardless of the application. The term *quality* means here the *visual impact of the artifacts* introduced by the computer graphics techniques. For instance, in the context of rendering, one needs to evaluate the level of annoyance due to the noise introduced by an approximate illumination algorithm. As another example, for level of details creation, one needs to measure the visual impact of the simplification on the appearance of a 3D shape. Figure 9.1 illustrates these two examples of artifacts encountered in computer graphics. The paragraphs below introduce several useful terms that also point out the main differences between existing approaches for quality assessment in graphics.

Artifact Visibility vs. Global Quality For a given signal to evaluate (e.g., an image), the term *quality* often refers to a single score (mean-opinion-score, MOS) that aims at reflecting a kind of global level of annoyance caused by all artifacts and distortions in an image. Such global quality index is relevant for many computer graphics applications, e.g. to reduce/augment the sampling density in ray-tracing rendering. However, beside this global information, it is also important in many

G. Lavoué (✉)
CNRS, University of Lyon, Insa-Lyon, LIRIS UMR 5205, Lyon, France
e-mail: glavoue@liris.cnrs.fr

R. Mantiuk
School of Computer Science, Bangor University, Bangor, Gwynedd LL57 2DG, UK
e-mail: mantiuk@gmail.com

C. Deng et al. (eds.), *Visual Signal Quality Assessment*,
DOI 10.1007/978-3-319-10368-6_9

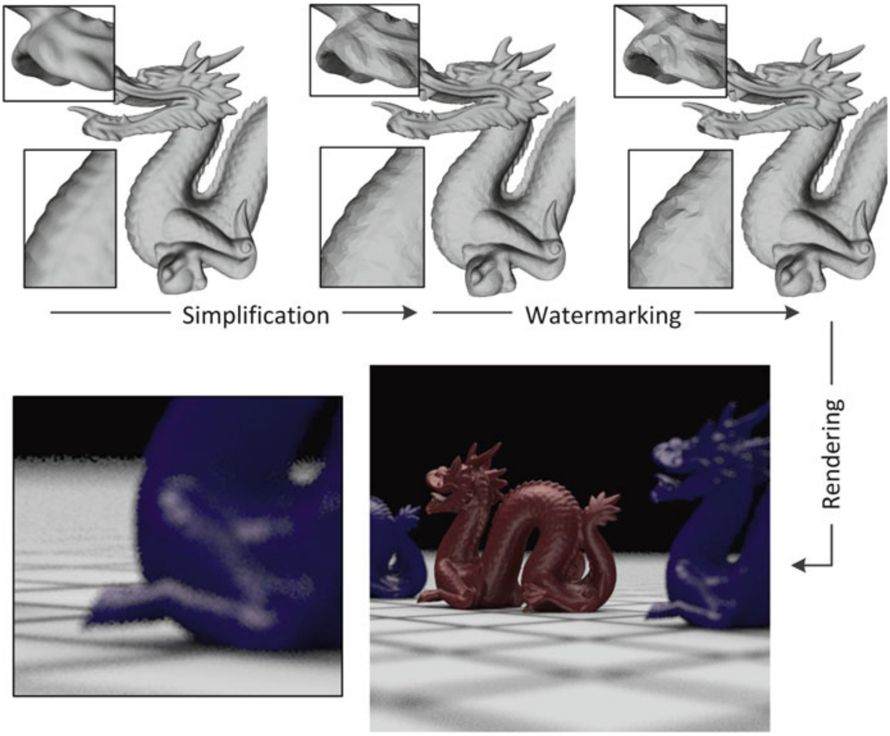


Fig. 9.1 Illustration of a typical computer graphics work-flow and its different sources of artifacts. *Top row, from left to right:* An original scanned 3D model (338K vertices); result after simplification (50K vertices) which introduces a rather uniform high frequency noise; result after watermarking [95] which creates some local bumps on the surface. *Bottom row:* Result after rendering (radiance caching) which introduces a nonuniform structured noise

cases to obtain an information about the local *visibility* of the artifacts (i.e., predicting their spatial localization in the image). Such local information may allow, for instance, an automatic local corrections of the detected artifacts, like in [30].

Objective vs. Subjective Quality Assessment The quality evaluation of a given stimulus can be done directly by gathering the opinion of some observers by means of a *subjective* experiment. However, this kind of study is obviously time-consuming, expensive and cannot be integrated into automatic processes. Hence researchers have focused on *objective* and automatic metrics that aim to predict this subjective visibility and/or quality. Both approaches are presented in this chapter.

Reference vs. No Reference Objective quality metrics can be classified according to the availability of the reference image (resp. video or 3D models): full-reference (FR), reduced reference (RR), and no-reference (NR). FR and RR metrics require at the quality evaluation stage that full or partial information on both images is present, the reference and the distorted one. NR metrics are much more challenging because they only have access to the distorted data; however, they are particularly relevant in computer graphics of which many techniques do not only *modify* but also *create* visual content from abstract data. For instance, a rendering process generates a synthetic image from a 3D scene, hence to evaluate the rendering artifacts the metric will have access only to the test image since a perfect reference image without artifact is often unavailable.

Image Artifacts vs. Model Artifacts Computer graphics involves coarsely two main types of data: 3D data, i.e. surface and volume meshes issued from geometric modeling or scanning processes and 2D images and videos created/modified by graphical processes like rendering, tone-mapping, and so on. Usually, in a computer graphics work-flow (e.g., see Fig. 9.1), 3D data are first created (geometric modelling), processed (e.g., filtering, simplification), and then images/videos are generated from this 3D content (by rendering) and finally they can be post-processed (tone-mapped, for instance). In such scenario, the visual defects at the very end of the processing chain may be due to artifacts introduced both on the 3D geometry (what we call model artifacts) and on the 2D image/video (what we called image artifacts). Since these two types of artifacts are introduced in very distinct processes and evaluated using very distinct metrics, each part of this chapter is divided according to this classification (except Sects. 9.2 and 9.3, respectively, dedicated to each of them).

Black-Box Metrics vs. White-Box Metrics There are two main approaches to modeling quality and fidelity: a black-box approach, which usually involves machine learning techniques; and a white-box approach, which attempts to model processes that are believed to exist in the human visual system. The visual difference predictors (VDPs), such as VDP [20], are an example of a white-box approach, while the data-driven metrics for non-reference quality prediction [30] or color palette selection [65] are the examples of the black-box approach. Both approaches have their shortcomings. The black-box methods are good at fitting complex functions, but are prone to over-fitting. It is difficult to determine the right size of the training and testing data sets. Unless very large data sets are used, nonparametric models used in machine learning techniques cannot distinguish between major effects, which govern our perception of quality, and minor effects, which are unimportant. They are not suitable for finding a general patterns in the data and extracting a higher level understanding of the processes. Finally, the success of the machine learning methods depends on the choice of feature vectors, which need to be selected manually, relying in equal amounts on the expertise and a lucky guess.

White-box methods rely on the vast body of research devoted to modeling visual perception. They are less prone to over-fitting as they model only the effects that they are meant to predict. However, the choice of the right models is difficult. But even if the right set of models and right complexity is selected, combining and then calibrating them all together is a major challenge. Moreover, such white-box approaches are not very effective at accounting for higher level effects, such as aesthetics and naturalness, for which no models exist.

It is yet to be seen which approach will dominate and lead to the most successful quality metrics. It is also foreseeable that the metrics that combine both approaches will be able to benefit from their individual strengths and mitigate their weaknesses.

This chapter is organized as follows: Sects. 9.2 and 9.3, respectively, present objective quality assessment regarding image artifacts and model artifacts. Then Sect. 9.4 details the subjective quality experiments that have been conducted by the computer graphics community as well as quantitative evaluations of the objective metrics presented in Sects. 9.2 and 9.3. Finally Sect. 9.5 is dedicated to the emerging trends and future research directions on the subject of quality assessment in graphics.

9.2 Image Quality Metrics in Graphics

9.2.1 Metrics for Rendering Based on Visual Models

Computer graphics rendering methods often rely on physical simulation of light propagation in a scene. Due to complex interaction of light with the environment and massive amount of light particles in a scene, these simulations require huge amount of computation. However, it has been long recognized that most applications of computer rendering methods require perceptually plausible solution rather than physically accurate results [71]. Knowing the limitations of the visual system, it should be possible to simplify the simulation and reduce the computational burden [66].

When rendering a scene, two important problems need to be addressed: (a) how to allocate samples (computation) over the image to improve perceptual quality; and (b) when to stop collecting samples as further computation does not result in perceivable improvement. Both problems were considered in a series of papers on perceptually based rendering, intended for both an accurate off-line techniques [11, 12, 26, 30, 62–64, 72, 104] and interactive rendering [23, 53]. Although the perceptual metrics used in these techniques operate in the image space, they are different from the typical fidelity metrics, which compute the difference between reference and test images. Since the reference image is usually not available when rendering, these metrics aim at estimating error bounds based on approximation of the final image. This approximation can be computed using fast GPU methods [104], by simulating only direct light (ray-casting) [72], approximating an image in the frequency domain [11, 12], using textures [94], intermediate rendering results [62], or consecutive animation frames [63]. Such approximated images may not contain all the illumination and shading details, especially those that are influenced by indirect lighting. However, the approximation is good enough to estimate the influence of both contrast and luminance masking in each part of the scene.

The visual metrics used in the rendering methods are predominantly based on VDPs [20, 51], often extended to incorporate spatio-temporal contrast sensitivity function (CSF) [34, 63, 64], opponent color processing and chromatic CSF [61], and saliency models [14, 31]. Threshold versus elevation function [23, 72], photoreceptor non-linearity [62], or luminance-dependent CSF is used to model luminance masking, which accounts for the reduced sensitivity of the visual system at low luminance levels. Then, the image is decomposed into spatial-frequency and orientation selective bands using the Cortex transform [62, 99], wavelets [12], the DCT transform [94], or differences-of-Gaussians (DOGs) [26]. The spatial-sensitivity is incorporated either by pre-filtering the image with a CSF [62] or weighting each frequency band according to the CSF sensitivity value for its peak frequency [26, 72]. The multi-band decomposition is necessary to model contrast masking, which is realized either using a contrast transducer function [26, 102] or threshold elevation function [62, 72]. The VDPs can be further weighted by a saliency map, which accounts for low-level attention [31, 72] and/or task-driven high-level attention [14].

Overall, the work on perceptual rendering influenced the way in which the perception is incorporated in graphics. Most methods in graphics rely on the near-threshold visual models and the notion of the just-noticeable-difference (JND). Such near-threshold models offer high accuracy and good rigor since the near-threshold models are well studied in the human vision research. But they also tend to result in over-conservative predictions and are not flexible enough to allow for visible but not disturbing distortions.

9.2.2 Open Source Metrics

The algorithms discussed for far incorporated visual metrics into rendering algorithms, making them difficult to test, compare, or use as a fidelity metric on a pair of test and reference images. These metrics are also complex and hence challenging to reimplement with no source code publicly available. However, the graphics community have several alternative metrics to choose from if they wish to evaluate results without a need to reimplement visual models. *pdiff* [103] is a simple perceptual difference metrics, which utilizes the CIE L*a*b* color space for differences in color, CSF, and model of visual masking from Daly's VDP [20], and some speed improvements from [72]. The C source code is publicly available at http://pdiff.sourceforge.net/. A more complex visual model is offered by the series of HDR-VDP metrics [54, 55], which we discuss in more detail in Sect. 9.2.4. The older version of this metric (HDR-VDP-1.7.1) is available as a C/C++ code, while the latest version is provided as matlab sources (HDR-VDP-2.x). Both versions can be downloaded from http://hdrvdp.sf.net/.

9.2.3 Data-Driven Metrics for Rendering

The majority of image metrics used in graphics rely on the models of the low-level visual perception. These metrics are often constructed by combining components from different visual models, such as saliency models, CSFs, threshold elevation functions, and contrast transducers. While all these partial models well predict the individual effects, there is no guarantee that the combination of them will actually improve predictions. As shown in Sect. 9.4.4.1, complex metrics may actually perform worse in some tasks than a simple arithmetic difference. An alternative to such a white-box approach is the black-box approach, in which the metric is trained to predict differences based on a large data set. In this section we discuss two such metrics, one no-reference and one full-reference metric.

Both metrics rely on the data collected in an experiment, in which observers were asked to label visible artifacts in computer graphics renderings, both when the reference image is shown and when it was hidden. The data set was the same as the one used to metric comparison, discussed in Sect. 9.4.4.1, though the non-reference metric was trained with only ten images from that data set. Example of such manually marked images are shown in the left-most column in Fig. 9.2. As compared to typical image quality databases, such as TID2008 [69], the maps of localized distortions provide much more data for the data-driven training. Instead of assigning a single MOS to each image, the localized distortion maps provide up to a million of such numbers per image, as the labeling is associated with every image pixel. In practice a subsampled version of such a map is used because of limited accuracy of manual labeling. The limitation of localized distortion maps is that they do not provide the estimate of the perceived magnitude of distortion. Instead, the maps contain the probability of detecting an artifact by an average observer.

Since a reference image is usually not available when rendering 3D scenes, Herzog et al. [30] proposed a no-reference image quality metric (NoRM) for three types of rendering distortions: VPL clamping, glossy VPL noise, and shadow map aliasing. In contrast to other non-reference metrics, which can rely solely on a single color image, computer graphics method can provide additional information,

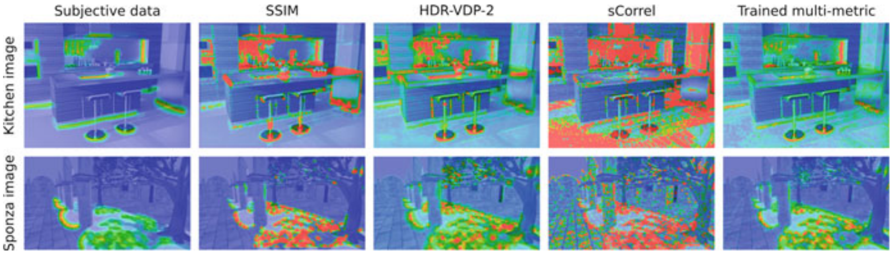


Fig. 9.2 Manually marked distortions in computer graphics rendering (*left*) and the predictions of image quality metrics: SSIM, HDR-VDP-2, sCorrel. Trained multi-metric uses the predictions of the existing metrics as a features for a decision forest classifier. It is trained to predict the subjective data

such as a depth-buffer, or a diffuse material buffer. Such additional information was used alongside the color buffer to solve a rather challenging problem: predict visibility of artifacts given no reference image to compare with. The authors trained a support-vector-machine (SVM) based classifier on ten images with manually labeled artifacts. The features used for training were an irradiance map with removed textures, screen-space ambient occlusion factor, unfolded textures described by the histogram of oriented gradients, a high-pass image with edges eliminated using the joint-bilateral filter and local statistics (mean, variance, skewness, kurtosis). Despite a rather small training set of ten images, the metric was shown to provide comparable or better prediction performance than the state-of-the-art full-reference metrics for the three types of targeted distortions. The authors describe also an application of this metric, in which detected artifacts are automatically corrected by inpainting. The regions with detected artifacts are looked up in a dictionary of artifact-free regions and replaced with a suitable substitute. The operation is illustrated in Fig. 9.3.

The non-reference metrics are specialized in predicting only a certain kind of artifacts as they solve heavily under-constraint problem. Their predictive strength comes from learning the characteristic of a given artifacts and differentiating it from a regular image content. If a metric is to be used for a general purpose and with a wide variety of distortions, it needs to be supplied with both test and reference images.

Čadík et al. [89] explored a possibility of building a more reliable full-reference metric for rendering images using a data-driven approach. The motivation for this work was a previous study, showing mixed performance of existing metrics in this task (discussed in Sect. 9.4.4.1). They identified 32 image difference features, some described by a single number, some by up to 62 dimensions. Features ranged

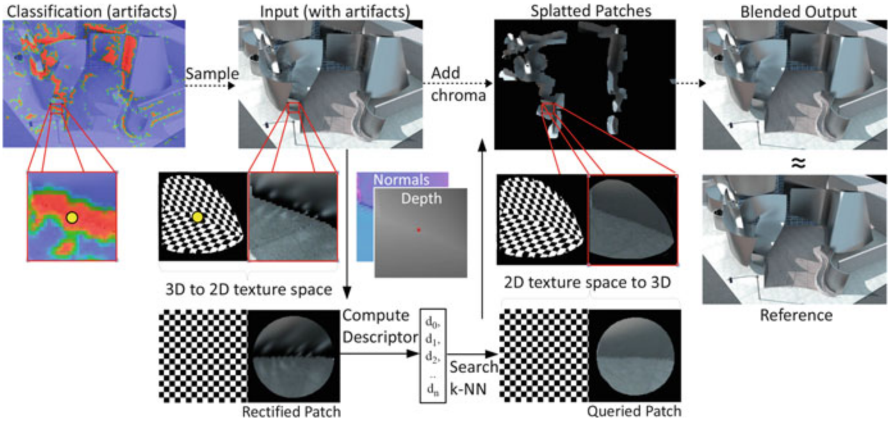


Fig. 9.3 Reduction of artifacts in rendered images by a metric-assisted inpainting [30]. Once the artifacts are detected in an image by a non-reference quality metric, the affected patches are replaced with similar non-distorted patches from the database. The operation is performed in an unfolded 2D texture space. The image courtesy of the authors

from a simple absolute difference to visual attention (measured with an eye-tracker) and included predictions of several major fidelity metrics (SSIM, HDR-VDP-2) and common computer vision descriptors (HOG, Harris corners, etc.). The metric was trained using 37 images with the manually labeled distortion maps. The best performance was achieved with ensembles of bagged decision trees (decision forest) used for classification. The classification was shown to perform significantly better than the best performing general purpose metric (sCorrel) as measured using the leave-one-out cross-validation procedure. Two examples of automatically generated distortion maps are shown in the right-most column of Fig. 9.2 and compared with the predictions of other metrics.

Another example of a data-driven approach to metric design is the no-reference metric for evaluating the quality of motion deblurring, proposed by Liu et al. [50]. Motion deblurring algorithms aim at removing from photographs the motion blur due to camera shake. This is a blind deconvolution problem, in which the blur kernel is unknown. Since usually only blurry image is unavailable, it is essential to provide a mean to measure quality without the need for a sharp reference image. The data for training the metric was collected in a large scale crowd-sourcing experiment, in which over one thousand users ranked in a pairwise comparison experiments 40 scenes, each processed with five different deblurring algorithms. The metric was trained as a logistic regression explaining the relation between a number of features and the scaled subjective scores. The features included several no-reference measures of noise, sharpness, ringing, and sharpness. In a dedicated validation experiment, the trained no-reference metric performed comparably or better than the state-of-the-art full-reference metrics. The authors suggested several applications of the new metric, such as automatic selection of the deblurring algorithm which performs the best for a given image, or, on a local level, fusing high quality image by picking different image fragments from the result of each deblurring algorithm.

9.2.4 HDR Metrics for Rendering

The majority of image quality metrics consider quality assessment for one particular medium, such as an LCD display or a print. However, the results of physically accurate computer graphics methods are not tied to any concrete device. They produce images in which pixels contain linear radiometric values, as opposed to the gamma-corrected RGB values of a display device. Furthermore, the radiance values corresponding to real-world scenes can span a very large dynamic range, which exceeds the contrast range of a typical display device. Hence the problem arises of how to compare the quality of such images, which represent actual scenes, rather than their tone-mapped reproductions.

Aydin et al. [6] proposed a simple luminance encoding that makes it possible to use PSNR and SSIM [97] metrics with HDR images. The encoding transforms physical luminance values (represented in $\mathrm{cd/m^2}$) into an approximately perceptually uniform representation (refer to Fig. 9.4). The transformation is derived

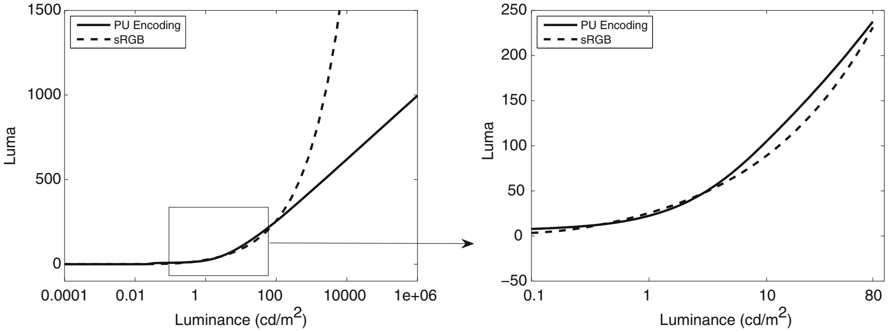


Fig. 9.4 Perceptually uniform (PU) encoding for evaluating quality of HDR images. The absolute luminance values are converted into luma values before they are used with standard image quality metrics, such as MSE, PSNR, or SSIM. Note that the PU encoding is designed to give a good fit to the sRGB non-linearity within the range 0.1–80 cd/m^2 so that the results for low dynamic range images are consistent with those performed in the sRGB color space

from luminance detection data using the threshold-integration method, similar to the one used for contrast transducer functions [102]. The transformation is further constrained so that the luminance values produced by a typical CRT display (in the range 0.1–80 cd/m^2) are mapped to 0–255 range to mimic the sRGB non-linearity. This way, the quality predictions for typical low-dynamic range images are comparable to those calculated using pixel values. However, the metric can also operate in a much greater range of luminance.

The pixel encoding of Aydin et al. accounts for luminance masking, but it does not account for other luminance-dependent effects, such as inter-ocular light scatter or the frequency shift of the CSF peak with luminance. Those effects were modeled in the visual difference predictor for high dynamic range images (HDR-VDP) [54]. The HDR-VDP extends Daly's VDP [20] to predict differences in HDR images. In 2011 the metric was superseded with a completely redesigned metric HDR-VDP-2 [55], which is discussed below.

HDR-VDP-2 is the visibility (discrimination) and quality metric capable of detecting differences in achromatic images spanning a wide range of absolute luminance values [55]. Although the metric originates from the classical VDP [20], and its extension—HDR-VDP [54], the visual models are very different from those used in those earlier metrics. The metric is also an effort to design a comprehensive model of the contrast visibility for a very wide range of illumination conditions.

As shown in Fig. 9.5, the metric takes two HDR luminance or radiance maps as input and predicts the probability of detecting a difference between the pair of images (P_{map} and P_{det}) as well as the quality (Q and Q_{MOS}), which is defined as the perceived level of distortion.

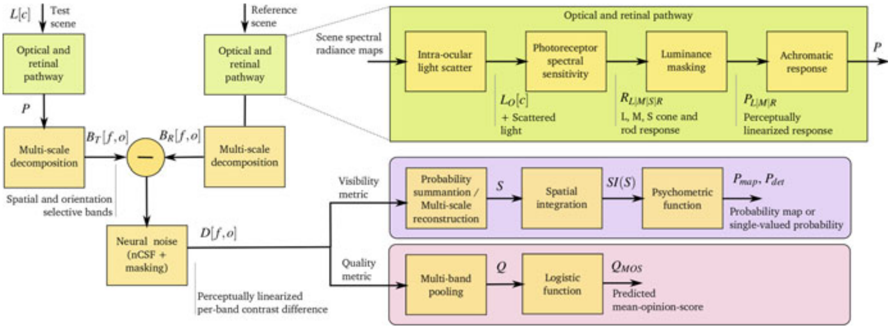

L[c]
Test scene
Optical and retinal pathway
P
Multi-scale decomposition
Spatial and orientation selective bands
$B_T[f,o]$
−
$B_R[f,o]$
Reference scene
Optical and retinal pathway
Multi-scale decomposition
Neural noise (nCSF + masking)
D[f,o]
Perceptually linearized per-band contrast difference
Scene spectral radiance maps
Optical and retinal pathway
Intra-ocular light scatter
Photoreceptor spectral sensitivity
Luminance masking
Achromatic response
P
$L_O[c]$
+ Scattered light
$R_{L|M|S|R}$
L, M, S cone and rod response
$P_{L|M|R}$
Perceptually linearized response
Visibility metric
Probability summantion / Multi-scale reconstruction
S
Spatial integration
SI(S)
Psychometric function
P_{map}, P_{det}
Probability map or single-valued probability
Quality metric
Multi-band pooling
Q
Logistic function
Q_{MOS}
Predicted mean-opinion-score

Fig. 9.5 The processing stages of the HDR-VDP-2 metric. Test and reference images undergo similar stages of visual modeling before they are compared at the level of individual spatial-and-orientation selective bands (B_T and B_R). The difference is used to predict both visibility (probability of detection) and quality (the perceived magnitude of distortion)

One of the major factors limiting the contrast perception in high contrast (HDR) scenes is the scattering of the light in the optics of the eye and on the retina [58]. The HDR-VDP-2 models it as a frequency-space filter, which was fitted to an appropriate data set (*inter-ocular light scatter* block in Fig. 9.5). The contrast perception deteriorates at lower luminance levels, where the vision is mediated mostly by night-vision photoreceptors—rods. This is especially manifested for small contrasts, which are close to the detection threshold. This effect is modeled as a hypothetical response of the photoreceptor (in steady state) to light (*luminance masking* block in Fig. 9.5). Such response reduces the magnitude of image difference for low luminance according to the contrast detection measurements. The masking model (*neural noise* block in Fig. 9.5) operates on the image decomposed into multiple orientation-and-frequency-selective bands to predict the threshold elevation due to contrast masking. Such masking is induced both by the contrast within the same band (intra-channel masking) and within neighboring bands (inter-channel masking). The same masking model incorporates also the effect of neural CSF, which is the contrast sensitivity function without the sensitivity reduction due to inter-ocular light scatter. Combining neural CSF with masking model is necessary to account for contrast constancy, which results in "flattening" of the CSF at the super-threshold contrast levels [27].

Figure 9.6 demonstrates the metric prediction for blur and noise. The model has been shown to predict numerous discrimination data sets, such as ModelFest [98], historical Blackwell's t.v.i. measurements [9], and newly measured CSF [35]. The source code of the metric is freely available for download from http://hdrvdp.sourceforge.net. It is also possible to run the metric using an on-line web service at http://driiqm.mpi-inf.mpg.de/.

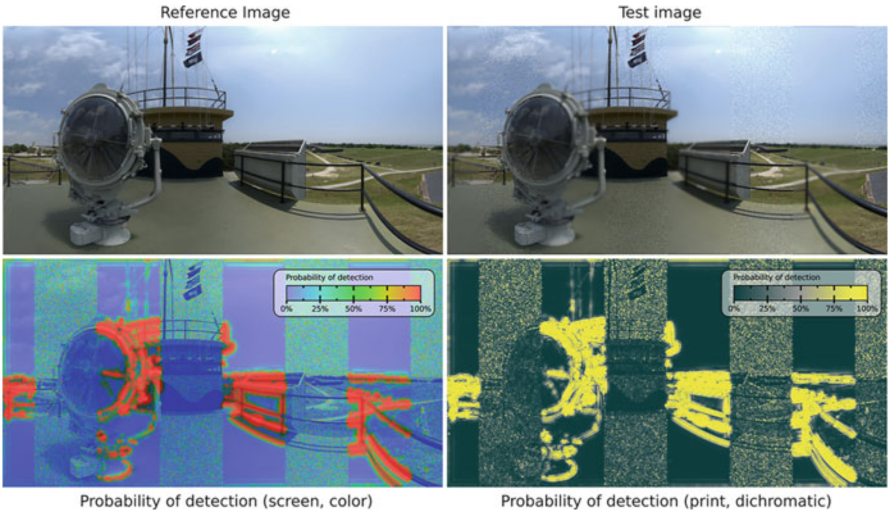


Fig. 9.6 Predicted visibility differences between the test and reference images. The test image contains interleaved vertical stripes of blur and white noise. The images are tone-mapped versions of an HDR input. The two color-coded maps on the right represent the probability that an average observer will notice a difference between the image pair. Both maps represent the same values, but use different color maps, optimized either for screen viewing or for grayscale/color printing. The probability of detection drops with lower luminance (luminance sensitivity) and higher texture activity (contrast masking). Image courtesy of HDR-VFX, LLC 2008

9.2.5 Tone-Mapping Metrics

Tone-mapping is the process of transforming an image represented in approximately physically accurate units, such as radiance and luminance, into pixel values that can be displayed on a screen of a limited dynamic range. Tone-mapping is a part of an image processing stack of any digital camera. A “raw” images captured by a digital sensor would produce unacceptable results if they were mapped directly to pixel values without any tone-mapping. But similar process is also necessary for all computer graphics methods that produce images represented in physical units. Therefore, the problem of tone-mapping and the quality assessment of tone-mapping results have been extensively studied in graphics.

Tone-mapping inherently produces images that are different from the original HDR reference. In order to fit the resulting image within available color gamut and dynamic range of a display, tone-mapping often needs to compress contrast and adjust brightness. Tone-mapped image may lose some quality as compared to the original seen on a HDR display, yet the images look often very similar and the degradation of quality is poorly predicted by most quality metrics. Smith et al. [82] proposed the first metric intended for predicting loss of quality due to local and global contrast distortion introduced by tone-mapping. However, the metric was only used in the context of controlling countershading algorithm and was not

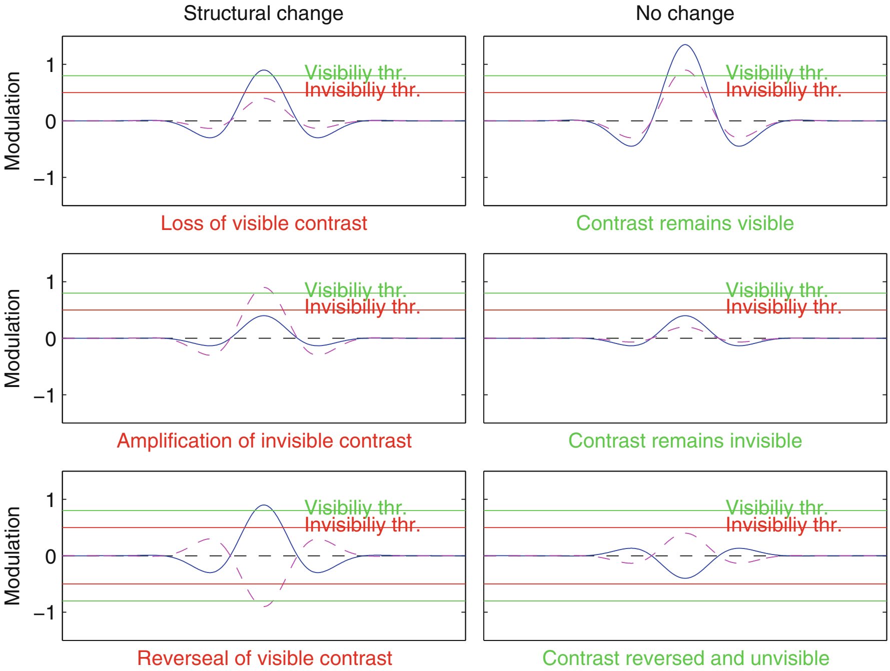


Fig. 9.7 The dynamic range independent metric [5] distinguished between the change of contrast that does and does not result in structural change. *Blue continuous line* shows a reference signal (from a band-pass pyramid) and *magenta dashed line* the test signal. When contrast remains visible or invisible after tone-mapping, no distortion is signalized (*top* and *middle right*). However, when the change of contrast alters the visibility of details, making visible details becoming invisible (*top-left*), it is signalized as a distortion

validated against experimental data. Aydin et al. [5] proposed a metric for comparing HDR and tone-mapped images that is robust to contrast changes. The metric was later extended to video [7]. Both metrics are invariant to the change of contrast magnitude as long as that change does not distort contrast (inverse its polarity) or affect its visibility. The metric classifies distortions into three types: loss of visible contrast, amplification of invisible contrast, and contrast reversal. All three cases are illustrated in Fig. 9.7 on an example of a simple 2D Gabor patch. These three cases are believed to affect the quality of tone-mapped images. Figure 9.8 shows the metric predictions for three tone-mapped images. The main weakness of this metric is that produced distortion maps are suitable mostly for visual inspection and qualitative evaluation. The metric does not produce a single-valued quality estimate and its correlation with subjective quality assessment has not been verified.

Yeganeh and Wang [105] proposed a metric for tone-mapping, which was designed to predict on overall quality of a tone-mapped image with respect to an HDR reference. The first component of the metric is the modification of the SSIM [97], which includes the contrast and structure components, but does not include the luminance component. The contrast component is further modified to detect only the cases in which invisible contrast becomes visible and visible contrast becomes invisible, in a similar spirit as in the dynamic range independent metric [5], described above. This is achieved by mapping local standard deviation values used in the contrast component into detection probabilities using a visual model, which consists of a psychometric function and a CSF. The second component of the metric describes "naturalness." The naturalness is captured by the measure of similarity between the histogram of a tone-mapped image and the distribution of histograms from the database of 3,000 low-dynamic range images. The histogram is approximated by the Gaussian distribution. Then, its mean and standard deviation is compared against the database of histograms. When both values are likely to be found in the database, the image is considered natural and is assigned a higher quality. The metric was tested and cross-validated using three databases, including one from [91] and authors' own measurements. The Spearman rank-order correlation coefficient (SROC) between the metric predictions and the subjective data was reported to be approximately 0.8. Such value is close to the performance of a random observer, which is estimated as the correlation between the mean and random observer's quality assessment.

Some visible distortions are desirable as long as they are not objectionable. An example of that is contrast enhancement through unsharp masking (high spatial frequencies) or countershading (low spatial frequencies) [37], commonly used in tone-mapping. In both cases, smooth gradients are introduced at both sides of an edge in order to enhance the contrast of that edge. This is demonstrated in Fig. 9.9 where the base contrast shown in the bottom row is enhanced by adding countershading profiles. Note that the brightness of the central part of each patch

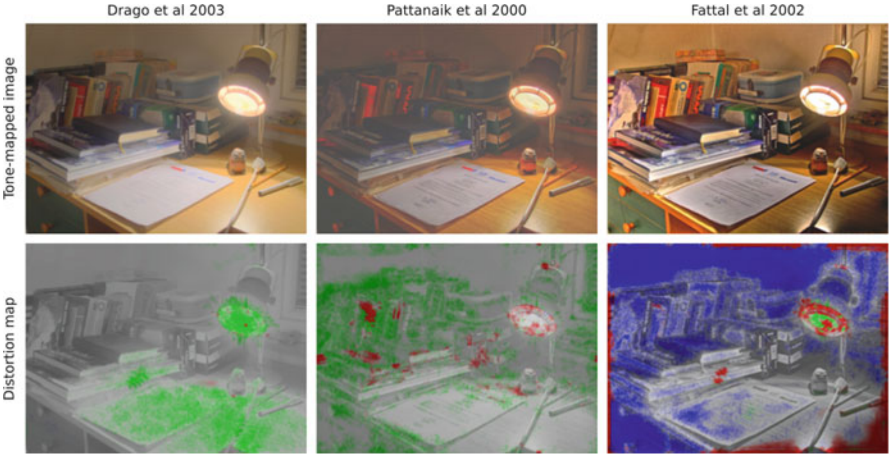


Fig. 9.8 Prediction of the dynamic range independent metric [5] (*top*) for tone-mapped images (*bottom*). The *green color* denotes the loss of visible contrast, the *blue color* the amplification of invisible contrast and the *red color* is contrast, reversal (refer to Fig. 9.7)

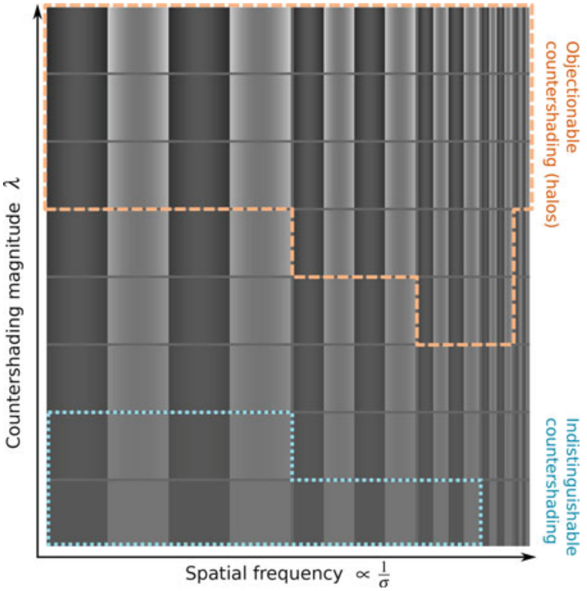

Countershading magnitude λ
Spatial frequency $\propto \frac{1}{\sigma}$
Objectionable countershading (halos)
Indistinguishable countershading

Fig. 9.9 Contrast enhancement by countershading. The figure shows the square-wave pattern with a reduced amplitude of the fundamental frequency, resulting in countershading profiles. The regions of indistinguishable (from a step edge) and objectionable countershading are marked with *dotted* and *dashed lines* of different color. The higher magnitude of countershading produces higher contrast edges. But if it is too high, the result appears objectionable. The marked regions are approximate and for illustration and actual regions will depend on the angular resolution of the figure

remains the same across all rows. The region marked with the blue dashed line denotes the range of the Cornsweet illusion, where the gradient remains invisible while the edge is still enhanced. Above that line the Cornsweet illusion breaks and the gradients become visible. In practice, when countershading is added to tone-mapped images, it is actually desirable to introduce such visible gradients. Otherwise, the contrast enhancement is too small and does not improve image quality. But too strong gradient results in visible contrast reversal, also known as "halo" artifact, which is disturbing and objectionable. Trentacoste et al. [86] measured the threshold when countershading profiles become objectionable in complex images. They found that the permissible strength of the countershading depends on the width of the gradient profile, which in turn depends on the size of an image. They proposed a metric predicting the maximum strength of the enhancement and demonstrated its application to tone-mapping. The metric is an example of a problem where it is more important to predict when an artifact becomes objectionable rather than just visible.

9.2.6 *Aesthetics and Naturalness*

Many quality assessment problems in graphics cannot be easily addressed by objective image and video metrics because they involve high-level concepts, such as aesthetics or naturalness. For example, there is no computational algorithm that could tell whether an animation of a human character looks natural, or whether a scene composition looks pleasing to the eye. Yet, such tasks are often the goals of graphics methods. The common approach to such problems is to find a suitable set of numerical features that could correlate with subjective assessment, collect a large data set of subjective responses and then use machine learning techniques to train a predictor. Such methods proved to be effective for selecting the best viewpoint of a mesh [78], or selecting color palettes for graphic designs [65]. Yet, it is hard to expect that a suitable metric will be found for each individual problem. Therefore, graphics more often needs to rely on efficient subjective methods, which are discussed in Sect. 9.4.

9.3 Quality Metrics for 3D Models

The previous section focused on the quality evaluation of 2D images coming from computer graphics methods, mostly from rendering, HDR imaging, or tone-mapping. Hence most of the involved metrics aimed to detect specific image artifacts like aliasing, structured noise due to global illumination or halo artifacts from tone-mapping. However, in computer graphics, visual artifacts do not come only from the final image creation process but they can occur on the 3D data themselves before the rendering. Indeed, 3D meshes are now subject to a wide range of processes which include transmission, compression, simplification, filtering, watermarking, and so on. These processes inevitably introduce distortions which alter the geometry or texture of these 3D data and thus their final rendered appearance. Hence quality metrics have been introduced to detect these specific 3D artifacts, i.e. geometric quantization noise, smooth deformations due to watermarking, simplification artifacts, and so on. A comprehensive review has been recently published about 3D mesh quality assessment [19]. Two kinds of approaches exist for this task: model-based and image-based approaches. Model-based approaches operate directly on the geometry and/or texture of the meshes being compared while image-based approaches consider rendered images of the 3D models (i.e., snapshots from different viewpoints) to evaluate their visual quality. Note that some image-based quality assessment algorithms consider only some specific viewpoints and thus are view-dependent.

9.3.1 Model-Based Metrics

In the fields of computer graphics, the first attempts to evaluate the visual fidelity of 3D objects were simple geometric distances, mainly used for driving mesh simplification [77]. A widely used metric is the Hausdorff distance, defined as follows:

$$H_a(M_1, M_2) = \max_{\mathbf{p} \in M_1} e(\mathbf{p}, M_2) \tag{9.1}$$

with M_1 and M_2, the two 3D objects to compare and $e(\mathbf{p}, M)$ the Euclidean distance from a point $\mathbf{p}$ in the 3D space to the surface M. This value is asymmetric; a symmetrical Hausdorff distance is defined as follows:

$$H(M_1, M_2) = \max \{H_a(M_1, M_2), H_a(M_2, M_1)\} \tag{9.2}$$

We can also define an asymmetric mean square error:

$$MSE_a(M_1, M_2) = \frac{1}{|M_1|} \int_{M_1} e(\mathbf{p}, M_2)^2 \, ds \tag{9.3}$$

The most widespread measurement is the Maximum Root Mean Square Error (MRMS):

$$MRMS(M_1, M_2) = \max \left\{ \sqrt{MSE_a(M_1, M_2)}, \sqrt{MSE_a(M_2, M_1)} \right\} \tag{9.4}$$

Cignoni et al. [16] provided the *Metro* software[1] with an implementation of Hausdorff and MRMS geometric distances between 3D models.

However these simple geometric measures are very poor predictor of the visual fidelity, like demonstrated in several studies [44, 88]. Hence, researchers have introduced perceptually motivated metrics. These full-reference metrics compare the distorted and original 3D models to compute a score which reflects the visual fidelity.

Karni and Gotsman [32], in order to evaluate properly their compression algorithm, consider the mean geometric distance between corresponding vertices and the mean distance of their geometric Laplacian values (which reflect a degree of smoothness of the surface) (this metric is abbreviated as $GL1$ in Table 9.1). Subsequently, Sorkine et al. [83] proposed a different version of this metric ($GL2$), which assumes slightly different values of the parameters involved. The performance of these metrics in terms of visual quality prediction remains low.

Several authors use the curvature information to derive perceptual quality metrics. Lavoué et al. [45] introduce the mesh structural distortion measure (MSDM)

[1] http://vcg.isti.cnr.it/activities/surfacegrevis/simplification/metro.html.

Table 9.1 Correlation between Mean-Opinion-Scores and values from the metrics for four publicly available subjective databases

	Masking [41]		Simplification [79]		General purpose [45]		Compression [88]	
	SROCC	LCC	SROCC	LCC	SROCC	LCC	SROCC	LCC
Hausdorff	0.27	0.20	0.49	0.51	0.14	0.11	0.25	0.14
RMS	0.49	0.41	0.64	0.59	0.27	0.28	0.52	0.49
GL1 [32]	0.42	0.40	NA	NA	0.33	0.35	0.67	0.71
GL2 [83]	0.40	0.38	NA	NA	0.39	0.42	0.74	0.76
3DWPM1 [18]	0.29	0.32	NA	NA	0.69	0.62	0.82	0.84
3DWPM2 [18]	0.37	0.43	NA	NA	0.49	0.50	0.81	0.82
MSDM [45]	0.65	0.69	NA	NA	0.74	0.75	**0.83**	**0.91**
MSDM2 [42]	**0.90**	**0.87**	**0.87**	**0.89**	**0.80**	**0.81**	**0.78**	**0.89**
FMPD [96]	0.80	0.81	**0.87**	**0.89**	**0.82**	**0.84**	**0.82**	**0.89**
DAME [88]	0.68	0.59	NA	NA	0.77	0.75	**0.86**	**0.94**
TPDM [85]	**0.87**	**0.87**	**0.87**	**0.86**	**0.85**	**0.85**	No data	No data
Learning [43]	**0.90**	**0.90**	NA	NA	**0.86**	**0.86**	No data	No data

Best values for each database are highlighted in bold

The correlations are computed for whole databases, with the exception of the compression database, where per-model averages were used, since the data acquiring procedure does not capture inter-model relations which follows the concept of structural similarity introduced for 2D image quality assessment by Wang et al. [97] (SSIM index). The local $LMSDM$ distance between two mesh local windows a and b is defined as follows:

$$LMSDM(a,b) = (\alpha L(a,b)^3 + \beta C(a,b)^3 + \gamma S(a,b)^3)^{\frac{1}{3}} \tag{9.5}$$

L, C, and S represent, respectively, curvature, contrast, and structure comparison functions:

$$\begin{aligned} L(a,b) &= \frac{\|\mu_a - \mu_b\|}{\max(\mu_a, \mu_b)} \\ C(a,b) &= \frac{\|\sigma_a - \sigma_b\|}{\max(\sigma_a, \sigma_b)} \\ S(a,b) &= \frac{\|\sigma_a \sigma_b - \sigma_{ab}\|}{\sigma_a \sigma_b} \end{aligned} \tag{9.6}$$

with μ_a, σ_a, and σ_{ab} are, respectively, the mean, standard deviation, and covariance of the curvature over the local windows a and b. A *local window* is defined as a connected set of vertices belonging to a sphere with a given radius. The global $MSDM$ measure between two meshes is then defined by a Minkowski sum of the local distances associated with all local windows; it is a visual distortion index ranging from 0 (objects are identical) to 1 (theoretical limit when objects are completely different). A multi-resolution improved version, named $MSDM2$, has recently been proposed in [42]. It provides better performance and allows one to compare meshes with arbitrary connectivities. Torkhani et al. [85] introduced a similar metric called $TPDM$ (Tensor-based Perceptual Distance Measure) which takes into account not only the mesh curvature amplitude but also the principal curvature directions. Their motivation is that these directions represent structural features of the surface and thus should be visually important. These metrics own the benefit of providing also a distortion map that predicts the perceived local artifacts visibility, like illustrated in Fig. 9.10.

Váša and Rus [88] consider the per-edge variations of oriented dihedral angles for visual quality assessment. The angle orientation allows to distinguish between convex and concave angles. Their metric ($DAME$ for Dihedral Angle Mesh Error) is obtained by summing up the dihedral angle variations for all edges of the mesh being compared, as follows:

$$DAME = \frac{1}{n_e} \sum_{n_e} \|\alpha_i - \bar{\alpha_i}\| .m_i.w_i \tag{9.7}$$

with n_e the number of edges of the meshes being compared, α_i and $\bar{\alpha_i}$ the respective dihedral angles of the ith edge of the original and distorted mesh. m_i is a weighting term relative to the masking effect (enhancing the distortion on smooth surfaces where they are the most visible). w_i sis a weighting term relative to the surface

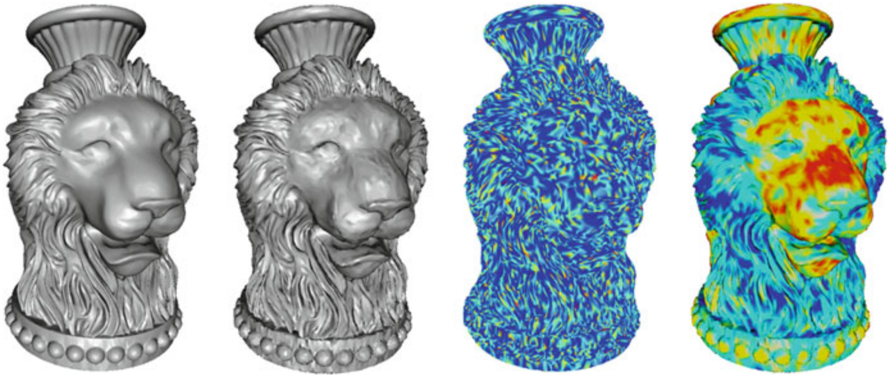


Fig. 9.10 *From left to right*: The Lion model; a distorted version after random noise addition; Hausdorff distortion map; MSDM2 distortion map. *Warmer colors* represent higher values

visibility; indeed, a region almost always invisible should not contribute to the global distortion. This metric has the advantage of being very fast to compute but only works for comparing meshes of shared connectivity.

The metrics presented above consider local variations of attribute values at vertex or edge level, which are then pooled into a global score. In contrast, Corsini et al. [18] and Wang et al. [96] compute one global roughness value per 3D model and then derive a simple global roughness difference to derive a visual fidelity value between two 3D models. Corsini et al. [18] propose two ways of measuring the global model roughness. The first one is based on statistical considerations (at multiple scales) about the dihedral angles and the second one computes the variance of the geometric differences between a smoothed version of the model and its original version. These metrics are abbreviated as $3DWPM1$ and $3DWPM2$ in Table 9.1. Wang et al. [96] define the global roughness of a 3D model as a normalized surface integral of the local roughness, defined as the Laplacian of the discrete Gaussian curvature. The local roughness is modulated to take into account the masking effect. Their metric ($FMPD$ for Fast Mesh Perceptual Distance) provides good results and is fast to compute. Moreover a local distortion map can be obtained by differencing the local roughness values. Figure 9.11 illustrates some distorted versions of the *Horse* 3D model, with their corresponding $MRMS$, $MSDM2$, and $FMPD$ values.

Given the fact that all metrics above rely on different features, e.g. curvature computations [42, 45, 85], dihedral angles [18, 88], Geometric Laplacian [32, 83], and Laplacian of Gaussian curvature [96]. Lavoué et al. [43] have hypothesized that a combination of these attributes could deliver better results that using them separately. They propose a quality metric based on an optimal linear combination of these attributes determined through machine learning. They obtained a very simple model which still provides good performance.

Some authors also proposed quality assessment metrics for textured 3D mesh [67, 84] dedicated to optimizing their compression and transmission. These metrics,

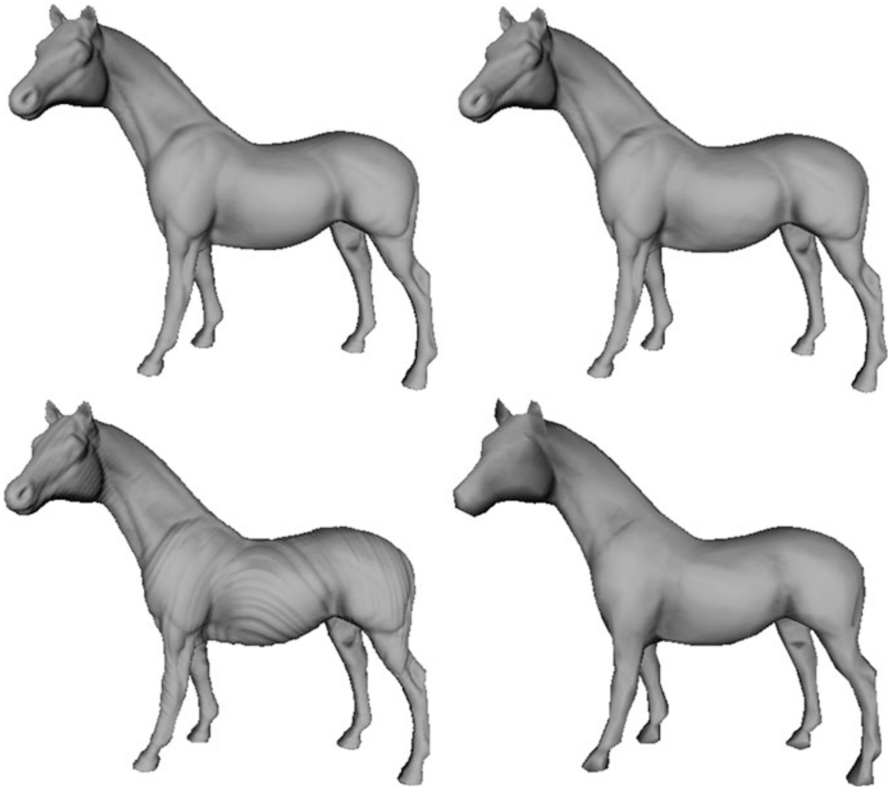


Fig. 9.11 Distorted versions of the *Horse* model, all associated with the same maximum root mean square error ($MRMS = 0.00105$). *From left to right, top to bottom*: Original model; result after watermarking from Wang et al. [95] (MSDM2 = 0.14, FMPD = 0.01); result after watermarking from Cho et al. [15] (MSDM2 = 0.51, FMPD = 0.40), result after simplification [48] from 113K vertices to 800 vertices (MSDM2 = 0.62, FMPD = 1.00)

respectively, rely on geometry and texture deviations [84] and on texture and mesh resolutions [67]. Their results underline the fact that the perceptual contribution of image texture is, in general, more important than the model's geometry, i.e. the reduction of the texture resolution is perceived more degraded than the reduction of model's polygons (geometry resolution).

For dynamic meshes, the most used metric is the KG error [33]. Given M_1 and M_2 the matrix representations ($3v \times f$ with v and f, respectively, the number of vertices and frames, 3 stands for the number of coordinates—x,y,z) of two dynamic meshes to compare, the KG error is defined as a normalized Frobenius norm of the matrix difference $\|M_1 - M_2\|$. Like the RMS for static meshes, this error metric does not correlate with the human vision. Váša and Skala have introduced a perceptual metric [87] for dynamic meshes, the STED error (Spatio-Temporal Edge Difference). The metric works on edges as basic primitives, and computes the relative change in length for each edge of the mesh in each frame of the animation. This quality metric is able to capture both spatial and temporal artifacts and correlates well with the human vision.

Guthe et al. [28] introduce a perceptual metric based on spatio-temporal CSF dedicated to bidirectional texture functions (BTFs), commonly used to represent the appearance of complex materials. This metric is used to measure the visual quality of the various compressed representations of BTF data.

Ramanarayanan et al. [71] proposed the concept of *visual equivalence* in order to create a metric that is more tolerant for non-disturbing artifacts. The authors proposed that two images are considered visually equivalent if object's shape and material are judged to be the same in both images and in a side-by-side comparison, an observer is unable to tell which image is closer to the reference. The authors proposed an experimental method and a metric (Visual Equivalence Predictor) based on the machine learning techniques (SVM). The metric associates simple geometry and material descriptors with the samples measured in the experiments. Then, a trained classifier determines whether the distortions in illumination map lead to visually equivalent results. The metric demonstrated an interesting concept, yet it can be used only with a very limited range of illumination distortions. This work is dedicated to the evaluation of illumination map distortion effect, and not to the evaluation of the 3D model quality. However, it relies on geometry and material information and thus can be classified as a model-based metric.

9.3.2 Image-Based Metrics

Apart from these quality metrics operating on the 3D geometry (that we call model-based), a lot of researchers have used 2D image metrics to evaluate the visual quality of 3D graphical models. Indeed, as pointed out in [49], the main benefit of using image metrics to evaluate the visual quality of 3D objects is that the complex interactions between the various properties involved in the appearance (geometry, texture, normals) are naturally handled, avoiding the problem of how to combine and weight them. Many image-based quality evaluation works have been proposed in the context of simplification and level-of-detail (LoD) management for rendering. Among existing 2D metrics, authors have considered the Sarnoff visual discrimination model (VDM) [51], the visible difference predictor (VDP) from Daly [20] (both provide local distortion maps that predict local perceived differences), but also the SSIM (Structural SIMilarity) index, introduced by Wang and Bovik [97] and the classical mean or root mean squared pixel difference.

Lindstrom and Turk [49] evaluate the impact of simplification using a fast image quality metric (RMS error) computed on snapshots taken from 20 different camera positions regularly sampled on a bounding sphere. Their approach is illustrated in Fig. 9.12. In his Ph.D. thesis [47], Lindstrom also proposed to replace the RMS by perceptual metrics including the Sarnoff VDM and surprisingly he found that the RMS error yields to better results. He also found that his image-based approach

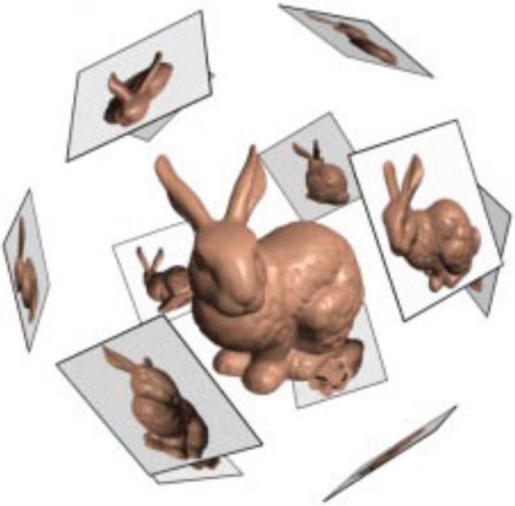


Fig. 9.12 Illustration of the image-based simplification approach from Lindstrom and Turk [49]. This algorithm considers the quality of 2D snapshots sampled around the 3D mesh as the main criterion for decimation. Image reprinted from [47]

provides better results than geometry-driven approaches, however he considered a similar image-based evaluation. Qu and Meyer [70] consider the visual masking properties of 2D texture maps to drive simplification and remeshing of textured meshes, they evaluate the potential masking effect of the surface signals (mainly the texture) using the 2D Sarnoff VDM [51]. The masking map is obtained by comparing, using VDM, the original texture map with a Gaussian filtered version. The final remeshing can be view-independent or view-dependent depending on the visual effects considered. Zhu et al. [109] studied the relationship between the viewing distance and the perceptibility of model details using 2D metrics (VDP and SSIM) for the optimal design of discrete LOD for the visualization of complex 3D building facades.

For animated characters, Larkin and O'Sullivan [40] ran an experiment to determine the influence of several types of artifacts (texture, silhouette, and lighting) caused by simplification; they found that silhouette is the dominant artifact and then devised a quality metric based on silhouette changes suited to drive simplification. Their metric is as follows: they render local regions containing silhouette areas from different viewpoints and compare the resulting images with a 2D quality metric [103].

Several approaches do not rely directly on 2D metrics but rather on psychophysical models of visual perception (mostly the CSF). One of the first study of this kind was that of Reddy [73], which analyzed the frequency content in several prerendered images to determine the best LOD to use in a real-time rendering system. Luebke and Hallen [52] developed a perceptually based simplification algorithm based on a simplified version of the CSF. They map the change resulting from a local simplification operation to a worst-case contrast and a worst-case frequency and then determine whether this operation will be imperceptible. Their method was then extended by Williams et al. [101] to integrate texture and lighting effects. These latter approaches are view-dependent. Menzel and Guthe [60] propose a perceptual model of JND (Just-Noticeable-Difference) to drive their simplification algorithm; it integrates CSF and masking effect. The strength of their algorithm is to be able to perform almost all the calculation (i.e., contrast and frequency) directly on vertices instead of rendered images. However, it still uses the rendered views to evaluate the masking effect, thus it can be classified as an hybrid image-based/model-based method.

9.4 Subjective Quality Assessment in Graphics

Quality assessment metrics presented in Sects. 9.2 and 9.3 aim at *predicting* the visual quality and/or the local artifact visibility in graphics images and 3D models. Both these local and global perceived qualities can also be directly and quantitatively assessed by means of *subjective quality assessment experiments*. In such experiments, human observers give their opinion about the perceived quality or artifact visibility for a corpus of distorted images or 3D models.

Subjective experiments also provide a mean to test objective metrics. The nonparametric correlation, such as Spearman's or Kendall's rank-order correlation coefficients, computed between subjective scores and the objective scores provides an indicator of the performance of these metrics and a way to evaluate them quantitatively. We discuss some work in graphics on evaluation of objective metrics in Sect. 9.4.4.

For global quality assessment, many protocols exist and have been used for graphics data. Usually, absolute rating, double stimulus rating, ranking or pairwise comparisons are considered. Mantiuk et al. [56] compared the sensitivity and experiment duration for four experimental methods: single stimulus with a hidden reference, double stimulus, pairwise comparisons, and similarity judgments. They found that the pairwise comparison method results in the lowest variation between observer's scores. Surprisingly, the method also required the shortest time to complete the experiments even for a large number of compared methods. This was believed to be due to the simplicity of the task, in which a better of two images was to be selected.

9.4.1 Scaling Methods

Once experimental data is collected, it needs to be scaled into a mean a quality measure for a group of observers. Because different observers are likely to use different scale when rating images, their results need to be unified. The easiest way to make their data comparable is to apply a linear transform that makes the mean and the standard deviation equal for all observers. The result of such a transform is called z-score and is computed as

$$z_{i,j,k,r} = \frac{d_{i,j,k,r} - \bar{d}_i}{\sigma_i}, \tag{9.8}$$

where the mean score $\bar{d}_i$ and standard deviation σ_i are computed across all stimuli rated by an observer i. The indexes correspond to i-th observer, j-th condition (algorithm), k-th stimuli (image, video, etc.), and r-th repetition.

Pairwise comparison experiments require different scaling procedures, usually based on Thurstone Case IV or V assumptions [25]. These procedures attempt to convert the results of pairwise comparisons into a scale of JNDs. When 75 % of observers select one condition over another, the quality difference between them is assumed to be 1 JND. The scaling methods that tend to be the most robust are based on the maximum likelihood estimation [3,81]. They maximize the probability that the scaled JND values explain the collected experimental data under the Thurstone Case V assumptions. The optimization procedure finds a quality value for each stimulus that maximizes the probability, which is modeled by the binomial distribution. Unlike standard scaling procedures, the probabilistic approach is robust to unanimous answers, which are common when a large number of conditions are compared. The detailed review of the scaling methods can be found in [25].

9.4.2 Specificity of Graphics Subjective Experiments

9.4.2.1 Global vs. Local

Artifacts coming from transmission or compression of natural images (i.e., blockiness, blurring, ringing) are mostly uniform. In contrast, artifacts from graphics processing or rendering are more often nonuniform. Therefore, this domain needs visual metrics able to distinguish local artifacts visibility rather than global quality. Consequently, many experiments involving graphical content involve locally marking noticeable and objectionable distortions [90] rather than judging an overall quality. This marking task is more complicated than a quality rating, thus it involves the creation of innovative protocols.

9.4.2.2 Large Number of Parameters

A subjective experiment usually involves a number of important parameters; for instance, for evaluating the quality of images or videos, one has to decide the corpus of data, the nature and amplitude of the distortions as well as the rating protocol (i.e., single or multiple stimulus, continuous or category rating, etc). However, the design of a subjective study involving 3D graphical content requires many additional parameters (as raised in [13]):

- Lighting. As raised in the experiment of Rogowitz and Rushmeier [74], the position and type of light source(s) have a strong influence on the perception of the artifacts.
- Materials and Shading. Complex materials and shaders may enhance the artifacts visibility, or on the contrary, act as a masker (in particular some texture patterns [26]).
- Background. The background may affect the perceived quality of the 3D model, in particular it influences the visibility of the silhouette, which strongly influences the perception.
- Animation and interaction. There exist different ways to display the 3D models to the observers, from the most simple (e.g., as a static image from one given viewpoint, as in [100]) to the most complex (e.g., by allowing free rotation, zoom, translation, as in [18]). Of course it is important for the observer to have access to different viewpoints of the objects, however the problem of allowing free interaction is the cognitive overload that may alter the results. A good compromise may be the use of animations, as in [67], however the velocity strongly influences the CSF [34], hence animations have to be reasonably slow.

9.4.2.3 Specifics of Tone-Mapping Evaluation

In this section we discuss the importance of selecting the right reference and an evaluation method for subjective evaluation of tone-mapping operators. This section serves as an example of the considerations that are relevant when considering quality assessment in graphics applications. Similar text has been published before in [24].

Figure 9.13 illustrates a general tone-mapping scenario and a number of possible evaluation methods. To create an HDR image, the physical light intensities (luminance and radiance) in a scene are captured with a camera or rendered using computer graphics methods. In the general case, “RAW” camera formats can be considered as HDR formats, as they do not alter captured light information given a linear response of a CCD/CMOS sensor. In the case of professional content production, the creator (director, artist) seldom wants to show what has been captured in a physical scene. The camera-captured content is edited, color-graded, and enhanced. This can be done manually by a color artist or automatically by color processing software. It is important to distinguish this step from actual tone-mapping, which, in our view, is meant to do “the least damage” to the appearance of carefully edited content. In some applications, such as simulators or realistic visualization, where faithful reproduction is crucial, the enhancement step is omitted.

Tone-mapping can be targeted for a range of displays, which may differ substantially in their contrast and brightness levels. Even HDR displays require tone-mapping as they are incapable of reproducing the luminance levels found in the real world. An HDR display, however, can be considered as the best possible reproduction available, or a “reference” display. Given such a tone-mapping pipeline, we can distinguish the following evaluation methods:

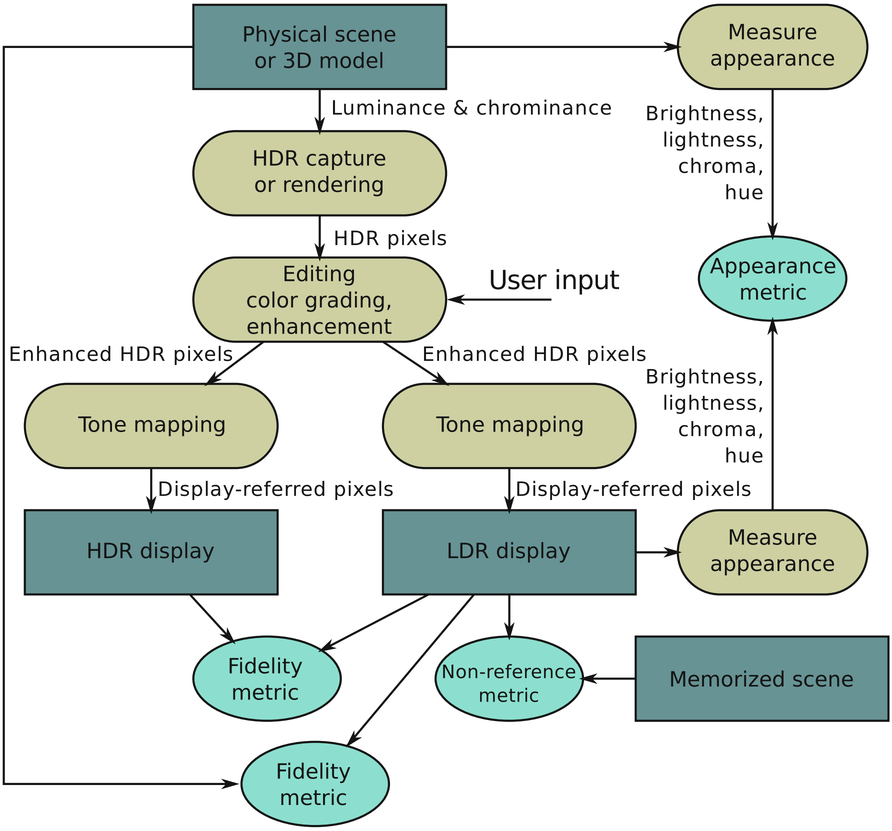


Fig. 9.13 Tone-mapping process and different methods of performing tone-mapping evaluation. Note that content editing has been distinguished from tone-mapping. The evaluation methods (subjective metrics) are shown as ovals

Fidelity with reality method, where a tone-mapped image is compared with a physical scene. Such a study is challenging to execute, in particular for video because it involves displaying both a tone-mapped image and the corresponding physical scene in the same experimental setup. Furthermore, the task is very difficult for observers as displayed scenes differ from real scenes not only in the dynamic range, but they also lack stereo depth, focal cues, and have restricted field of view and color gamut. These factors usually cannot be controlled or eliminated. Moreover, this task does not capture the actual intent when the content needs enhancement. Despite the above issues, the method directly tests one of the main objectives of tone-mapping and was used in a number of studies [4, 91, 92, 106, 107].

Fidelity with HDR reproduction methods, where content is matched against a reference shown on an HDR display. Although HDR displays offer a potentially large dynamic range, some form of tone-mapping, such as absolute luminance adjustment and clipping, is still required to reproduce the original content. This introduces imperfections in the displayed reference content. For example, an HDR display will not evoke the same sensation of glare in the eye as the actual scene. However, the approach has the advantage that the experiments can be run in a well-controlled environment and, given the reference, the task is easier. Because of the limited availability of HDR displays, only a few studies employed this method: [38, 46].

Non-reference methods, where observers are asked to evaluate operators without being shown any reference. In many applications there is no need for fidelity with "perfect" or "reference" reproduction. For example, the consumer photography is focused on making images look possibly good on a device or print alone as most consumers will rarely judge the images while comparing with real scenes. Although the method is simple and targets many applications, it carries the risk of running a "beauty contest" [59], where the criteria of evaluation are very subjective. In the non-reference scenario, it is commonly assumed that tone-mapping is also responsible for performing color editing and enhancement. But, since people differ a lot in their preference for enhancement [107], such studies lead to very inconsistent results. The best results are achieved if the algorithm is tweaked independently for each scene, or essentially if a color artist is involved. In this way we are not testing an automatic algorithm though, but a color editing tool and the skills of the artist. However, if these issues are well controlled, the method provides a convenient way to test TMO performance against user expectations and, therefore, it was employed in most of the studies on tone-mapping: [1, 4, 21, 39, 68, 91, 107].

Appearance match methods compare color appearance in both the original scene and its reproduction [59]. For example, the brightness of square patches can be measured in a physical scene and on a display using the magnitude estimation methods. Then, the best tone-mapping is the one that provides the best match between the measured perceptual attributes. Even though this seems to be a very precise method, it poses a number of problems. Firstly, measuring appearance for complex scenes is challenging. While measuring brightness for uniform patches is a tractable task, there is no easy method to measure the appearance of gloss, gradients, textures, and complex materials. Secondly, the match of sparsely measured perceptual attributes does not need to guarantee the overall match of image appearance.

None of the discussed evaluation methods is free of problems. The choice of a method depends on the application that is relevant to the study. The diversity of the methods shows the challenge of subjective quality assessment in tone-mapping, and is one of the factors that contribute to volatility of the results.

9.4.2.4 Volatility of the Results

It is not uncommon to find quality studies in graphics, which arrive with contradicting or inconclusive results. For example, two studies [8, 57] compared inverse tone-mapping operators. Both studies asked to rate or rank the fidelity of the processed image with the reference shown on an HDR display. The first study [8] demonstrated that the performance of complex operators is superior to that of a simple linear scaling. The second study [57] arrived with the opposite conclusion that the linear contrast scaling performs comparably or better than the complex operators. Both studies compared the same operators, but images, parameter settings for each algorithm, evaluation methods and experimental conditions were different. This two conflicting results show the volatility of many subjective experiments performed on images. The statistical testing employed in these studies can ensure that the results are likely to be the same if the experiment is repeated for a different group of observers, but with exactly the same images and in exactly the same conditions. The statistical testing, however, cannot generalize the results to the entire population of possible images, parameters, experimental conditions, and evaluation procedures.

9.4.3 Subjective Quality Experiments

This subsection presents the subjective tests conducted by the scientific community related to quality assessment of graphics data. The first and second parts detail, respectively, experiments related to image and 3D model artifact evaluation.

9.4.3.1 Image and Video Quality Assessment

Evaluating computer graphics methods is inherently difficult, as the results can often be only evaluated visually. This poses a challenge for the authors of new algorithms, who are expected to compare their results with the state of the art. For that reason, many recent papers in graphics include a short section with experimental validation. Such a trend shows that subjective quality assessment becomes a standard practice and a part of the research methodology in graphics. The need to validate methods also motivates comparative studies, in which several state-of-the-art algorithms are evaluated in a subjective experiment. Studies like this have been performed for image aspect ratio retargeting [75], image deghosting [29], or inverse tone-mapping [8, 57]. However, probably the most attention has attracted the problem of tone-mapping, which is discussed below.

Currently (as of 2014) Google Scholar search reports over 7,000 papers with the phrase "tone-mapping" in the title. Given this enormous choice of different algorithms, which accomplish a very similar task, one would wish to know which algorithm performs the best in a general case. In Sect. 9.2.5 we discussed a few objective metrics for tone-mapping. However, because their accuracy still needs to be validated, they are not commonly recognized method for comparing tone-mapping operators. Instead, the operators have been compared in a large number of subjective studies evaluating both tone-mapping for static images [1, 2, 4, 21, 22, 36, 38, 39, 46, 91, 92, 106, 107] and tone-mapping for video [10, 24, 68]. None of these studies provided a definite ranking of the operators since such a ranking strongly depends on the scene content and the parameters passed to a tone-mapping operator. Interestingly, many complex tone-mapping methods seem to perform comparable or worse than even a simple method, provided that it is fine-tuned manually [1, 38, 91]. This shows the importance of per-image parameter tuning. Furthermore, the objective (intent) of tone-mapping can be very different between operators. Some operators simulate the performance of the visual system with all its limitation; other operators minimize color differences between the HDR image and its reproduction; and some produce the most pleasing images [24, 59]. Therefore, a single ranking and evaluation criteria do not seem to be appropriate for evaluation of all types of tone-mapping. The studies have identified the factors that affect overall quality of the results, such as naturalness and detail [22], overall contrast and brightness reproduction [106, 107], color reproduction and visible artifacts [91]. In case of video tone-mapping, the overall quality is also affected by flickering, ghosting, noise, and consistency of colors across a video sequence [10, 24]. Evaluating all these attributes provides the most insight into the performance of the operators but it also requires the most effort and expertise and, therefore, is often performed by expert observers [24]. Overall, the subjective studies have not identified a single operator that would perform well a general case. But they helped to identify common problems in tone-mapping, which will help in guiding further research on this topic.

9.4.3.2 3D Model Quality Assessment

Several authors have made subjective tests involving 3D static or dynamic models [17, 18, 41, 45, 67, 74, 76, 79, 80, 87, 88, 100]. Their experiments, detailed below, had different purposes and used different methodologies. Bulbul et al. [13] recently provided a good overview and comparison of their environments, methodologies, and materials.

Subjective tests from Watson et al. [100] and Rogowitz and Rushmeier [74] focus on a mesh simplification scenario; their test databases were created by applying different simplification algorithms at different ratios on several 3D models. They considered a double stimulus rating scenario, i.e. observers had to rate the fidelity of simplified models regarding the original ones. The purposes of their experiments were, respectively, to compare image-based metrics and geometric ones to predict the perceived degradation of simplified 3D models [100] and to study if 2D images of a 3D model are really suited to evaluate its quality [74].

Rushmeier et al. [76] and Pan et al. [67] also considered a simplification scenario; however, their 3D models were textured. These experiments provided useful insights on how resolution of texture and resolution of mesh influence the visual appearance of the object. Pan et al. [67] also provided a perceptual metric predicting this visual quality and evaluated it quantitatively by studying the correlation with subjective MOS from their experiment.

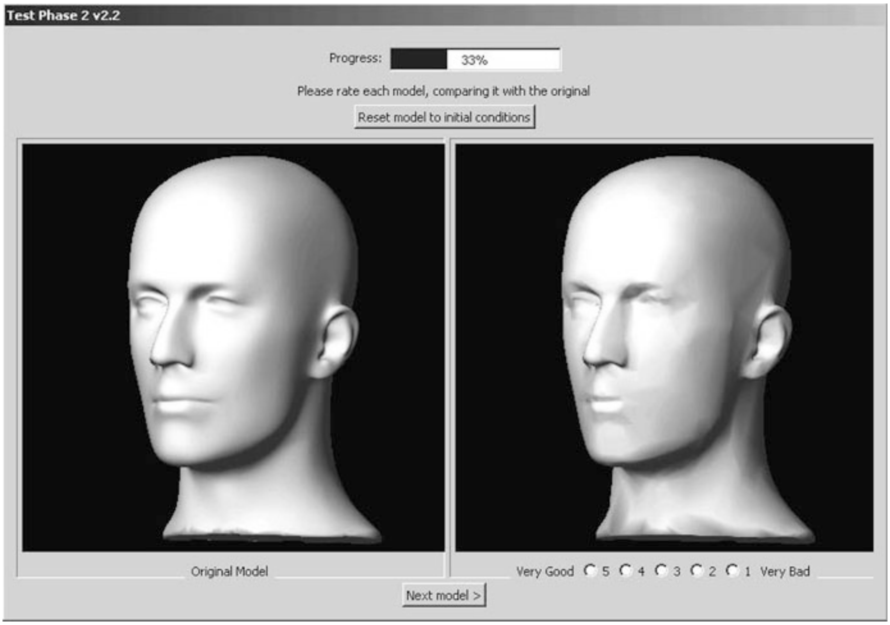
Lavoué [41] conducted an experiment involving 3D objects specifically chosen because they contain significantly smooth and rough areas. The author applied noise addition with different strengths either on smooth or rough areas. The specific objective of this study was to evaluate the *visual masking* effect. It turns out that the noise is indeed far less visible on rough regions. Hence, the metrics should follow this perceptual mechanism. The data resulting from this experiment (Masking Database in Table 9.1) are publicly available.[2]

To the best of our knowledge, the only experiment involving dynamic meshes was the one performed by Váša and Skala [87] in their work proposing the STED metric. They considered five dynamic meshes (chicken, dance, cloth, mocap, and jump) and applied different kinds of both spatial and temporal distortion of varying types: random noise, smooth sinusoidal dislocation of vertices, temporal shaking, and results of various compression algorithms. All the versions (including the original) were displayed at the same time to the observers, and they were asked to rate them using a continuous scale from 0 to 10.

In all the studies presented above, the observers are asked to rate the fidelity of a distorted model regarding a reference one, displayed at the same time (usually a double stimulus scenario). However some experiments consider a *single stimulus absolute rating* scenario. Corsini et al. [18] proposed two subjective experiments focusing on a watermarking scenario; the material was composed of 3D models processed by different watermarking algorithms introducing different kinds of artifacts. On the contrary to the studies presented above, they consider an absolute rating with hidden reference (i.e., the reference is displayed among the other stimuli). The authors then used the mean-opinion-scores to evaluate the effectiveness of several geometric metrics and proposed a new perceptual one (see Sect. 9.3.1) to assess the quality of watermarked 3D models. Lavoué et al. [45] follow the same protocol for their study; their material is composed of 88 models generated from 4 reference objects (Armadillo, Dyno, Venus and RockerArm). Two types of distortion (noise addition and smoothing) are applied with different strengths and nonuniformly on the object surface. The resulting MOS were originally used to evaluate the performance of the MSDM perceptual metric (see Sect. 9.3.1). The corresponding database (General-Purpose Database in Table 9.1) and MOS data are publicly available (see Footnote 2).

Rating experiments have the benefit of directly providing a mean-opinion-score for each object from the corpus, however the task of assigning a quality score to each stimulus is difficult for the observers and may lead to inaccurate results. That is why many experiments now rely on the simpler task of *Paired Comparison* where observers just have to provide a preference between a pair of stimuli (usually as a binary forced choice). Silva et al. [79] proposed an experiment involving both rating and preference tasks. Their corpus contains 30 models generated from 5 reference objects. The reference models have been simplified using three different methods and two levels. For the rating task, observers were asked to provide a score from 1 (very bad) to 5 (very good). Along with this rating, in another phase of the test, the observers were asked about their preference among several simplified models presented together. Figure 9.14 illustrates the evaluation interface for the rating task, the stimulus to rate is presented with its reference stimulus. The data resulting

[2] http://liris.cnrs.fr/guillaume.lavoue/data/datasets.html.


Test Phase 2 v2.2
Progress:
33%
Reset model to initial conditions
Original Model
Very Good ○ 5 ○ 4 ○ 3 ○ 2 ○ 1 Very Bad
Next model >

Please rate each model, comparing it with the original

Fig. 9.14 Evaluation interface for the subjective test of Silva et al. [80]. The observers were asked to compare the target stimulus (*right*) with the referential stimuli (*left*) and assign it a category rating from 1 (very bad) to 5 (very good). Reprinted from [80]

from these subjective experiments are publicly available[3] (Simplification Database in Table 9.1). The same authors did another subjective experiment using a larger corpus of models [80] where they only collected preferences.

Váša and Rus [88] conducted a subjective study focusing on evaluating compression artifacts. Their corpus contains 65 models from 5 references. The applied distortions are uniform and Gaussian noise, sine signal, geometric quantization, affine transform, smoothing and results from three compression algorithms. The observer's task is a binary forced choice, in the presence of the reference; i.e. triplets of meshes were presented, with one mesh being designated as original, and two randomly chosen distorted versions. A scalar quality value for each object from the corpus is then derived from the user choices. The data (Compression Database in Table 9.1) are publicly available.[4]

[3] http://www.ieeta.pt/~sss/repository/.

[4] http://compression.kiv.zcu.cz/.

9.4.4 Performance of Quality Metrics

9.4.4.1 Image Quality Assessment for Rendering

VDP-like metrics are, which are dominant in graphics, often considered to be too sensitive to small, barely noticeable, and often negligible differences. For example, many computer graphics methods result in a bias, which makes the part of a rendered scene brighter or darker than the physically accurate reference. Since such a brightness change is local, smooth, and spatially consistent, most observers are unlikely to notice it unless they scrupulously compare the image with a reference. Yet, such a difference will be signalized as significant by most VDP-like metrics, which will correctly predict that the difference is in fact visible when scrutinized. As a result, the distortion maps produced by objective metrics often do not correspond well with subjective judgment about visible artifacts.

Cadík et al. [90] investigated this problem by comparing the performance of the state-of-the-art fidelity metrics in predicting rendering artifacts. The selected metrics were based on perceptual models (HDR-VDP-2), texture statistics (SSIM, MS-SSIM), color differences (sCIE-Lab), and simple arithmetic difference (MSE). The performance was compared against experimental data, which was collected by asking observers to label noticeable artifacts in images. Two examples of such manually labeled distortion maps are shown in Fig. 9.2.

The same group of observers completed the experiment for two different tasks. The first task involved marking artifacts without revealing the reference (artifact free) image. It relied on the observers being able to spot objectionable distortions. In the second task the reference image was shown next to the distorted and the observers were asked to find all visible differences. The results for both tasks were mostly consistent across observers resulting in similar distortion maps for each individual.

When subjective distortion maps were compared against the metric predictions, they revealed weaknesses of both simple (PSNR, sCIE-Lab [108]) and advanced (SSIM, MS-SSIM [97], HDR-VDP-2) quality metrics. The results for the two separate data sets (NORM [30] and LOCCG[90]) and two experimental conditions (with-reference and no-reference) are shown in Fig. 9.15. The results show that the metrics that performed the best for one data set (HDR-VDP and SSIM for NORM) ended up in the middle or the end of the ranking for the other data set (LOCCG). This is another example of the volatility of the comparison experiments, discussed in Sect. 9.4.2.4. Because of the large differences in metric performance between images, no metric could be said to be statistically significantly better (in terms of AUC) than any other metric in a general case. More helpful was the detailed analysis of the results for particular images, which revealed the issues that reduced the performance of the advanced metrics. One of those issues was excessive sensitivity to brightness and contrast changes, which are common in graphics due to the bias of rendering methods (refer to Fig. 9.16). The simple metrics failed to distinguish between imperceptible and well visible noise levels in complex scenes

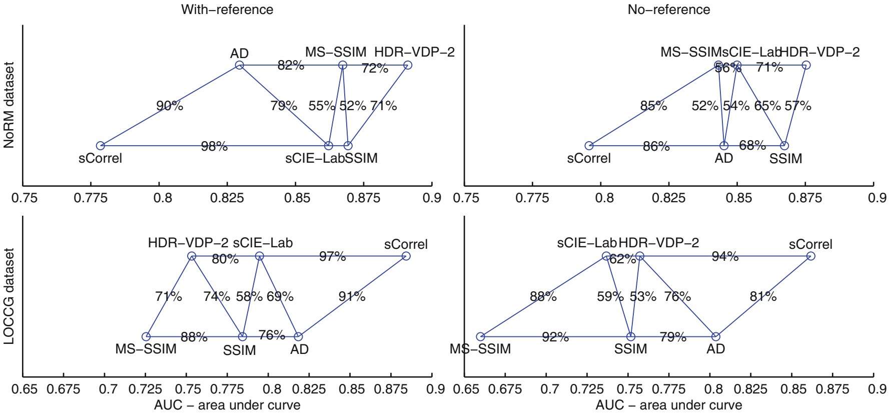


Fig. 9.15 The performance of quality metrics according to the area-under-curve (AUC) (the higher the AUC, the better the classification into distorted and undistorted regions). The *top row* shows the results for the NoRM data set [30] and *bottom row* the LOCCG data [90]. The columns correspond to the experiments in which the reference non-distorted image was shown (*left column*) or hidden (*right column*). The percentages indicate how frequently the metric on the right results in higher AUC when the image set is randomized using a bootstrapping procedure. The metrics: AD—absolute difference (equivalent to PSNR); SSIM—Structural Similarity Index; MS-SSIM—multi-scale SSIM; HDR-VDP-2—refer to Sect. 9.2.4; sCIE-Lab—spatial CIELab; sCorrel—per-block Spearman's nonparametric correlation

(refer to Fig. 9.17). The multi-scale metrics revealed problems in localizing small-area and high-contrast distortions (refer to Fig. 9.18). But the most challenging are the distortions that appeared as a plausible part of the scene, such as darkening in corners, which appeared as soft shadows (refer to Fig. 9.19).

Overall, the results revealed that the metrics are not as universal as they are believed to be. Complex metrics employing multi-scale decompositions can better predict visibility of low contrast distortions but they are less successful with super-threshold distortions. Simple metrics, such as PSNR, can localize distortions well, but they fail to account for masking effects.

9.4.4.2 3D Model Quality Assessment

For model-based metrics (i.e., relying on the geometry), recent studies [19, 44] have provided extensive quantitative comparisons of existing metrics by computing correlations with MOS from several databases. Studies generally consider two correlation coefficients: the SROC which measures the monotonic association between the MOS and the metric values and the Pearson linear correlation coefficient (LCC), which measures the prediction accuracy. The Pearson correlation is computed after performing a non-linear regression on the metric values as suggested by the video quality experts group (VQEG) [93], usually using a cumulative Gaussian function.

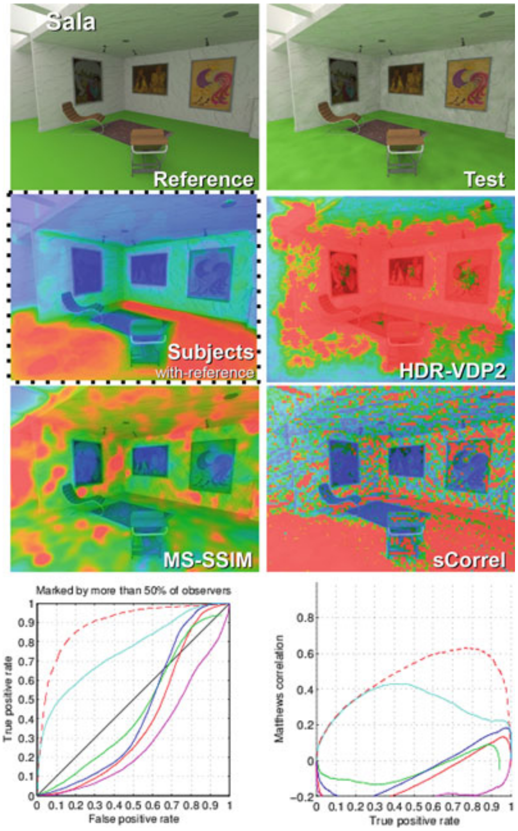


Fig. 9.16 Scene *sala* (*top*), distortion maps for selected metrics (second and third rows), ROC and correlation plots (*bottom*). Most metrics are sensitive to brightness changes, which often remain unnoticed by observers. *sCorrel* (block-wise Spearson correlation) is the only metric robust to these artifacts. Refer to the legend in Fig. 9.15 to check which lines correspond to which metrics in the plots

Table 9.1 summarizes these correlation results; best metrics are highlighted for each database. Note that many metrics cannot be applied to evaluating simplification distortions because they need the compared objects to share the same connectivity—[32, 43, 45, 83, 88]—or the same level of details—[18].

We can observe that classical geometric distances, like Hausdorff and RMS, provide a very poor correlation with human judgment, while most recent ones [42, 43, 85, 88, 96] provide much better performance. Unfortunately, image-based metrics have not been quantitatively tested on these public databases, hence a legitimate question remains: which is the best to predict 3D mesh visual fidelity, image-based or model-based metrics? Rogowitz and Rushmeier [74] argue for model-based metrics since they show that 2D judgments do not provide a good predictor of 3D object quality, implying that the quality of 3D objects cannot be correctly predicted by the quality of static 2D projections. To demonstrate that, the authors have conducted two subjective rating experiments; in the first one, the observers rated the quality of 2D static images of simplified 3D objects, while in the second one they rated an animated sequence of these images, showing a rotation of the 3D objects. Results show that (1) the lighting conditions strongly influence

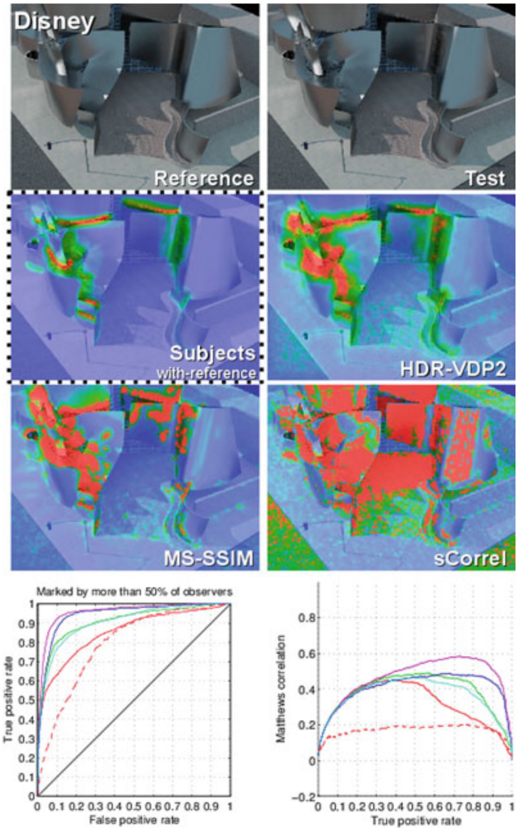


Fig. 9.17 Scene *disney*: simple metrics, such as sCorrel and AD, fail to distinguish between visible and invisible amount of noise resulting in worse performance

the perceived quality and (2) the observers perceive differently the quality of the 3D objects if they observe still images or animations. Watson et al. [100] also compared the performance of several image-based (Bolin-Meyer [12] and Mean Squared Error) and model-based (mean, max, and RMS) metrics. They conducted several subjective experiments to study the visual fidelity of simplified 3D objects, including perceived quality rating. Their results showed a good performance of 2D metrics (Bolin-Meyer [12] and MSE) as well as the mean 3D geometric distance as predictor of the perceived quality. The main limitation of this study is that the authors only consider one single view of the 3D models. More recently, Cleju and Saupe [17] designed another subjective experiment for evaluating the perceived visual quality of simplified 3D models and found that generally image-based metrics perform better than model-based metrics. In particular, they found that 2D mean squared error and SSIM provide good results, whereas SSIM's performance being more sensitive to the 3D model type. For model-based metric, like Watson et al. [100], they showed that the mean geometric distance performs better than RMS which is better than Hausdorff (i.e., maximum distance). The main limitation of these studies (mostly from 10 years ago) is that they consider one single type

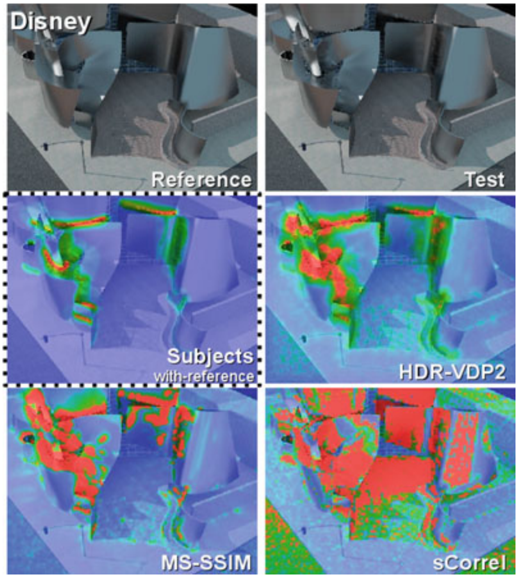


Fig. 9.18 *Dragons* scene contains artifacts on the dragon figures but not in the *black* background. Multi-scale IQMs, such as MS-SSIM and HDR-VDP-2, mark much larger regions due to the differences detected at lower spatial frequencies. Pixel-based AD (absolute differences) can better localize distortions in this case

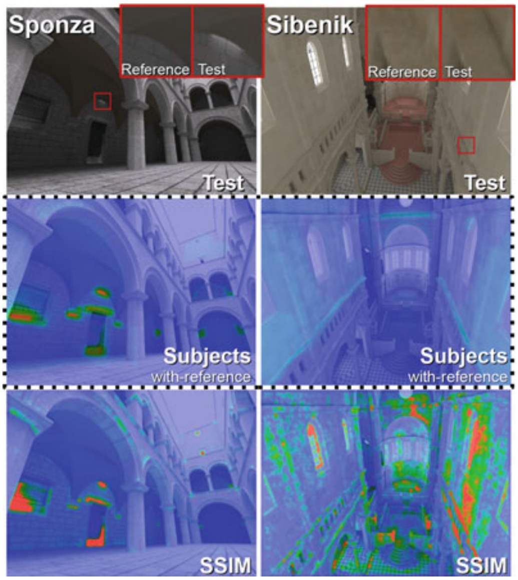


Fig. 9.19 Photon leaking and VPL clamping artifacts in scenes *sponza* and *sibenik* result in either brightening or darkening of corners. *Darkening* is subjectively acceptable, whereas *brightening* leads to objectionable artifacts

of distortion (only simplification) and very simple image-based and model-based metrics.

For dynamic meshes, a study presented by Váša and Skala [87] demonstrates an excellent prediction performance of the STED metric, while others (e.g., the KG error [33]) provide very poor results. Another open question concerns the quantitative evaluation of quality metrics for colored or textured meshes; indeed per-vertex colors or texture play a very important role in the appearance of a 3D model, however very few metrics still exist and no comparison study is still available.

9.5 Emerging Trends

9.5.1 Machine Learning

The objective of a quality assessment metric is to predict the visual quality of a signal, hence it basically needs to mimic the psychophysical process of the HSV, or at least relies on some features related to perceptual mechanisms. However modeling these complex principles and/or choosing appropriate characteristics may be hard. Hence it may appear convenient to treat the HVS as a black box which we wish to learn the input–output relationship. Such learning approaches were proposed recently [30, 43, 89]; they compute a large number of features and train classifiers on subjective ground-truth data. Such kinds of metrics are usually very efficient, however their ability to generalize depends on the richness of the ground-truth data. A very interesting point is that crowd-sourcing is developing as an excellent way to gather quickly a huge set of human opinions, that can then feed a classifier. As stated in the introduction, the future of quality metrics could lie in a combination of machine learning techniques with accurate psychophysical models.

9.5.2 3D Animation

There still exist very few works about quality assessment for dynamic meshes (i.e., sequence of meshes) and articulated meshes (i.e., one single mesh + animated skeleton) while these types of data are present in a wide range of computer graphics applications. The perceived visual quality of such 3D animation depends not only on the geometry, texture, and other visual attributes but also, to a large extent, on the nature of the movement and its velocity. This temporal dimension carries a whole range of additional cognitive phenomena. The CSF, for instance, is completely modified in a dynamic setting [34]. This is easily understandable since a rapid movement will be able to hide a geometrical artifact which would have been visible in a static case. In the case of human or animal animations, the *realism* of the animation is also a critical factor in the perception from the user. All these factors should be taken into account to devise efficient quality metrics, many progresses still remain to be achieved in this field.

9.5.3 Material and Lighting

The need of photorealistic rendering of 3D content has led to embed complex material and lighting information together with the geometric information. For instance, the bi-directional reflectance distribution function (BRDF) describes how much light is reflected when light makes contact with a certain material. More complex nonuniform materials can be represented by more complex reflectance functions acquired through sophisticated photometric systems, including surface light field (SLF) which represents the color of a point depending on the viewing direction (hence assuming a fixed lighting direction), BTF that extends the SLF for any incident lighting direction, and finally bidirectional subsurface scattering reflectance distribution function (BSSRDF) which is basically a BTF plus a model of the surface scattering. There still exist no metric to assess the quality of these complex attributes (mapped or not onto the surface). In particular, it could be very useful to integrate them into existing model-based metrics (e.g., MSDM2) which are currently too much independent of the rendering conditions.

9.5.4 Toward Merging Image and Model Artifacts

We have seen all along this chapter that visual defects may appear at several stages of a computer graphics work-flow (as illustrated in Fig. 9.1) and may concern different types of data: either the 3D models, or the final rendered or tone-mapped images. We have seen that there exist specific metrics dedicated to the detection of these model or image artifacts. Their use depends on the application, e.g. a 3D mesh compression approach has to be driven by a metric operating on the geometry, while a global illumination algorithm will be tuned using an image quality metric. What has been ignored until now is that these visual defects introduced either onto the geometry or onto the final images do have a visual interplay. For instance, the nature of the rendering algorithm obviously influences the perceptibility of a geometric artifact; similarly, some types of rendering artifact could be avoided by a proper modelling or a specific geometry processing algorithm. Hence it appears obvious that these two types of quality assessment (i.e., respectively, applied on models and images) should be connected. Integrating lighting and material information into model-based metrics (like mentioned in the above paragraph) could be a way to take into account these both processes (modeling and rendering). Considering the 3D scene for detecting image-based artifacts could be another way to model efficiently this interplay.

References

1. Akyüz, A.O., Fleming, R., Riecke, B.E., Reinhard, E., Bulthoff, H.H.: Do HDR displays support LDR content? A psychophysical evaluation. ACM Transactions on Graphics **26**(3), article no. 38 (2007)
2. Akyüz, A.O., Reinhard, E.: Perceptual evaluation of tone reproduction operators using the Cornsweet-Craik-O'Brien illusion. ACM Transactions on Applied Perception **4**(4), 1–29 (2008)
3. Allan, R., Terry, M.E.: Rank Analysis of Incomplete Block Designs : I . The Method of Paired Comparisons. Biometrica **39**(3/4), 324–345 (1952)
4. Ashikhmin, M., Goyal, J.: A reality check for tone mapping operators. ACM Transactions on Applied Perception **3**(4), 399–411 (2006)
5. Aydın, T.O., Mantiuk, R., Myszkowski, K., Seidel, H.P.: Dynamic range independent image quality assessment. ACM Transactions on Graphics (Proc. of SIGGRAPH) **27**(3), 69 (2008)
6. Aydın, T.O., Mantiuk, R., Seidel, H.P.: Extending quality metrics to full luminance range images. In: Proceedings of SPIE, pp. 68,060B–10. Spie (2008). DOI 10.1117/12.765095
7. Aydın, T.O., Čadík, M., Myszkowski, K., Seidel, H.P.: Video quality assessment for computer graphics applications. ACM Transactions on Graphics **29**(6), 1 (2010). DOI 10.1145/1882261.1866187
8. Banterle, F., Ledda, P., Debattista, K., Bloj, M., Artusi, A., Chalmers, A.: A Psychophysical Evaluation of Inverse Tone Mapping Techniques. Computer Graphics Forum **28**(1), 13–25 (2009). DOI 10.1111/j.1467-8659.2008.01176.x
9. Blackwell, H.: Contrast thresholds of the human eye. Journal of the Optical Society of America **36**(11), 624–632 (1946)
10. Boitard, R., Cozot, R., Thoreau, D., Bouatouch, K.: Temporal coherency in video tone mapping, a survey. In: HDRi2013 - First International Conference and SME Workshop on HDR imaging, Xx, p. no. 1 (2013)
11. Bolin, M.R., Meyer, G.W.: A frequency based ray tracer. In: Proc. of SIGGRAPH '95, pp. 409–418 (1995)
12. Bolin, M.R., Meyer, G.W.: A perceptually based adaptive sampling algorithm. In: Proceedings of the 25th annual conference on Computer graphics and interactive techniques - SIGGRAPH '98, pp. 299–309. ACM Press, New York, New York, USA (1998). DOI 10.1145/280814.280924
13. Bulbul, A., Capin, T., Lavoue, G., Preda, M.: Measuring Visual Quality of 3D Polygonal Models. IEEE Signal Processing Magazine **28**(6), 80–90 (2011)
14. Cater, K., Chalmers, A., Ward, G.: Detail to Attention: Exploiting Visual Tasks for Selective Rendering. Proc. of Eurographics workshop on Rendering pp. 270–280 (2003)
15. Cho, J., Prost, R., Jung, H.: An oblivious watermarking for 3-D polygonal meshes using distribution of vertex norms. IEEE Transactions on Signal Processing **55**(1), 142–155 (2007)
16. Cignoni, P., Rocchini, C., Scopigno, R.: Metro: Measuring Error on Simplified Surfaces. Computer Graphics Forum **17**(2), 167–174 (1998). DOI 10.1111/1467-8659.00236
17. Cleju, I., Saupe, D.: Evaluation of supra-threshold perceptual metrics for 3D models. In: Symposium on Applied Perception in Graphics and Visualization. ACM Press (2006). DOI 10.1145/1140491.1140499
18. Corsini, M., Gelasca, E.D., Ebrahimi, T., Barni, M.: Watermarked 3-D Mesh Quality Assessment. IEEE Transactions on Multimedia **9**(2), 247–256 (2007)
19. Corsini, M., Larabi, M.C., Lavoué, G., Petřík, O., Váša, L., Wang, K.: Perceptual Metrics for Static and Dynamic Triangle Meshes. Computer Graphics Forum **32**(1), 101–125 (2013). DOI 10.1111/cgf.12001
20. Daly, S.: The Visible Differences Predictor: An Algorithm for the Assessment of Image Fidelity. In: A.B. Watson (ed.) Digital Images and Human Vision, pp. 179–206. MIT Press (1993)

21. Delahunt, P.B., Zhang, X., Brainard, D.H.: Perceptual image quality: Effects of tone characteristics. Journal of Electronic Imaging **14**(2), 1–12 (2005). DOI 10.1117/1.1900134
22. Drago, F., L.Martens, W., Myszkowski, K., Sidel, H.P.: Perceptual Evaluation of Tone Mapping Operators with Regard to Similarity and Preference. Tech. rep., MPI Informatik (2002)
23. Dumont, R., Pellacini, F., Ferwerda, J.A.: Perceptually-driven decision theory for interactive realistic rendering. ACM Transactions on Graphics **22**(2), 152–181 (2003). DOI 10.1145/636886.636888
24. Eilertsen, G., Wanat, R., Mantiuk, R.K., Unger, J.: Evaluation of Tone Mapping Operators for HDR-Video. Computer Graphics Forum (Proc. of Pacific Graphics) **32**(7), 275–284 (2013)
25. Engeldrum, P.: Psychometric scaling: a toolkit for imaging systems development. Imcotek Press (2000)
26. Ferwerda, J.A., Shirley, P., Pattanaik, S.N., Greenberg, D.P.: A model of visual masking for computer graphics. In: Proc. of SIGGRAPH '97, pp. 143–152. ACM Press, New York, New York, USA (1997). DOI 10.1145/258734.258818
27. Georgeson, M.A., Sullivan, G.D.: Contrast constancy: deblurring in human vision by spatial frequency channels. J. Physiol. **252**(3), 627–656 (1975)
28. Guthe, M., Müller, G., Schneider, M., Klein, R.: BTF-CIELab: A Perceptual Difference Measure for Quality Assessment and Compression of BTFs. Computer Graphics Forum **28**(1), 101–113 (2009). DOI 10.1111/j.1467-8659.2008.01299.x
29. Hadziabdic, K.K., Telalovic, J.H., Mantiuk, R.: Comparison of deghosting algorithms for multi-exposure high dynamic range imaging. In: Proc. of Spring Conference on Computer Graphics, pp. 1–8 (2013)
30. Herzog, R., Čadík, M., Aydın, T.O., Kim, K.I., Myszkowski, K., Seidel, H.P.: NoRM: No-Reference Image Quality Metric for Realistic Image Synthesis. Computer Graphics Forum **31**(2pt3), 545–554 (2012). DOI 10.1111/j.1467-8659.2012.03055.x
31. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence **20**(11), 1254–1259 (1998). DOI 10.1109/34.730558
32. Karni, Z., Gotsman, C.: Spectral compression of mesh geometry. In: ACM Siggraph, pp. 279–286 (2000)
33. Karni, Z., Gotsman, C.: Compression of Soft-Body animation sequences. Computers & Graphics **28**(1), 25–34 (2004)
34. Kelly, D.H.: Motion and vision. {II}. Stabilized spatiotemporal threshold surface. Journal of Optical Society of America **69**(10), 1340–1349 (1979)
35. Kim, K.J., Mantiuk, R., Lee, K.H.: Measurements of achromatic and chromatic contrast sensitivity functions for an extended range of adaptation luminance. In: B.E. Rogowitz, T.N. Pappas, H. de Ridder (eds.) Human Vision and Electronic Imaging, p. 86511A (2013). DOI 10.1117/12.2002178
36. Korshunov, P., Ebrahimi, T.: Influence of Context and Content on Tone-mapping Operators. In: HDRi2013 - First International Conference and SME Workshop on HDR imaging, p. no. 2 (2013)
37. Krawczyk, G., Myszkowski, K., Seidel, H.P.: Contrast Restoration by Adaptive Countershading. Computer Graphics Forum **26**(3), 581–590 (2007). DOI 10.1111/j.1467-8659.2007.01081.x
38. Kuang, J., Heckaman, R., Fairchild, M.D.: Evaluation of HDR tone-mapping algorithms using a high-dynamic-range display to emulate real scenes. Journal of the Society for Information Display **18**(7), 461–468 (2010). DOI 10.1889/JSID18.7.461
39. Kuang, J., Yamaguchi, H., Johnson, G.M., Fairchild, M.D.: Testing HDR image rendering algorithms. In: Proc. IS&T/SID 12th Color Imaging Conference, pp. 315–320. Scotsdale, Arizona (2004)
40. Larkin, M., O'Sullivan, C.: Perception of Simplification Artifacts for Animated Characters. In: symposium on Applied perception in graphics and visualization, pp. 93–100 (2011)

41. Lavoué, G.: A local roughness measure for 3D meshes and its application to visual masking. ACM Transactions on Applied Perception (TAP) **5**(4) (2009)
42. Lavoué, G.: A Multiscale Metric for 3D Mesh Visual Quality Assessment. Computer Graphics Forum **30**(5), 1427–1437 (2011)
43. Lavoué, G., Cheng, I., Basu, A.: Perceptual Quality Metrics for 3D Meshes: Towards an Optimal Multi-Attribute Computational Model. In: IEEE International Conference on Systems, Man, and Cybernetics (SMC) (2013)
44. Lavoué, G., Corsini, M.: A comparison of perceptually-based metrics for objective evaluation of geometry processing. IEEE Transactions on Multimedia **12**(7), 636–649 (2010)
45. Lavoue, G., Drelie Gelasca, E., Dupont, F., Baskurt, A., Ebrahimi, T.: Perceptually driven 3D distance metrics with application to watermarking. In: SPIE, vol. 6312, pp. 63,120L–63,120L–12. SPIE (2006)
46. Ledda, P., Chalmers, A., Troscianko, T., Seetzen, H.: Evaluation of tone mapping operators using a high dynamic range display. ACM Transactions on Graphics **24**(3), 640–648 (2005)
47. Lindstrom, P.: Model Simplification using Image and Geometry-Based Metrics. Ph.D. thesis, Georgia Institute of Technology (2000)
48. Lindstrom, P., Turk, G.: Evaluation of memoryless simplification. IEEE Transactions on Visualization and Computer Graphics **5**(2), 98–115 (1999). DOI 10.1109/2945.773803
49. Lindstrom, P., Turk, G.: Image Driven Simplification. ACM Transactions on Graphics **19**(3), 204–241 (2000)
50. Liu, Y., Wang, J., Cho, S., Finkelstein, A., Rusinkiewicz, S.: A no-reference metric for evaluating the quality of motion deblurring. ACM Transactions on Graphics **32**(6), 1–12 (2013). DOI 10.1145/2508363.2508391
51. Lubin, J.: A visual discrimination model for imaging system design and evaluation. In: E. Peli (ed.) Vision Models for Target Detection and Recognition, pp. 245–283. World Scientific Publishing Company (1995)
52. Luebke, D., Hallen, B.: Perceptually driven simplification for interactive rendering. In: Rendering Techniques 2001: Proceedings of the Eurographics Workshop, p. 223 (2001)
53. Luebke, D., Hallen, B., Newfield, D., Watson, B.: Perceptually Driven Simplification Using Gaze-Directed Rendering. In: EGSR, pp. 223–234 (2001)
54. Mantiuk, R., Daly, S., Myszkowski, K., Seidel, H.: Predicting visible differences in high dynamic range images: model and its calibration. In: Human Vision and Electronic Imaging, pp. 204–214 (2005)
55. Mantiuk, R., Kim, K.J., Rempel, A.G., Heidrich, W.: HDR-VDP-2: a calibrated visual metric for visibility and quality predictions in all luminance conditions. ACM Trans. Graph (Proc. SIGGRAPH) **30**(4), 1 (2011). DOI 10.1145/2010324.1964935
56. Mantiuk, R.K., Tomaszewska, A., Mantiuk, R.: Comparison of four subjective methods for image quality assessment. Computer Graphics Forum **31**(8), 2478–2491 (2012)
57. Masia, B., Agustin, S., Fleming, R.W., Sorkine, O., Gutierrez, D.: Evaluation of reverse tone mapping through varying exposure conditions. ACM Transactions on Graphics **28**(5), 1 (2009). DOI 10.1145/1618452.1618506
58. McCann, J., Rizzi, A.: Veiling glare: the dynamic range limit of hdr images. In: Proc. of HVEI XII, vol. 6492, pp. 649,213–649,213. International Society for Optics and Photonics (2007)
59. McCann, J.J., Rizzi, A.: The Art and Science of HDR Imaging (Google eBook). John Wiley & Sons (2011)
60. Menzel, N., Guthe, M.: Towards Perceptual Simplification of Models with Arbitrary Materials. Computer Graphics Forum **29**(7), 2261–2270 (2010). DOI 10.1111/j.1467-8659.2010.01815.x
61. Mullen, K.T.: The contrast sensitivity of human colour vision to red-green and blue-yellow chromaic gratings. Journal of Physiolohy **359**, 381–400 (1985)
62. Myszkowski, K.: The visible differences predictor: Applications to global illumination problems. In: Rendering techniques' 98: proceedings of the Eurographics Workshop in Vienna, Austria, June 29-July 1, 1998, p. 223. Springer Verlag Wien (1998)

63. Myszkowski, K., Rokita, P., Tawara, T.: Perceptually-informed accelerated rendering of high quality walkthrough sequences. In: Eurographics Workshop on Rendering, vol. 99, pp. 5–18 (1999)
64. Myszkowski, K., Tawara, T., Akamine, H., Seidel, H.P.: Perception-guided global illumination solution for animation rendering. In: Proc. of SIGGRAPH'01, pp. 221–230. ACM, New York, NY, USA (2001). DOI 10.1145/383259.383284
65. O'Donovan, P., Agarwala, A., Hertzmann, A.: Color compatibility from large datasets. ACM Transactions on Graphics **30**(4), 1 (2011). DOI 10.1145/2010324.1964958
66. O'Sullivan, C., Howlett, S., Morvan, Y.: Perceptually adaptive graphics. Eurographics State of the Art Reports pp. 141–164 (2004)
67. Pan, Y., Cheng, I., Basu, A.: Quality metric for approximating subjective evaluation of 3-D objects. IEEE Transactions on Multimedia **7**(2), 269–279 (2005)
68. Petit, J., Mantiuk, R.K.: Assessment of video tone-mapping : Are cameras ' S-shaped tone-curves good enough? Journal of Visual Communication and Image Representation **24**, 1020–1030 (2013). DOI 10.1016/j.jvcir.2013.06.014
69. Ponomarenko, N., Lukin, V., Zelensky, A., Egiazarian, K., Carli, M., Battisti, F.: TID2008 - A database for evaluation of full-reference visual quality assessment metrics. Advances of Modern Radioelectronics **10**, 30–45 (2009)
70. Qu, L., Meyer, G.: Perceptually guided polygon reduction. IEEE Transactions on Visualization and Computer Graphics **14**(5), 1015–1029 (2008). DOI 10.1109/TVCG.2008.51
71. Ramanarayanan, G., Ferwerda, J., Walter, B.: Visual equivalence: towards a new standard for image fidelity. ACM Transactions on Graphics (TOG) **26**(3), 76 (2007). DOI 10.1145/1276377.1276472
72. Ramasubramanian, M., Pattanaik, S.N., Greenberg, D.P.: A perceptually based physical error metric for realistic image synthesis. In: Proceedings of the 26th annual conference on Computer graphics and interactive techniques - SIGGRAPH '99, pp. 73–82. ACM Press, New York, New York, USA (1999). DOI 10.1145/311535.311543
73. Reddy, M.: SCROOGE: Perceptually-Driven Polygon Reduction. Computer Graphics Forum **15**(4), 191–203 (1996)
74. Rogowitz, B.E., Holly E. Rushmeier: Are image quality metrics adequate to evaluate the quality of geometric objects? Proceedings of SPIE pp. 340–348 (2001)
75. Rubinstein, M., Gutierrez, D., Sorkine, O., Shamir, A.: A comparative study of image retargeting. ACM Transactions on Graphics **29**(6), 1 (2010). DOI 10.1145/1882261.1866186
76. Rushmeier, H., Rogowitz, B., Piatko, C.: Perceptual issues in substituting texture for geometry. In: SPIE, pp. 372–383. International Society for Optical Engineering; 1999 (2000)
77. Schroeder, W., Zarge, J., Lorensen, W.: Decimation of triangle meshes. In: ACM Siggraph, pp. 65–70 (1992)
78. Secord, A., Lu, J., Finkelstein, A., Singh, M., Nealen, A.: Perceptual models of viewpoint preference. ACM Transactions on Graphics **30**(5), 1–12 (2011). DOI 10.1145/2019627.2019628
79. Silva, S., Santos, B., Ferreira, C.: Comparison of methods for the simplification of mesh models using quality indices and an observer study. SPIE pp. 64,921L–64,921L–12 (2007)
80. Silva, S., Santos, B.S., Ferreira, C., Madeira, J.: A Perceptual Data Repository for Polygonal Meshes. 2009 Second International Conference in Visualisation pp. 207–212 (2009)
81. Silverstein, D., Farrell, J.: Efficient method for paired comparison. Journal of Electronic Imaging **10**, 394 (2001). DOI 10.1117/1.1344187
82. Smith, K., Krawczyk, G., Myszkowski, K.: Beyond tone mapping: Enhanced depiction of tone mapped HDR images. Computer Graphics Forum **25**(3), 427–438 (2006)
83. Sorkine, O., Cohen-Or, D., Toldeo, S.: High-pass quantization for mesh encoding. In: Eurographics Symposium on Geometry Processing, pp. 42–51 (2003)
84. Tian, D., AlRegib, G.: FQM: A Fast Quality Measure for Efficient Transmission of Textured 3D Models. In: ACM Multimedia, pp. 684–691 (2004)
85. Torkhani, F., Wang, K., Chassery, J.m.: A Curvature Tensor Distance for Mesh Visual Quality Assessment. In: International Conference on Computer Vision and Graphics (2012)

86. Trentacoste, M., Mantiuk, R., Heidrich, W., Dufrot, F.: Unsharp Masking, Countershading and Halos: Enhancements or Artifacts? Computer Graphics Forum **31**(2pt3), 555–564 (2012). DOI 10.1111/j.1467-8659.2012.03056.x
87. Vasa, L., Skala, V.: A Perception Correlated Comparison Method for Dynamic Meshes. IEEE Trans. on Visualization and Computer Graphics **17**(2), 220–230 (2011)
88. Váša, L., Rus, J.: Dihedral Angle Mesh Error: a fast perception correlated distortion measure for fixed connectivity triangle meshes. Computer Graphics Forum **31**(5) (2012)
89. Čadík, M., Herzog, R., Mantiuk, R., Mantiuk, R., Myszkowski, K., Seidel, H.P.: Learning to Predict Localized Distortions in Rendered Images. Computer Graphics Forum (Proc. of Pacific Graphics) **32**(7), 401–410 (2013)
90. Čadík, M., Herzog, R., Mantiuk, R.K., Myszkowski, K., Seidel, H.P., Čadík, M.: New Measurements Reveal Weaknesses of Image Quality Metrics in Evaluating Graphics Artifacts. ACM Trans. Graph (Proc. SIGGRAPH Asia) **31**(6), 147 (2012). DOI 10.1145/2366145.2366166
91. Čadík, M., Wimmer, M., Neumann, L., Artusi, A.: Evaluation of HDR tone mapping methods using essential perceptual attributes. Computers & Graphics **32**(3), 330–349 (2008). DOI 10.1016/j.cag.2008.04.003
92. Villa, C., Labayrade, R.: Psychovisual assessment of tone-mapping operators for global appearance and colour reproduction. In: Proc. of Colour in Graphics Imaging and Vision 2010, pp. 189–196. Joensuu, Finland (2010)
93. VQEG: Final report from the video quality experts group on the validation of objective models of video quality assessment. Tech. rep., Video Quality Experts Group (2000)
94. Walter, B., Pattanaik, S.N., Greenberg, D.P.: Using Perceptual Texture Masking for Efficient Image Synthesis. Computer Graphics Forum **21**(3), 393–399 (2002). DOI 10.1111/1467-8659.t01-1-00599
95. Wang, K., Lavoué, G., Denis, F., Baskurt, A.: Robust and blind mesh watermarking based on volume moments. Computers & Graphics **35**(1), 1–19 (2011)
96. Wang, K., Torkhani, F., Montanvert, A.: A Fast Roughness-Based Approach to the Assessment of 3D Mesh Visual Quality. Computers & Graphics (2012)
97. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004)
98. Watson, A., Ahumada Jr, A.: A standard model for foveal detection of spatial contrast. Journal of Vision **5**(9), 717–740 (2005)
99. Watson, A.B.: The cortex transform: Rapid computation of simulated neural images. Computer Vision, Graphics, and Image Processing **39**(3), 311–327 (1987). DOI 10.1016/S0734-189X(87)80184-6
100. Watson, B., Friedman, A., McGaffey, A.: Measuring and predicting visual fidelity. ACM Siggraph pp. 213–220 (2001)
101. Williams, N., Luebke, D., Cohen, J., Kelley, M., Schubert, B.: Perceptually Guided Simplification of Lit, Textured Meshes. In: ACM Symposium on Interactive 3D Graphics, pp. 113–121 (2003)
102. Wilson, H.R.: A transducer function for threshold and suprathreshold human vision. Biological Cybernetics **38**(3), 171–178 (1980). DOI 10.1007/BF00337406
103. Yee, H.: Perceptual Metric for Production Testing. Journal of Graphics Tools **9**(4), pages 33–40 (2004)
104. Yee, H., Pattanaik, S., Greenberg, D.P.: Spatiotemporal sensitivity and visual attention for efficient rendering of dynamic environments. ACM Transactions on Graphics **20**(1), 39–65 (2001). DOI 10.1145/383745.383748
105. Yeganeh, H., Wang, Z.: Objective quality assessment of tone-mapped images. IEEE Transactions on Image Processing **22**(2), 657–67 (2013). DOI 10.1109/TIP.2012.2221725
106. Yoshida, A., Blanz, V., Myszkowski, K., Seidel, H.P.: Perceptual evaluation of tone mapping operators with real world scenes. In: Proc. of SPIE Human Vision and Electronic Imaging X, vol. 5666, pp. 192–203. San Jose, CA (2005)

107. Yoshida, A., Mantiuk, R., Myszkowski, K., Seidel, H.P.: Analysis of reproducing real-world appearance on displays of varying dynamic range. Computer Graphics Forum **25**(3), 415–426 (2006)
108. Zhang, X., Wandell, B.A.: A spatial extension of CIELAB for digital color-image reproduction. Journal of the Society for Information Display **5**(1), 61 (1997). DOI 10.1889/1.1985127
109. Zhu, Q., Zhao, J., Du, Z., Zhang, Y.: Quantitative analysis of discrete 3D geometrical detail levels based on perceptual metric. Computers & Graphics **34**(1), 55–65 (2010). DOI 10.1016/j.cag.2009.10.004

Chapter 10
Conclusions and Perspectives

Chenwei Deng, Shuigen Wang, and Lin Ma

10.1 Summary

The main contribution of this book is offering an overview of current status, challenges, and new trends of visual quality assessment, from subjective assessment models to objective metrics, covering full-reference (FR), reduced-reference (RR), and no-reference (NR), multiply distorted images, contrast-changed images, mobile media, high dynamic range (HDR) images and videos, medical images, stereoscopic/3D videos, retargeted images and videos, computer graphics and animation quality assessment. Figure 10.1 diagrams the content presented in this book.

With the rapid development of digital technologies in the past decades, visual communications, broadcasting, entertainment, and recreation of video and photography have been completely transformed from analogue based devices, products, systems, and services to increasing diverse forms of digital counterparts, such as digital cameras, digital video cameras, and digital TV services. The development of visual quality assessment has also changed from quality of service (QoS) to quality of experience (QoE), to adapt to various applications. Numerous subjective and objective quality guaging approaches have been proposed, including entropy and rate-distortion based methods for pixel-based evaluation, e.g., mean-square-error (MSE) and peak signal-to-noise ratio (PSNR), feature-driven algorithms, e.g., structural similarity (SSIM) and its variants, natural scene statistics (NSS) based

C. Deng (✉) • S. Wang
School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China
e-mail: cwdeng@bit.edu.cn; sgwang@bit.edu.cn

L. Ma
Huawei Noah's Ark Lab, Hong Kong, China
e-mail: forest.linma@gmail.com

C. Deng et al. (eds.), *Visual Signal Quality Assessment*,
DOI 10.1007/978-3-319-10368-6_10

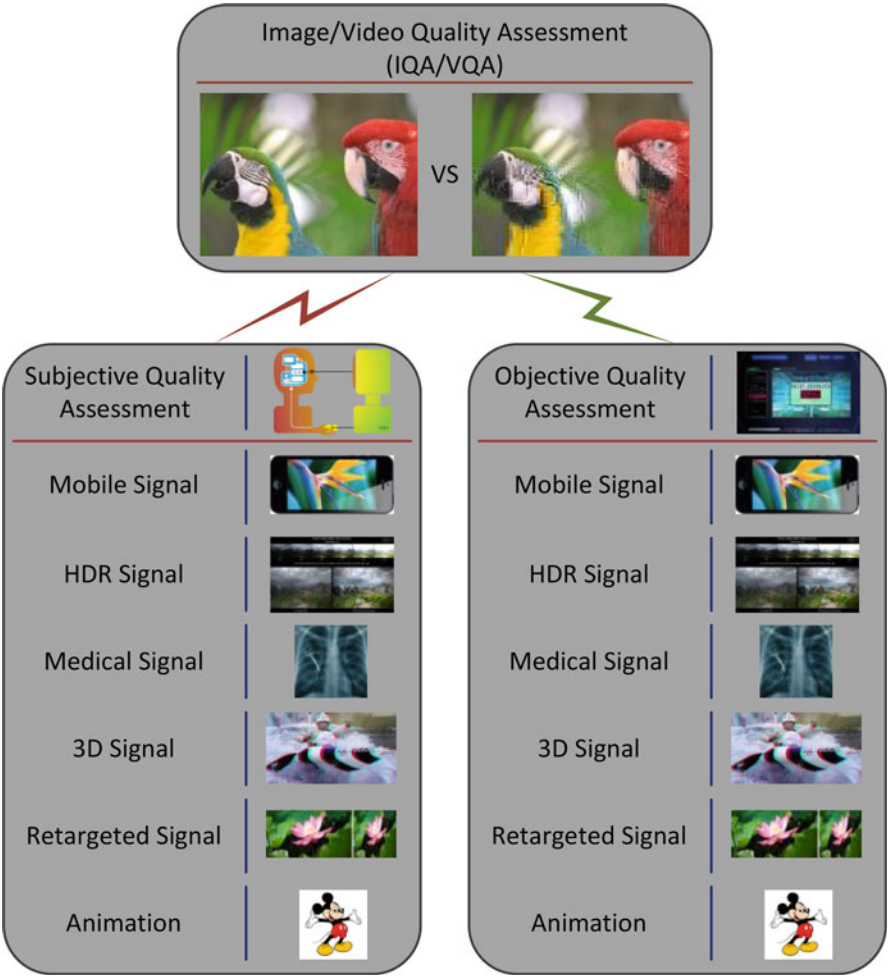


Fig. 10.1 The content of this book

models, just-noticeable-difference (JND) models, and multi-channel contrast gain control (CGC) metrics, etc. With these efforts, tremendous advances have been achieved for perceptual visual signal quality assessment.

Subjective assessment of QoE is defined as "the overall acceptability of an application or service, as perceived subjectively by an end-user," and it is the most reliable way for visual quality assessment. The direct results of subjective tests are the IQA databases. The popular IQA databases include Laboratory for Image & Video Engineering (LIVE) database, Tampere Image Database 2008 (TID2008), and Categorical Subjective Image Quality (CSIQ) database, etc. These three databases cover almost all (if not all) the distortion types, such as JPEG2000 compression distortion, JPEG compression distortion, White noise, Gaussian blur, and contrast change, etc. Recently, several new image databases have been built for publicly use. The TID2008 database was extended to the TID2013 database which contains total number of 3,000 images and is generated by corrupting 25 original images with 24 types of distortion (17 types for TID2008) at five different levels (four for TID2008). Apart from the aforementioned single distortion databases, multiple distortions image database, i.e., LIVE multiply distorted image database (LIVEMD), was built with each image corrupted by two distortion types. In order to address the problems in contrast change evaluation, a dedicated and more comprehensive contrast-changed image quality database (CID2013) was established for contrast change assessment with 15 natural images taken from Kodak database and 400 contrast-changed versions of mean luminance shift and contrast change. Furthermore, HDR imaging has attracted a lot of attention and enthusiasm in the last decade. A new and dedicated HDR image quality database (HDR2014) was also proposed for HDR IQA studies.

Recently, numerous relevant researches have been carried out varying from traditional impairment-centric methods to QoE. For a long period of time, the research of subjective assessment was targeted at determining user sensitivity to impairments induced in the media by suboptimal delivering, and the media recipient was considered as a passive observer, whose appreciation of the video material was determined primarily by the degree of annoyance due to the visual impairments. Visual impairments are often produced by limited spatial, temporal, and bit rate resolutions in displays, bandwidth and storage constraints, or error-prone transmission channels. As a result, multimedia material is often delivered along with impairments disrupting the overall appearance of the visual contents, and provoking a sense of dissatisfaction in the users [1–3]. Great efforts [4–12] have been devoted to the development of technologies that can either prevent the appearance of impairments, or repair it when needed.

One representative work, Engeldrum's image quality circle (IQC) framework [8], aims at providing an effective methodology for linking experienced visual quality to the settings of technological variables of a multimedia system. The IQC proposes a divide-and-conquer approach, involving three intermediate steps: (1) linking overall image quality to the combination of underlying perceived attributes of the image; (2) linking each image attribute to the physical characteristics of the system output; (3) linking the physical description of the system output to the system technological variables. It can adapt to different problem domains, including display quality assessment, signal processing algorithm optimization, and network parameter optimization.

As for displays, the RaPID method [13] and the IQC constituted a solid framework to improve display quality. With regard to the visual quality preference of processed signals, the signal processing community evolved in rather distinct aspects, and its researchers largely focused on a signal fidelity approach. The goal is to understand the impacts on perceived quality of a specific type (e.g., compression, scaling, denoising) of processing algorithm, by identifying the optimized parameters of the algorithm to produce the best visual quality. A large number of works utilizing low-level features of the human visual system (HVS), like [7, 10, 14, 15], have been proposed to figure out the relations between coding artifacts and quality preference. After that, subjective studies aimed at generating ground-truth data, and the modeling of contrast [16–18], luminance masking [19, 20], spatial pooling strategies [21], and image structure perception [22, 23] have been conducted. However, it was interestingly found that HVS-based models cannot predict and describe image quality accurately even at threshold level, since non-expert observers are less sensitive to compression-related artifacts while the well-informed experts are very sensitive to artifact visibility in the images. Therefore, some attempts were raised to explore the role of higher-level HVS features in signal impairment annoyance and quality appreciation targeted visual attention mechanisms [24–27].

As for the network-related impairments, e.g., bandwidth limitations, along with network unreliability (i.e., the possibility of packets loss), can cause frame freezes, deformations of the spatial and temporal structure of the content, and long stalling times. And researchers have been working towards correlating the QoS parameters (e.g., packet loss ratio, delay, jitter, and available bandwidth) to QoE measurements by using fitting functions [28–30]. Generally, QoS metrics work well in estimating QoE from a network efficiency point of view, but they do not accurately reflect the overall viewing experience. The impacts of signal impairments such as blockiness and blur are not taken into account in these approaches.

However, all the impairment-centric models mentioned above cannot meet the requirements and developments of multimedia applications. The media recipient becomes an active user instead of a passive one, who creates content, interacts with the system, and selects the media he/she wants to be delivered. Elements such as visual semantics, user personality, preferences and intent, social and environmental context of media fruition also play important roles in the final experience assessment, which is illustrated in Fig. 10.2. In order to adapt the traditional visual quality gauging metrics to QoE, a few models [7, 31–35] have been proposed throughout the last decade, and significant improvements have been achieved, though there still exists a long way to go. Keelan [7] defined visual

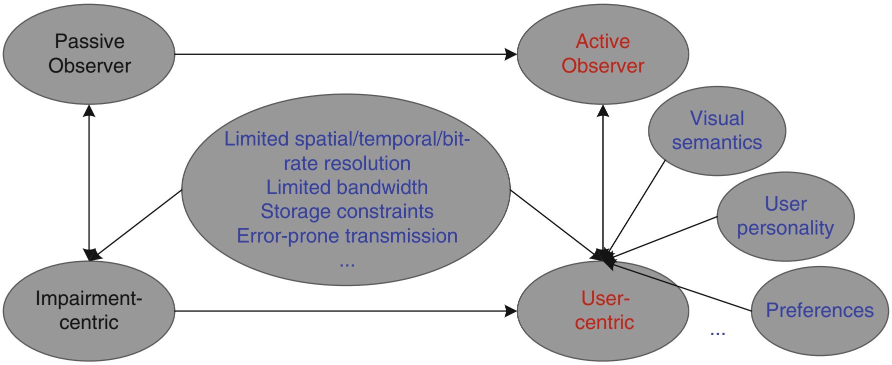


Fig. 10.2 The development of subjective quality assessment models from impairment-centric approaches to user-centric ones

quality as a multidimensional quantity and distinguished four different families of attributes: artifactual, preferential, aesthetic, and personal ones. Ghinea and Thomas attempted to present a more encompassing definition of visual quality, by proposing the concept of quality of perception (QoP) [31]. The FUN model of de Ridder and Endrikhovski [32] can also be considered as a milestone in the road that took visual quality to be evolved into QoE.

With respect to objective quality assessment, using the aforementioned publicly available IQA databases, numerous metrics have been developed for image quality evaluation in the past decades. Depending on the availability of the reference image, objective IQA can be categorized into three groups: (1) full-reference (FR); (2) reduced-reference (RR); and (3) no-reference (NR).

Besides the two classic metrics, i.e., MSE and PSNR, a large variety of FR-IQA approaches have been proposed and achieved remarkable performances. These FR-IQA metrics include scale transform-based ones (e.g., MS-SSIM), saliency-based ones (e.g., IW-SSIM), gradient magnitude-based ones (e.g., FSIM, GSIM), and others (e.g., VIF, IGM).

As for the RR-IQA, only partial information of the original image is available. A free energy based distortion metric (FEDM), RR entropic-difference indexes (RRED), Fourier transform based quality measure (FTQM), RR-SSIM, and structural degradation model (SDM) have been developed based on different theories.

However, in real-world applications, information of the original images is not always available. In these cases, NR-IQA metrics are required for measuring image quality. In the last decade, some NR methods were proposed for specific distortion types, such as CPDB for blur distortion, FISH for sharpness evaluation. In recent years, general-purpose NR-IQA has been intensively studied and can be categorized into two types: (1) extracting effective features followed by a regression process (e.g., NSS based DIIVINE, BLIINDS-II, BRISQUE); (2) operating without human ratings (e.g., NIQE, QAC).

Regarding the emerging fields in objective quality assessment, the free energy model based comparative IQA (C-IQA) approach was developed, which is inherently different from FR, RR, and NR methods. The C-IQA takes an image pair as input and predicts their relative quality without using any knowledge about the original image. For the multiply distorted IQA, a FIve-Step BLInd Metric (FISBLIM) was proposed using several common image processing blocks to simulate the image perceiving process of the human eyes. A novel reduced-reference image quality metric (RIQMC) was proposed for contrast change evaluation using entropies and order statistics of the image histograms.

Apart from the study of general-purpose IQA, mobile video quality evaluation is becoming an important research area these days, due to the technological advances of high speed wireless communication networks, and mobile devices being capable of producing and consuming high quality videos. Chapter 4 presented a review of recent researches of subjective and objective mobile videos quality assessment, for maximizing the QoE of the delivered video contents. Various factors affect the quality of mobile videos, including blockiness, blurring, ringing, packet losses, spatial, temporal and quality resolution changes, etc. To address these issues, many researches have been carried out on subjective and objective quality assessment for mobile videos. Since the mobile environments are different in various aspects, in terms of the types of used devices, the degree of concentration of users, lighting conditions, etc., mobile video based subjective VQA has to consider both these factors and the International standard ITU-R BT.500-13 from International Telecommunication Union Radiocommunication Sector. With the subjective tests for video scalability, it was found that there is a bit rate threshold at which the preference of scalability options is switched. Below the threshold, enhancing the frame quality has the priority with improvements in either the SNR or spatial dimension. Above the threshold, the frame quality reaches a certain satisfactory level, and thus, the frame rate becomes more critical for perceived quality. Note that the threshold mainly depends on the content characteristics.

As for the objective quality evaluation of mobile scenarios, the existing general objective metrics, e.g., PSNR, MSE, SSIM, VQM, and motion-based video integrity evaluation (MOVIE), cannot achieve high correlation between measured quality and subjective scores. Mobile quality gauging has to consider its unique properties, such as network and displays. Specially, due to the limitations of network capacity, scalability for mobile videos is useful to handle the network terminal capability issue. Several objective metrics for mobile video scalability have been developed from spatial, temporal, and quality dimensions, and experimental results demonstrated that their performances are even better than those of PSNR and SSIM. Nevertheless, it is also important to evaluate their relative performance via thorough benchmarking studies. And there still exists a huge space for mobile video quality assessment, such as 3D videos, HDR videos.

One newly developing multimedia technology is HDR display. HDR is opposite to low dynamic range (LDR) which suffers the drawbacks that real physical luminance cannot be captured in a natural scene. However, HDR is able to capture or reproduce higher contrast and luminance ranges, and represents the dynamic range of the visual stimuli presented in practical applications. The popularization of HDR largely improves the visual QoE of the end users.

However, on the other hand, HDR also faces some challenges in generation, storage, processing, and display, etc. One can obtain HDR content in three ways: (1) fusing multi-exposure LDR images/frames [36]; (2) using specialized cameras [37]; (3) using renderers from virtual environments. In general, each pixel of the generated HDR content requires 12 bytes for storage. For one 512×512 HDR image, it requires 3,145,728 bytes memory for storage, which is too expensive to sustain. Therefore, a number of compression methods [38–44] have been proposed for HDR content. The proposed method in [38] only needs 4 bytes per pixel by storing a shared exponent among three color channels. The compression approach in [39] is called as LogLUV encoding. The authors in [40] defined the format of HDR as a half-floating point format. An interesting and novel algorithm in [44] converts the HDR content into LDR by tone mapping operators (TMOs). The resultant LDR image is thus compressed through encoder and decoder. Finally, the decoded LDR image is re-converted into its HDR format. The HDR data is efficiently compressed in this way, even though it is difficult to convert an HDR image/video to LDR without losing perceivable visual information. With TMOs technique, HDR contents can be displayed using common LDR devices, such as CRT, LCD monitors, and printers. Since tone mapping reduces the dynamic range, it will inevitably lead to the loss of visual details and further affect the perceived appearance of the HDR content. It is therefore necessary to analyze how they affect the visual experience of the processed HDR content. A detailed discussion about the relationship between tone mapping and image quality has been presented from perceptual visual quality, visual attention, and naturalness aspects. It was found that tone mapping not only degrades visual quality by destroying scene details but also affects the natural appearance of the content. Nevertheless, how it affects naturalness remains as an open problem. Experimental results have demonstrated that apart from the increasing or decreasing of attention magnitude [45–51], the tone mapping changes attentional regions, since most of TMOs sacrifice visual information by reducing the dynamic range. As a consequence, a non-attentional (attentional) region in the HDR image becomes an attentional (non-attentional) one in the tone mapped version. In addition, there exist similar conclusions for video signals in terms of visual attention maps. Since HDR content may suffer from multiple distortions (tone mapping, compression artifacts, inverse tone mapping artifacts), the quality measurement for HDR is challenging and few researches have been conducted for both subjective and objective ones [52–54].

Medical IQA is another IQA research topic for real applications. The IQA for medical image is rather important and has great practical significance, such as improving the image quality in mammography, optimizing X-ray tube voltage and tube current, etc. In medical imaging, image quality is governed by a variety of factors such as contrast, resolution (sharpness), noise, artifacts, and distortion. A number of studies have been conducted to establish image quality standards and develop quality assessment methods, including conventional medical image quality metrics, e.g., PSNR, contrast-to-noise ratio (CNR), contrast improvement ratio (CIR), detective quantum efficiency (DQE), modulation transfer function (MTF) and noise power spectrum (NPS), and recently proposed metrics, like mutual information (MI) [55]. Each metric has its corresponding application for medical images. The MTF is widely recognized as the most relevant metric of resolution performance in radiographic imaging [56]. The NPS is one of the most common metric describing the noise properties of imaging systems. The DQE is a spatial frequency based measurement, to evaluate the ability of the imaging devices converting the spatial information contained in the incident X-ray fluence into useful image information [57–59]. In addition, medical IQA has been utilized to improve the quality of mammography image with wavelet-based approaches for image denoising and enhancing, and to evaluate the effects of radiation dose reduction on image quality in digital radiography. The issue of radiation dose exposure to patients from digital radiography is a major public health concern. In particular, it is important to keep radiation dose exposure to a minimum in female patients during their reproductive period, who frequently undergo repeated radiation exposure during the course of diagnostic imaging and treatment follow-up.

With the rapid growth in the quantity of stereoscopic/3D content created by cinema, television, and entertainment industry, visual quality assessment of stereoscopic/3D image and video has become an increasingly important and active field of research. However, due to the diversity of stereoscopic display technologies and the profundity of how human perceives and processes 3D information, understanding the QoE of stereoscopic image and video is a complex and multidisciplinary problem. Most existing objective stereoscopic quality assessment algorithms can be regarded as the extended versions of 2D QA algorithms, while few of them consider some aspects of depth perception and utilize either computed or measured depth/disparity information from the stereo pairs. Studies have shown that NSS can efficiently distinguish original images from images distorted by different distortions. And therefore, the properties of HVS are important for evaluating the QoE. Some works have been conducted based on NSS and HVS, for stereoscopic image and video quality assessment, which achieve better performance than those based on 2D algorithms. Furthermore, visual discomfort and fatigue when viewing stereoscopic images and videos should also be considered for better QoE evaluation.

To adjust the abovementioned visual signals into arbitrary sizes and resolutions displayed in various devices, signals are then needed to be transformed into related versions by some retargeting techniques, e.g., cropping (CROP), scaling (SCAL), seam carving (SEAM), Optimized seam carving and scale (SCSC), Non-homogeneous retargeting (WARP), and Scale and stretch (SCST), etc. Since the resolutions of retargeted images are changed, the objective shapes may be distorted, and some content information may be discarded. All these changes will have remarkable influences on the image quality. Most of the existing full-reference (FR) assessment methods require the same image sizes for both reference and distorted images, but this is not always available for retargeted images. In Chap. 8, the newly developed subjective quality assessment metrics for retargeted images were reviewed as well as the objective ones. Two public subjective databases, i.e., RetargetMe and CUHK retargeting database have been built for retargeted image quality assessment. To automatically and reliably evaluate the retargeted image quality, bidirectional warping (BDW), quality metric scale space matching (SSM), color layout (CL), earth mover's distance (EMD), bidirectional similarity (BDS), edge histogram (E-H), and SIFTflow have been proposed but achieved poor performances on both RetargetMe and CUHK retargeting databases. Almost all these methods are not applicable in some specific distortions or induce information loss. Moreover, it was found that in most cases, the human subjects tend to sacrifice the information loss rather than the shape distortion for recognizing a good quality image. If the information loss and shape distortion can be combined together, better performance would be achieved for retargeted image quality assessment.

Many aforementioned images and videos, such as HDR contents, and stereoscopic/3D visual signals, are generated by computer graphics techniques, as demonstrated in Fig. 10.3. Therefore, the quality assessment of computer graphics covers a broad area for the evaluation of artifacts visual impacts induced by various computer graphic techniques, e.g., geometry processing, rendering, tone mapping, and animation, etc. Chapter 9 presents a review of subjective and

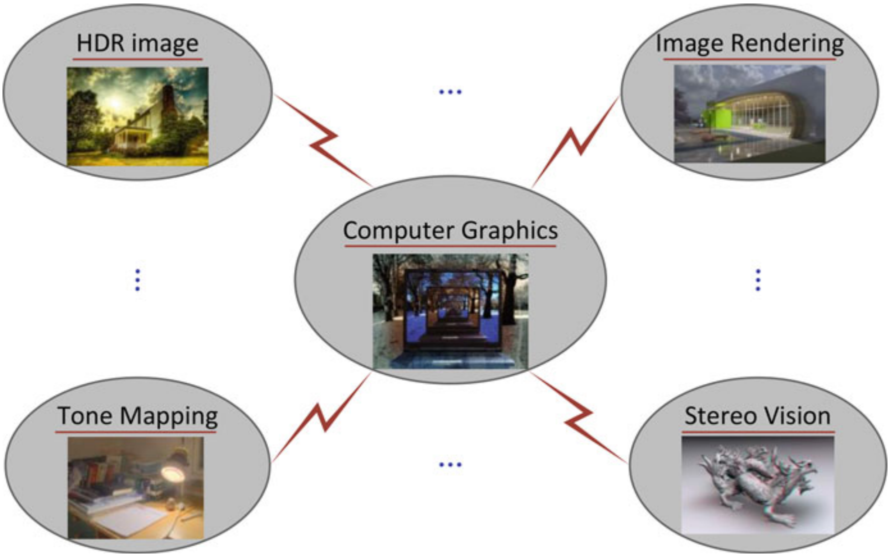


Fig. 10.3 The broad research area of computer graphics

objective evaluation methods for graphics and animations. The existing works can be classified into image-based (i.e., evaluating artifacts in 2D rendered images and videos) and model-based approaches (i.e., artifacts in the 3D models), since computer graphics involves two main types of data: 3D data, i.e. surface and volume meshes issued from geometric modeling or scanning processes, and 2D images and videos created/modified by graphical processes like rendering and tone mapping. In Chap. 9, different image quality metrics have been reviewed, including visual models based metrics, open source metrics, data-driven metrics, HDR metrics for rendering, and the metrics for aesthetics and naturalness evaluation. In terms of rendering, two important problems need to be addressed: (1) how to allocate samples, and (2) when to stop collecting samples. As for HDR content, luminance-dependent effects are modeled in the visual difference predictor for HDR images (i.e., HDR-VDP). Aydin et al. [60, 61] classified the distortions of tone mapping into the loss of visible contrast, amplification of invisible contrast and contrast reversal. Both color and binocular disparity are important factors that have strong influences on the depth perception of 3D scenes, and geometry and/or texture of the meshes are considered in model-based approaches for 3D models. Apart from methods mentioned above, the detailed tips of subjective assessment including scaling methods, parameter settings have also been discussed. Moreover, it also points out the corresponding quality affecting factors for different types of graphics contents.

In the above, we summarize high-level concepts, ideas, and algorithms of the recent development of visual QoE, and we believe that a good understanding of all the chapters of this book can help the readers make the right choice. In the following, an overall perspective is given for each research topics described in this book.

10.2 Perspectives

Despite the existing works of subjective QoE, unveiling a reliable model of user QoE preference is still beyond reach. The profound transformation that media consumption underwent in the last decade opens countless questions and applications, in which affecting factors and features of the viewing experience still have to be determined. Therefore, the existing knowledge developed in other fields, such as human computer interaction, affective computing, behavioral psychology, media production, computer graphics, and lighting design, should be incorporated into subjective QoE assessment.

New display technologies, such as HDR displays, stereoscopic and autostereoscopic displays, can provide a more immersive viewing experience by enhancing specific experience features. To ensure the full enjoyment of the enhanced experience, immersive technologies need optimization both at display and signal levels. For example, in the case of HDR imaging, backlight dimming [62] and tone mapping algorithms that can display an HDR image in a regular display [63] are still under investigation. Furthermore, to drive the optimization of such immersive technologies, it is essential to: (1) properly understand the impact of an enhanced dimension on the eventual QoE; and (2) assess whether such attribute enhancement modifies the impact of other attributes on QoE.

The second important evolution in subjective QoE assessment is the inclusion of affective evaluations within QoE measurements. The affective state of the user (i.e., his/her mood or specific emotional state) may have impacts on the way a viewing experience is appreciated [64]. In turn, the emotion of human and aesthetics should be considered in QoE assessment paradigms. However, two challenges need to be overcome: (1) appropriate methodologies to measure the affective impact of media in relation to QoE have yet to be determined; (2) the effects of affective states pre-exists from that of the emotional state induced by the viewing experience itself. Furthermore, understanding and quantifying the relationship among physical properties of the image, their perception, and their impact on the user affective state is therefore a key challenge for upcoming QoE research.

Another change for subjective QoE is the shifting of the approaches employed for subjective tests conducting, from traditional lab-based psychometric ones to Internet-based environments. Lately, more interests around paired comparison (PC) [65–67] are growing for the QoE, and a few methods have been developed for establishing confidence intervals to the quality scores provided by the PC tests. With respect to the evolution of multimedia technology, new paradigms are in demand for carrying subjective experiments in which crowdsourcing [68] has become more and more eye-catching.

Apart from the remarkable existing works, both subjective and objective assessments still have large space for improvements, especially the image quality evaluation for emerging applications. More subjective tests are in need to discover more potential principles and relations between image quality and human perception. Furthermore, new image databases are required to be built for newly emerging applications. These new databases are further used for testing the objective IQA metrics developed. Up to now, FR IQA approaches have already had high consistency with subjective quality scores. However, without any reference image information, it is extremely challenging for NR IQA. It is very important to figure out the representations to accurately measure image quality. On the other hand, with the development of multimedia technology, some new IQA approaches should be proposed to deal with the new applications, such as multiply distorted images, mobile based quality assessment, HDR images, and retargeted images, etc. In addition, successful NR IQA metrics exploiting some knowledge of human brain is also a new trend and needed to be well developed.

In the mobile signal quality evaluation, diverse distortion factors affect the quality of mobile videos in different ways and have quite different characteristics. Although many subjective and objective studies have been conducted and performed, further studies are needed to investigate subjective testing methodologies and environments, by considering characteristics of mobile devices and viewing behavior. And for objective assessment, it is also important to evaluate their relative performance via thorough benchmarking studies, which would be more desirable in the future. Moreover, HVS properties and video contents should be considered into the assessment. Perceptual algorithms developed for mobile videos may be more consistent with human viewing results. Due to fast technological development, new types of media are being introduced to consumers such as 3D videos, HDR videos, and ultra high definition (UHD) videos. There is or will be a high demand to consume these types of media in the mobile environment, for which perceptual quality assessment will also play an important role. Future research in these fields will be valuable.

HDR is a newly developing multimedia technique for producing high contrast and luminance ranges. The major difference between HDR and the traditional LDR is the much more bits for a luminance component representation in HDR images. The resultant consequence is that it becomes more difficult for HDR content storage and display. TMOs can efficiently reduce the storage memory for HDR content, but also lose some visual information with distortion artifacts. Therefore, the first open issue need to be tackled for HDR is trying to decrease information loss while maintaining the high compression ratio. Improved TMOs and/or other new compression algorithms can be considered in the future. The next open problem is to develop new display techniques for HDR content. On the one hand, the price of specialized monitors for HDR needs to be down. On the other hand, with the development of electronics and material, perhaps no more specialized devices are required. These two open issues are badly in need to be addressed, because they deeply affect the visual QoE and quality measurement. Moreover, due to the complicated relations of HDR quality and distortions, and the fact that mathematical models for LDR are not suitable for HDR, very few methods for HDR quality assessment have been proposed, and much more efforts are required in both theories and implementations. Available standard databases are in urgent need as well.

With respect to medical IQA, most of the existing assessment metrics are pixel-based ones without considering HVS properties, even the recently proposed mutual information (MI) method. Therefore, one of the new trends for medical IQA could be developing HVS-based quality measurement algorithms. In addition, by considering the real medical applications, such as image denoising, enhancement, and deblur, more relevant studies should be conducted, and more practical assessment metrics are needed to be proposed. Considering the newly developed HDR or mobile applications, relevant researches can be conducted in the future.

When the 2D images and videos transfer to 3D, the properties of images and videos have a significant change, especially that 3D contents have depth information while not for 2D ones. However, how to use the depth information is still under exploration. Therefore, novel 3D multimedia quality assessment methods should be developed by considering both the basic structure information and the additional depth information. In addition, more attention should be attracted for real applications of 3D visual quality assessment.

As demonstrated in Chap. 8, the performances of the objective quality metrics for retargeted images are still not good enough. The statistical correlations between subjective MOS values and the objective metric outputs are inconsistent. It is in badly need to figure out how to capture and use the source image content, the retargeting scale, the shape distortion, the content information loss, and the HVS properties. As for the shape distortion description, the recently developed metrics tried to capture object shape of the image, but do not have good performances. In order to accurately depict perceptual quality of the retargeted images, shape distortions by retargeting processing need to be captured more precisely. Gabor filter is believed to have the ability to well extract image structures, such as edges and bars. Another important point is how to fuse the shape distortion and content information loss together by considering their corresponding contributions to the final image quality. Apart from the two aforementioned factors, source image quality and retargeting scale also affect the image quality. In addition, image content and HVS saliency should be considered to predict the retargeted image quality, which is expected to develop a more effective quality metric for retargeted images, as the image content correlates closely to the crop margin of the source image, and the shape distortions and content information loss in the salient regions are more sensitively perceived by the viewers than those in the non-salient regions.

Nowadays, there are several emerging trends of computer graphics quality assessment: (1) using machine learning technique to predict the visual quality by learning the input–output relationship; (2) paying more attention to 3D animation quality evaluation. There still exist very few works about quality assessment for dynamic meshes and articulated meshes; (3) material and lighting information should be considered together with geometric information; (4) merging image and model artifacts is also necessary, since visual defects may appear at several stages of a computer graphics work-flow, and may contain different types of data: either the 3D models, or the final rendered or tone mapped images. Integrating the lighting and material information, considering both image and model-based metrics, and utilizing the machine learning approach could be another way for graphics quality assessment.

From all the discussions mentioned above, we can see that various quality assessment researches have similar difficulties, e.g., how to capture good representations of objects, how to combine the HVS properties, etc. In the future, more researches and studies are expected to handle these issues in a better way.

References

1. D. Wang, F. Speranza, A. Vincent, T. Martin, and P. Blanchfield, “Toward optimal rate control: a study of the impact of spatial resolution, frame rate, and quantization on subjective video quality and bit rate,” in *Visual Communications and Image Processing.* International Society for Optics and Photonics, pp. 198–209, 2003.
2. P. Pérez, M. Jesús, J. R. Jaime, and G. Narciso, “Effect of packet loss in video quality of experience,” *Bell Labs Technical Journal.* vol. 16, no. 1, pp. 91–104, 2011.
3. L. Goldmann, D. S. Francesca, D. Frederic, E. Touradj, T. Rudolf, and L. Mauro, “Impact of video transcoding artifacts on the subjective quality,” In Quality of Multimedia Experience (QoMEX), 2010 Second International Workshop on, pp. 52–57. IEEE, 2010.
4. B. Girod, “What’s wrong with mean-squared error?” In *Digital Images and Human Vision*, pp. 207–220. MIT press, 1993.
5. D. M. Chandler, “Seven challenges in image quality assessment: past, present, and future research” ISRN Signal Processing, 2013.
6. J. Allnatt, “Transmitted-picture assessment” Chichester, UK: Wiley, 1983.
7. B. Keelan, “Handbook of image quality: characterization and prediction,” CRC Press, 2002.
8. P. G. Engeldrum, “Psychometric scaling: a toolkit for imaging systems development,” Imcotek Press, 2000.
9. “Methodology for the subjective assessment of the quality of television pictures,” ITU-R Recommendation BT.500-11, Geneva, 2002.
10. “Subjective audiovisual quality assessment methods for multimedia applications,” ITU-T Recommendation P.911, Geneva, 1998.
11. “Subjective methods for the assessment stereoscopic 3DTV systems,” International Telecommunication Union, Geneva, 2012.
12. B. Keelan, and H. Urabe, “ISO 20462, A psychophysical image quality measurement standard,” Proc. SPIE 5294, pp. 181–189, 2004.
13. S. Bech, H. Roelof, N. Marco, T. Kees, L. D. J. Henny, H. Paul, and K. P. Sakti, “Rapid perceptual image description (RaPID) method,” In *Electronic Imaging: Science and Technology*, pp. 317–328. International Society for Optics and Photonics, 1996.
14. A. B. Watson, “Efficiency of a model human image code,” JOSA A, vol. 4, no. 12, pp. 2401–2417, 1987.
15. J. A. Redi, and I. Heynderickx, “Image integrity and aesthetics: towards a more encompassing definition of visual quality,” In Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 8291, No. 5, p. 35, 2012.
16. J. A. Solomon, A. B. Watson, and A. Ahumada, “Visibility of DCT basis functions: Effects of contrast masking,” In Data Compression Conference, DCC’94. Proceedings IEEE, pp. 361–370, Mar, 1994.
17. A. M. Haun, and E. Peli, “Is image quality a function of contrast perception?” In IS&T/SPIE Electronic Imaging. International Society for Optics and Photonics, pp. 86510C–86510C, 2013.
18. P. G. Barten, “Contrast sensitivity of the human eye and its effects on image quality,” Washington: SPIE Optical Engineering Press, vol. 21, 1999.
19. T. N. Pappas, R. J. Safranek, and J. Chen, “Perceptual criteria for image quality evaluation,” Handbook of image and video processing, pp. 669–684, 2000.

20. H. Liu, and I. Heynderickx, “A perceptually relevant no-reference blockiness metric based on local image characteristics,” EURASIP Journal on Advances in Signal Processing, 2009.
21. Z. Wang, and X. Shang, “Spatial pooling strategies for perceptual image quality assessment,” In Image Processing, 2006 IEEE International Conference on, pp. 2945–2948, 2006.
22. Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” Image Processing, IEEE Transactions on, vol. 13, no. 4, pp. 600–612, 2004.
23. R. Ferzli, and L. J. Karam, “A no-reference objective image sharpness metric based on the notion of just noticeable blur (JNB),” Image Processing, IEEE Transactions on, vol. 18, no. 4, pp. 717–728, 2009.
24. U. Engelke, H. Kaprykowsky, H. J. Zepernick, and P. Ndjiki-Nya, “Visual attention in quality assessment,” Signal Processing Magazine, IEEE, vol. 28, no. 6, pp. 50–59, 2011.
25. J. Redi, H. Liu, R. Zunino, ans I. Heynderickx, “Interactions of visual attention and quality perception,” In IS&T/SPIE Electronic Imaging, International Society for Optics and Photonics, pp. 78650S–78650S, Feb, 2011.
26. R. Desimone, and J. Duncan, “Neural mechanisms of selective visual attention,” Annual review of neuroscience, vol. 18, no. 1, pp. 193–222, 1995.
27. H. Alers, J. Redi, H. Liu, I. Heynderickx, “Studying the effect of optimizing image quality in salient regions at the expense of background content,” J. Electron. Imaging, vol. 22, no. 4, pp. 043012–043012, 2013.
28. M. Fiedler, T. Hossfeld, and P. Tran-Gia, “A generic quantitative relationship between quality of experience and quality of service,” Network, IEEE, vol. 24, no. 2, pp. 36–41, 2010.
29. H. J. Kim, D. H. Lee, J. M. Lee, K. H. Lee, W. Lyu, and S. G. Choi, “The QoE evaluation method through the QoS-QoE correlation model,” In Networked Computing and Advanced Information Management, 2008. NCM’08. Fourth International Conference on Network, IEEE, vol. 2, pp. 719–725, 2008.
30. M. Siller, and J. Woods, “Improving quality of experience for multimedia services by QoS arbitration on a QoE framework,” In Proc. of the 13th Packed Video Workshop, 2003.
31. G. Ghinea, and J. P. Thomas, “Quality of perception: user quality of service in multimedia presentations,” Multimedia, IEEE Transactions on, vol. 7, no. 4, pp. 786–789, 2005.
32. H. Ridder, and S. Endrikhovski, “Image quality is FUN: reflections on fidelity, usefulness and naturalness,” In SID Symposium Digest of Technical Papers, Blackwell Publishing Ltd, vol. 33, no. 1, pp. 986–989, May, 2002.
33. E. Fedorovskaya, C. Neustaedter, and W. Hao, “Image harmony for consumer images,” In Image Processing, 15th IEEE International Conference on, 2008.
34. P. Kortum, and M. Sullivan, “The effect of content desirability on subjective video quality ratings,” Human factors: the journal of the human factors and ergonomics society, vol. 52, no. 1, pp. 105–118, 2010.
35. W. A. Mansilla, A. Perkis, ans T. Ebrahimi, “Implicit experiences as a determinant of perceptual quality and aesthetic appreciation,” In Proceedings of the 19th ACM international conference on Multimedia, pp. 153–162, Nov, 2011.
36. S. Mann and R. Picard, “Being ’Undigitial’ with Digital Cameras: Extending Dynamic Range by Combining Differently Exposed Pictures,” In: Proceedings of IS&T 48th Annual Conference, Society for Imaging Science and Technology, pp. 422–428, 1995.
37. Spheron, “Spheron HDR VR,” 2008, Available at: http://www.spheron.com/home.html.
38. G. Ward, “Real Pixels,” Graphic Gems, pp. 15–31, 1991.
39. G. Ward, “LogLuv Encoding for Full-Gamut High Dynamic Range Images,” Journal of Graphics Tools, vol. 3, no. 1, 1998.
40. “Industrial Light & Magic,” OpenEXR, 2008, Available at: http://www.openexr.com/.
41. G. Ward and M. Simmons, “JPEG-HDR: A Backwards-Compatible High Dynamic Range Extension to JPEG,” In: Proceedings of ACM SIGGRAPH 2006 Courses, 2006.
42. N. Sugiyama, H. Kaida, X. Xue, T. Jinno, N. Adami, and M. Okuda, “HDR Compression Using Optimized Tone Mapping Model,” In: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1001–1004, 2009.

43. R. Mantiuk, A. Efremov, K. Myszkowski, and H. Seidel, “Backward Compatible High Dynamic Range MPEG Video Compression,” ACM Transactions on Graphics, vol. 25, no. 3, pp. 713–723, 2006.
44. F. Banterle, K. Debattista, A. Artusi, S. Pattanaik, K. Myszkowski, P. Ledda, and A. Chalmers, “High Dynamic Range Imaging and Low Dynamic Range Expansion for Generating HDR Content,” Computer Graphics Forum, vol. 28, no. 8, 2009.
45. M. Cadik, M. Wimmer, L. Neumann, and A. Artusi, “Evaluation of HDR tone mapping methods using essential perceptual attributes,” Computers & Graphics, vol. 32, pp. 330–349, 2008.
46. F. Drago, WL. Martens, K. Myszkowski, and H. Seidel, “Perceptual evaluation of tone mapping operators,” In: Proceedings of the SIGGRAPH 2003 conference on sketches & applications, New York, NY, USA: ACM Press, 2003.
47. J. Kuang, H. Yamaguchi, C. Liu, G. Johnson, and M. Fairchild, “Evaluating HDR rendering algorithms,” ACM Transactions on Applied Perception, vol. 4, no. 9, 2007.
48. A. Yoshida, V. Blanz, K. Myszkowski, and H. Seidel, “Perceptual evaluation of tone mapping operators with real-world scenes,” Human Vision & Electronic Imaging X, San Jose, CA, USA: SPIE, pp. 192–203, 2005.
49. P. Ledda, A. Chalmers, T. Troscianko, and H. Seetzen, “Evaluation of tone mapping operators using a high dynamic range display,” In: Proceedings of the 32nd annual conference on computer graphics and interactive techniques, ACM Press, pp. 640–648, 2005.
50. M. Ashikhmin, J. Goyal, “A reality check for tone-mapping operators,” ACM Transactions on Applied Perception, vol. 3, no. 4, pp. 399–411, 2006.
51. G. Eilertsen, R. Wanat, R. Mantiuk, and J. Unger, “Evaluation of tone mapping operators for HDR-video,” In: Computer Graphics Forum Special Issue Proceedings of Pacific Graphics, 2013.
52. M. Narwaria, M. Silva, P. Callet, and R. Pepion, “Tone mapping Based High Dynamic Range Image Compression: Study of Optimization Criterion and Perceptual Quality,” Optical Engineering (Special Issue on High Dynamic Range Imaging), vol. 52, no. 10, 2013.
53. M. Narwaria, M. Silva, P. Callet, and R. Pepion, “Impact of Tone Mapping In High Dynamic Range Image Compression,” In: Proc. Eighth International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM), 2014.
54. R. Mantiuk, K. Jim, A. Rempel, and W. Heidrich, “HDR-VDP-2: a calibrated visual metric for visibility and quality predictions in all luminance conditions,” in ACM Transactions on Graphics (TOG), vol. 30, no. 4, 2011.
55. D. Tsai, Y. Lee, and E. Matsuyama, “Information entropy measure for evaluation of image quality,” J Digit Imaging, vol. 21, pp. 338–347, 2008.
56. E. Samei, T. R. Nicole, T. D. James, and C. Ying, “Intercomparison of methods for image quality characterization. I. Modulation transfer functiona,” Medical physics, vol. 33, no. 5, pp. 1454–1465, 2006.
57. U. Neitzel, G.-K. Susanne, B. Giovanni, and S. Ehsan, “Determination of the detective quantum efficiency of a digital x-ray detector: Comparison of three evaluations using a common image data set,” Medical physics, vol. 31, no. 8, pp. 2205–2211, 2004.
58. M. Spahn, “Flat detectors and their clinical applications,” Eur Radiol, vol. 15, pp. 1934–1947, 2005.
59. K. Fettery, and N. Hangiandreou, “Effect of x-ray spectra on the DQE of a computed radiography system,” Med Phys, vol. 28, pp. 241–249, 2001.
60. T. O. Aydin, R. Mantiuk, K. Myszkowski, and H. P. Seidel, “Dynamic range independent image quality assessment,” ACM Transactions on Graphics (Proc. of SIGGRAPH), vol. 27, no. 3, 2008.
61. T. O. Aydin, M. Cadik, K. Myszkowski, and H. P. Seidel, “Video quality assessment for computer graphics applications,” ACM Transactions on Graphics (Proc. of SIGGRAPH), vol. 29, no. 6, 2010.

62. J. Korhonen, C. Mantel, N. Burini, and S. Forchhammer, “Searching for the preferred backlight intensity in liquid crystal displays with local backlight dimming,” In Quality of Multimedia Experience (QoMEX), 2013 Fifth IEEE International Workshop on, July, 2013.
63. A. Yoshida, V. Blanz, K. Myszkowski, and H. P. Seidel, “Perceptual evaluation of tone mapping operators with real-world scenes,” In Electronic Imaging 2005 International Society for Optics and Photonics, 2005.
64. I. Wechsung, M. Schulz, K. P. Engelbrecht, J. Niemann, ans S. Moller, “All users are (not) equal-the influence of user characteristics on perceived quality, modality choice and performance,” In Proceedings of the Paralinguistic Information and its Integration in Spoken Dialogue Systems Workshop, Springer New York, Jan, 2011.
65. J-S Lee, F. D. Simone, and T. Ebrahimi, “Subjective quality evaluation via paired comparison: application to scalable video coding,” IEEE Transactions on Multimedia, vol. 13, no. 5, pp: 882–893, 2011.
66. C-C Wu, K-T Chen, Y-C Chang, and C-L Lei, “Crowdsourcing multimedia qoe evaluation: A trusted framework,” IEEE transactions on multimedia, vol. 15, no. 5, pp: 1121–1137, 2013.
67. Q. Xu, Q. Huang, T. Jiang, B. Yan, W. Lin, and Y. Yao, “Hodgerank on random graphs for subjective video quality assessment,” IEEE Transactions on Multimedia, vol. 14, no. 3, pp: 844–857, 2012.
68. J. Howe, “The rise of crowdsourcing,” Wired magazine, vol. 14, no. 6, pp: 1–4, 2006.