MULTIMEDIA QUALITY OF EXPERIENCE (QoE)

MULTIMEDIA QUALITY OF EXPERIENCE (QoE)

CURRENT STATUS AND FUTURE REQUIREMENTS

Edited by

Chang Wen Chen
State University of New York at Buffalo, USA

Periklis Chatzimisios
Alexander Technological Educational Institute, Thessaloniki, Greece

Tasos Dagiuklas
Hellenic Open University, Greece

Luigi Atzori
University of Cagliari, Italy

WILEY

This edition first published 2016

Registered office
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

For details of our global editorial offices, for customer services and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Designations used by companies to distinguish their products are often claimed as trademarks. All brand names and product names used in this book are trade names, service marks, trademarks or registered trademarks of their respective owners. The publisher is not associated with any product or vendor mentioned in this book.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. It is sold on the understanding that the publisher is not engaged in rendering professional services and neither the publisher nor the author shall be liable for damages arising herefrom. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

The advice and strategies contained herein may not be suitable for every situation. In view of ongoing research, equipment modifications, changes in governmental regulations, and the constant flow of information relating to the use of experimental reagents, equipment, and devices, the reader is urged to review and evaluate the information provided in the package insert or instructions for each chemical, piece of equipment, reagent, or device for, among other things, any changes in the instructions or indication of usage and for added warnings and precautions. The fact that an organization or Website is referred to in this work as a citation and/or a potential source of further information does not mean that the author or the publisher endorses the information the organization or Website may provide or recommendations it may make. Further, readers should be aware that Internet Websites listed in this work may have changed or disappeared between when this work was written and when it is read. No warranty may be created or extended by any promotional statements for this work. Neither the publisher nor the author shall be liable for any damages arising herefrom.

Library of Congress Cataloging-in-Publication Data

Dagiuklas, Tasos, author.
Multimedia quality of experience (QoE) : current status and future requirements / Tasos Dagiuklas, Luigi Atzori, Periklis Chatzimisios, Chang Wen Chen.
pages cm
Includes index.
ISBN 978-1-118-48391-6 (hardback)
1. Multimedia communications–Quality control. I. Title.
TK5105.15.D34 2015
006.7068′5–dc23
2015033891

A catalogue record for this book is available from the British Library.

ISBN: 9781118483916

Cover image: scanrail/iStockphoto

Set in 10/12pt Times by Aptara Inc., New Delhi, India

1 2016

Contents

About the Editors

Chang Wen Chen is Professor of Computer Science and Engineering at the University at Buffalo, State University of New York. He was Allen Henry Endow Chair Professor at the Florida Institute of Technology from July 2003 to December 2007; on the Electrical and Computer Engineering faculty at the University of Rochester from 1992 to 1996; and on the Electrical and Computer Engineering faculty at the University of Missouri-Columbia from 1996 to 2003.

Professor Chen has been the Editor-in-Chief for *IEEE Transactions on Multimedia* since January 2014. He has also served as the Editor-in-Chief for *IEEE Transactions on Circuits and Systems for Video Technology* from 2006 to 2009, and been an editor for several other major IEEE transactions and journals.

Professor Chen received a BS from the University of Science and Technology of China in 1983, an MSEE from the University of Southern California in 1986, and a PhD from the University of Illinois at Urbana-Champaign in 1992. He and his students have received eight Best Paper Awards or Best Student Paper Awards over the past two decades. He has also received several research and professional achievement awards, including the Sigma Xi Excellence in Graduate Research Mentoring Award in 2003, the Alexander von Humboldt Research Award in 2009, and the State University of New York at Buffalo Exceptional Scholar – Sustained Achievement Award in 2012. He is an IEEE Fellow and a SPIE Fellow.

Tasos Dagiuklas received his first degree from the University of Patras (Greece) in 1989, an MSc from the University of Manchester (UK) in 1991, and a PhD from the University of Essex (UK) in 1995, all in Electrical Engineering. He is Assistant Professor at the School of Science and Technology at the Hellenic Open University, Greece. He currently leads the Converged Networks and Services (CONES) Research Group (http://cones.eap.gr), carrying out research in the areas of Future Internet Architectures, Media Optimization, and Cloud Infrastructures and Services. Dr. Dagiuklas is a Senior Member of the IEEE, the Chair for IEEE MMTC 3DIG WG, and an IEEE MMTC E-Board Member. He has served as Vice-Chair for IEEE MMTC QoE WG and a Key Member for IEEE MMTC MSIG and 3DIG WGs. He is Associate Technical Editor for *IEEE Communications Magazine*. He has served as guest editor for many scientific journals. He is a reviewer for journals such as *IEEE Transactions on Multimedia*, *IEEE Communication Letters*, and *IEEE Journal on Selected Areas in Communications*. His research interests include FTV, 3DV, media optimization across heterogeneous networks, QoE, and cloud infrastructures and services.

Luigi Atzori is Associate Professor at the Department of Electrical and Electronic Engineering at the University of Cagliari (Italy) and Research Associate at the Multimedia Communications Laboratory of CNIT (Consorzio Nazionale Interuniversitario per le Telecomunicazioni). His research interests are in multimedia communications and computer networking (wireless and wireline), with emphasis on multimedia QoE, multimedia streaming, NGN service management, service management in wireless sensor networks, architecture and services in the Internet of Things.

Professor Atzori is a Senior Member of the IEEE (since 2009), Chair of the Steering Committee of the IEEE Multimedia Communications Committee (MMTC), and Co-Chair of the IEEE 1907.1 standard on "Network-Adaptive Quality of Experience (QoE) Management Scheme for Real-Time Mobile Video Communications". He is the coordinator of the Marie Curie Initial Training Network on QoE for multimedia services, involving ten European institutions and one in South Korea.

Professor Atzori has been an Editor for *Wireless Networks Journal* and a Guest Editor for *IEEE Communications Magazine, Monet Journal, and Signal Processing: Image Communications Journal.* He is a member of the editorial board of *IEEE Internet of Things Journal* and *Ad Hoc Networks.* He has served as technical program chair for various international conferences and workshops. He has also served as a reviewer and panelist for many funding agencies, including FP7, Cost, the Italian Ministry of Education, Universities and Research (MIUR), and Regional.

Periklis Chatzimisios (SMIEEE) is Associate Professor at the Department of Informatics at Alexander TEI of Thessaloniki (Greece). Recently, he has been a Visiting Academic/Researcher at the University of Toronto (Canada) and the Massachusetts Institute of Technology (USA). Dr. Chatzimisios is involved in several standardization activities, serving as a Member of the Standards Development Board for the IEEE Communication Society (ComSoc) (2010–today), as Secretary of the IEEE 1907.1 Standardization Working Group, and lately as an active member of IEEE Research Groups on the Internet of Things, Communications & Networking Infrastructure, and Software Defined & Virtualized Wireless Access.

Dr. Chatzimisios has served as Organizing/TPC Committee Member for more than 150 conferences and Founder/Organizer/Co-Chair for many workshops co-allocated with flagship IEEE/ACM conferences. He also holds editorial board positions for several IEEE/non-IEEE journals and is the Director (Co-Director during 2012–2014) for the E-letter of the IEEE Technical Committee on Multimedia Communications (MMTC). He is the author/editor of eight books and more than 85 peer-reviewed papers and book chapters on the topics of multimedia communications, performance evaluation, and standardization activities of mobile/wireless communications – with more than 1300 citations by other researchers. Dr. Chatzimisios received his PhD from Bournemouth University (UK) (2005) and his BSc from Alexander TEI of Thessaloniki, Greece (2000).

List of Contributors

Luigi Atzori, University of Cagliari, Italy

Alan C. Bovik, University of Texas at Austin, USA

Periklis Chatzimisios, Alexander Technological Educational Institute, Thessaloniki, Greece

Chang Wen Chen, State University of New York at Buffalo, USA

Tasos Dagiuklas, Hellenic Open University, Greece

Katrien De Moor, NTNU, Trondheim, Norway

Yuming Fang, Nanyang Technological University, Singapore

Chaminda T.E.R. Hewage, Kingston University, UK

Utsaw Kumar, Intel Corporation, USA

Weisi Lin, Nanyang Technological University, Singapore

Maria G. Martini, Kingston University, UK

Anish Mittal, Nokia Research Center, USA

Anush K. Moorthy, Qualcomm Inc., USA

Rana Morsi, Intel Corporation, USA

Moustafa M. Nasralla, Kingston University, UK

Ognen Ognenoski, Kingston University, UK

Ozgur Oyman, Intel Corporation, USA

Vishwanath Ramamurthi, Intel Corporation, USA

Mohamed Rehan, Intel Corporation, USA

Peter Reichl, Université Européenne de Bretagne/Télécom Bretagne, France and University of Vienna, Austria

Lea Skorin-Kapov, University of Zagreb, Croatia

Martín Varela, VTT Technical Research Centre, Finland

Stefan Winkler, Advanced Digital Sciences Center (ADSC), Singapore

Hong Ren Wu, RMIT, Australia

Preface

The effectiveness of distributed multimedia applications as well as mobile computing services – which are becoming dominant in the modern telecommunications era – is primarily based on the networking protocols and communication systems that deliver content to the end-user. Research and development in these protocols and delivery systems is currently being driven from a technical perspective for the end-user's benefit. However, it is a fact that the effectiveness of any service presentation is ultimately measured by the end-user's experience in terms of aesthetic quality, accuracy of information, system responsiveness, and many other impacting factors.

Quality of Experience (QoE) can be defined as the overall acceptability of an application or service strictly from the end-user's point of view. It is a subjective measure of end-to-end service performance from the user's perspective, and it is an indication of how well any system and network components meet the user's needs. Encompassing many different aspects, QoE rivets on the true feelings of end-users when they watch streaming video and podcasts, listen to digitized music, and browse the Internet through a plethora of methods and devices.

The problem of understanding and enhancing QoE in complex, distributed, and diverse environments has been and is continuing to be the subject of intense research investigation. Considerable effort has been devoted to assessing QoE via objective or subjective means for new and emerging multimedia services over modern fixed/mobile devices (e.g., IPTV/HDTV/3DTV, tablet video calls, 3D smartphones). Many researchers have looked at this as a usability problem, while others have studied the correlation between specific technological settings and user-perceived QoE. However, as of today, we do not know how to manage and control QoE in a diverse heterogeneous environment. The variables that affect QoE are just too many and span several interdisciplinary areas, including multiple technologies, but also psychological and sociological factors. Despite the effort devoted to QoE study, managing and controlling user QoE is still an open issue. Currently, services and applications offer QoE as a by-product of QoS management. Most commonly, QoE is achieved by over-provisioning and over-committing network and computational resources. Therefore, QoE is still a best-effort service, which is not a viable option when applications become multimodal (a complex combination of voice, video, and data). In these cases, resources have to be managed and controlled more accurately and proactively for a successful, QoE-assured, service delivery.

1

Introduction

Tasos Dagiuklas[1], Luigi Atzori[2], Chang Wen Chen[3] and
Periklis Chatzimisios[4]
[1]Hellenic Open University, Greece
[2]University of Cagliari, Italy
[3]State University of New York at Buffalo, USA
[4]Alexander Technological Educational Institute, Thessaloniki, Greece

During recent years, Quality of Experience (QoE) has established itself as a topic in its own right for both industrial and academic research. With its focus on the end-user in terms of acceptability, delight, and performance, it is about to take over the role of Quality of Service (QoS) as the key paradigm for provisioning and managing services and networks.

According to Wikipedia, "Quality in business, engineering and manufacturing has a pragmatic interpretation as the non-inferiority or superiority of something; it is also defined as fitness for purpose. Quality is a perceptual, conditional, and somewhat subjective attribute and may be understood differently by different people. Consumers may focus on the specification quality of a product/service, or how it compares to competitors in the marketplace." Quality is a term that has been defined since ancient times. In philosophy, quality (from the Latin *qualitas*) is an attribute or property. The Ancient Greek philosopher Aristotle analyzed qualities in his logical work *Categories*, where all objects of human comprehension are classified into ten categories. Quality is one of these categories.

The term "quality" appears in various standardization fora. As an example, the International Organization for Standardization (ISO) has defined various standards related to quality, as indicated below:

1. ISO 8402-1986 standard defines quality as "the totality of features and characteristics of a product or service that bears its ability to satisfy stated or implied needs."
2. ISO 9000 is a series of standards that define, establish, and maintain a quality assurance system for manufacturing and service industries. The ISO 9000 family addresses

Multimedia Quality of Experience (QoE): Current Status and Future Requirements, First Edition.
Edited by Chang Wen Chen, Periklis Chatzimisios, Tasos Dagiuklas and Luigi Atzori.

various aspects of quality management and contains some of the ISO's best-known standards. The standards provide guidance and tools for companies and organizations who want to ensure that their products and services consistently meet customers' requirements, and that quality is consistently improved. Standards in the ISO 9000 family include the following:

- ISO 9001:2008 – sets out the requirements of a quality management system.
- ISO 9000:2005 – covers the basic concepts and language.
- ISO 9004:2009 – focuses on how to make a quality management system more efficient and effective.
- ISO 19011:2011 – sets out guidance on internal and external audits of quality management systems.

In order to lead in today's communications services market, network operators and content/service providers must offer customers the best user experience for premium services. In the past, networks have been examined objectively by measuring a number of criteria to determine network quality using QoS. In effect, QoS refers to the ability of a network to achieve more deterministic behavior, so data can be transported by optimizing parameters such as packet loss, delay, jitter, and bandwidth consumption. One should note that QoS does not consider the end-user's perception.

The perceived quality of media services is a crucial subject for service and content providers, with growing market competition. QoE for next-generation multimedia applications will be a combination of measurable QoS parameters considering both network and service environment and non-quality parameters. Network-based parameters include bandwidth, delay, jitter, packet loss, and PER (Packet Error Rate); service-oriented parameters (especially in the case of video services) may be metrics such as PSNR (Peak Signal-to-Noise Ratio) or MSE (Mean Square Error). Other factors that are not quality based but are important to quantify QoE include: screen size (e.g., mobile phone vs. large TV set), screen illumination (e.g., mobile terminal on a cloudy day vs. high contrast in a cinema environment), viewing distance, content (e.g., movie with action vs. news from a broadcaster), application (e.g., social networking vs. medical vs. distance learning), price (Skype videoconferencing vs. regular cellular video call), user profile (e.g., teenager vs. professional).

QoE is a new research topic that is currently being addressed by the various standardization fora. In the networked media delivery industry, guarantee of user experience is a key factor for many media-aware applications and services. Therefore, contrary to QoS, the concept of QoE has highlighted concerns for the media delivery industry – referring to "the overall acceptability of an application or service, as perceived subjectively by the end user." The media delivery industry views end-user QoE monitoring as either "critical" or "very important" to their media initiatives; meanwhile, the foremost issue reported by industry is that the current QoE assessment solutions are too costly and not accurate enough to measure end-user experience.

QoE is a subjective metric that involves human dimensions; it ties together user perceptions, expectations, and experiences of applications and network performance. It is now widely acknowledged that the adoption of new multimodal media necessitates mechanisms to assess and evaluate perceived multimedia quality. QoE is defined as a metric to assess end-user experience at the perceptual pseudo-layer located above application and network layers.

Considerable effort has been devoted to assessing QoE via objective or subjective means for new and emerging multimedia services over modern fixed/mobile devices (e.g., IPTV/HDTV/3DTV, tablet, 3D smartphone). Many researchers have looked at this as a usability problem, while others have studied the correlation between specific technological settings and user-perceived QoE. As of today, we do not know how to manage and control QoE in a diverse heterogeneous environment. The variables that affect QoE are just too wide and too many. Hence, managing and controlling user QoE is still an open issue. Currently, services and applications offer QoE as a by-product of QoS management. Most commonly, QoE is achieved by over-provisioning and over-committing network and computational resources. Therefore, QoE is still a best-effort service. As applications become multimodal, resources will have to be managed and controlled more accurately and proactively for successful QoE-assured service delivery.

The overall structure of the book is as follows:

Chapter 2 outlines the QoE defining a user-centric concept of service quality. It provides QoE in various standardization fora such as ITU, ETSI, and IETF. It provides factors influencing QoE such as human Influencing Factors (IFs), system IFs, and context IFs. QoE is defined for different services such as speech, video, HTTP streaming, and cloud-based services. Moreover, it outlines a set of factors at the human, system, and context level that – either independently or interlinked – may influence QoE. Finally, the role of QoE in communication ecosystems is defined so that the user experience is optimized.

Chapter 3 reviews existing objective QoE methodologies and provides a taxonomy of objective quality metrics that may be grouped using the characteristics of the human visual system and the availability of the original signal. The chapter also presents the basic computational modules for perceptual quality metrics; quality metrics for images, video, and audio/speech; and joint audiovisual quality metrics.

Chapter 4 describes QoE for HTTP adaptive streaming services and presents QoE-based optimization strategies for Internet video. As a relatively new technology in comparison with traditional push-based adaptive streaming techniques, the deployment of Http Adaptive Streaming (HAS) services presents new challenges and opportunities for content developers, service providers, network operators, and device manufacturers. One of these important challenges is developing evaluation methodologies and performance metrics to accurately assess user QoE for HAS services, as well as effectively utilizing these metrics for service provision and optimization of network adaptation.

Chapter 5 emphasizes visual quality assessment covering both opinion-aware and opinion-unaware models. Most of the approaches are based on understanding and modeling the underlying statistics of natural images and/or distortions using perceptual principles. These approaches measure deviations from statistical regularities and quantify such deviations, leading to estimates of quality. The chapter presents the motivation and principles underlying such statistical descriptions of quality, and describes such algorithms in detail. Exhaustive comparative analysis of these approaches is provided, together with a discussion of the potential applications of no-reference algorithms.

The discussions so far have highlighted an increasing emphasis on QoE compared with QoS in audiovisual communication, broadcasting, and entertainment applications, which signals a transition from technology-driven services to user-centric (or perceived) quality-assured services. **Chapter 6** focuses on the issues underpinning the theoretical framework/models and methodologies for QoE subjective and objective evaluation of visual signal communication services. Issues relevant to human visual perception and quality scoring or rating for television and multimedia applications are discussed, while readers are referred to the standards documents and/or other monographs regarding specific details of the aforementioned standards.

In recent years, the concept of QoS has been extended to the new concept of QoE, reflecting the experience of the end-user accessing the provided service. Experience is user- and context-dependent. However, subjective QoE evaluation is time consuming and not suitable for use in closed-loop adaptations. Hence, objective (rather than subjective) QoE evaluation enables optimal use of available resources based on the defined objective utility index. The main aim of achieving a satisfactory QoE for the users of a system can be afforded at different layers of the protocol stack. On this basis, **Chapter 7** presents a review of recent strategies for QoE monitoring, control, and management, including new solutions for a variety of different service types. The chapter also considers QoE management and control in different scenarios, including wireless scenarios, adaptive streaming over HTTP, and transmission to multiple users.

Finally, **Chapter 8** completes the book by providing conclusions drawn from each of the previous chapters.

2

QoE–Defining a User-Centric Concept for Service Quality

Martín Varela[1], Lea Skorin-Kapov[2], Katrien De Moor[3] and Peter Reichl[4]
[1] VTT Technical Research Centre, Finland
[2] University of Zagreb, Croatia
[3] NTNU, Trondheim, Norway
[4] Université Européenne de Bretagne/Télécom Bretagne, France and University of Vienna, Austria

2.1 Introduction

Quality of Experience (QoE) has, in recent years, gained a prominent role in the research and related work of several fields, notably networking and multimedia, but also in other domains such as network economics, gaming, or telemedicine. It has, to some extent, also become a buzzword for marketers, network operators, and other service providers.

Historically, its origins can be traced to several sources, which are nowadays converging toward a common, mature understanding of what QoE actually is. Several of the key ideas behind QoE can be traced several decades back to research done, for example, by telephone operators into the quality of calls made through their systems, and of TV broadcasters in a quest to understand how users perceived the quality of television pictures. The issues involved here relate not only to the transmission aspects, but also to coding and equipment ones.

With the advent of Internet-based multimedia communication services, such as Voice over IP (VoIP), video streaming, video conferencing, etc., the role of the network's performance (often referred to as Quality of Service, QoS) became more important in determining the perceived quality of those services, and thus a part of the networking community also became involved in the research of *perceived QoS,* which has itself evolved to be called Quality of Experience (QoE) within the networking community.

Multimedia Quality of Experience (QoE): Current Status and Future Requirements, First Edition.
Edited by Chang Wen Chen, Periklis Chatzimisios, Tasos Dagiuklas and Luigi Atzori.

The prominence of the users and their experience within the QoE research community has risen as time goes by. This makes sense, since at the end of the day, what really matters for service providers is that users keep buying their services, and providing good QoE is a sure-fire way to keep users satisfied. Beyond this rather mundane rationale lies an interesting scientific challenge: QoE is actually all about the user. The technical (network/application) aspects involved in understanding it are but a fraction of the overall QoE picture. Granted, those technical aspects are critical, and more importantly, are tractable; but if we had to predict what the "next frontier" for QoE is, it is bound to be at the user level. The question then becomes, how can we move from codecs and packets to an understanding of how users actually experience those services, and which elements make or break the quality of these experiences. Recently, psychologists, sociologists, and researchers in other humanities-related fields – with long traditions in investigating different aspects related to human experiences – have also begun to look into how users perceive and are affected by the quality of the services they use. The scope here is quite wide, ranging from research of emotion and affective aspects, to economics.

It is no wonder that with people from such different backgrounds involved in what conceptually is the same topic, views on what QoE actually is can be, and to some extent often are, radically different. In this chapter we will discuss the most prominent existing definitions of QoE, QoS and their relation, factors that may influence QoE, the need for understanding it in the context of services, social and psychological aspects of QoE, and its role in future telecommunications ecosystems.

2.2 Definitions of QoE

As mentioned above, QoE – as a research domain – lies at the junction of several disciplines, and this is to some extent reflected in the different definitions that have been given for it throughout the years.

The International Telecommunications Union (ITU)'s definition, given in [1], states that QoE is

> *The overall acceptability of an application or service, as perceived subjectively by the end-user.*

The above definition is accompanied by two notes, which state that:

1. *QoE includes the complete end-to-end system effects, and*
2. *The overall acceptability may be influenced by user expectations and context.*

The definition given by the ITU clearly focuses on the acceptability of a service, but does not address the actual experience of the user of the said service. Moreover, it does not specify the complex notions of "user expectations" and "context," and the possible influence of other human factors is ignored.

An alternative definition was produced at a Dagstuhl seminar on QoE held in 2009 ("From Quality of Service to Quality of Experience") [2, 3], whereby QoE

> *… describes the degree of delight of the user of a service, influenced by content, network, device, application, user expectations and goals, and context of use.*

In contrast to the ITU definition, the Dagstuhl seminar's definition takes a purely hedonistic view of the quality experienced by the user, and tacitly puts the system under consideration in the role of an influencing factor, whereas in the ITU definition, it was the focus.

In 2012, a large collaborative effort within COST Action IC1003 – European Network on Quality of Experience in Multimedia Systems (Qualinet) produced a white paper on definitions of QoE [4],[1] which delved deeper than previous, efforts into providing a wider, and yet more precise, definition of QoE. The proposed definition states that 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.*

As with the Dagstuhl definition, the one proposed by Qualinet has a clear focus on the user. It does, however, take both a hedonistic and a utilitarian perspective, and more importantly, it explicitly introduces the user's personality and current state (mood). This reflects a visible shift in the state of QoE research that sees the user aspects become more important, and is accompanied by the addition of the humanities' perspectives and approaches.

An interesting fact regarding the evolution of these definitions of QoE is the increasing importance of the user with time. From the systems-oriented ITU definition, through the user-centric Dagstuhl one, and ending in the more explicit consideration of the user's personality and mood in the Qualinet definition, the changes in perspective reflect those that can to some degree already be observed in the state of the art as well.

2.3 Differences Between QoE and QoS

In contrast to QoE, the concept of QoS has a rich history spanning more than two decades, during which it has been subject to an immense number of related publications [5, 6]. Quality of service has been defined by the ITU-T [7] as

> *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.*

An earlier version of the definition has also been adopted by ETSI as their own definition of QoS [8].

With respect to the ITU-T's own definition of QoE, there is a clearly narrower scope in the definition of QoS, which limits itself to the *characteristics of a telecommunications service.* When we consider the newer and more inclusive Qualinet definition of QoE, the gap between the two widens significantly. The definition of QoS has a clear focus on the technical aspects of telecommunication services, whereas the QoE definition is focused on the user, and the services and applications it considers need not be telecommunications-related. The IETF takes an even more systems-oriented approach with their definition of QoS [9], which states that QoS is

> *...a set of service requirements to be met by the network while transporting a flow.*

[1] As a disclaimer, all the authors of this chapter contributed to that work.

The IETF's definition of QoS has a purely network-oriented focus, and arguably no connection with the concept of QoE.

Beyond the semantic distinction in terms of definitions, QoS has in practice (i.e., in the extensive literature dedicated to it) mostly connotations of network performance, and no real relation to the *stated and implied needs of the user*. These network performance connotations can be related to both measures of network performance (such as delay, throughput, or jitter), and architectures and mechanisms for improving network performance (such as differentiated services [10]). Under both of those meanings, the relation of QoS to the end-user is tenuous at best. Oftentimes, works in the literature refer to QoE, when they are really addressing QoS (for instance, a common occurrence is that of equating lower delays to higher QoE). While the network QoS can certainly have a significant bearing on QoE, the type of implications often seen in the literature do not always hold in practice, mainly for two reasons. Firstly, QoE is itself a multi-dimensional construct, of which only some dimensions can be affected by network performance.[2] Secondly, unless subjective user studies are carried out, or suitable models of user perception are used, it is very easy to overstate the impact of the network performance on actual QoE (i.e., the aforementioned lowered delays can, in fact, have no significant impact on the quality experienced by the users).

As a rule of thumb, we could say that we are only talking about QoE if at least the users' perception of quality is considered, be it via subjective assessment, or via suitable quality models.[3]

This is not to say that QoS is a concept unrelated to QoE. In practice, the performance aspects of a service and its underlying infrastructure are deeply – and non-trivially – related to the user's perception of the system's quality, and therefore to QoE. For media services in particular, there is a rather clear relation between the performance of the transport network (i.e., QoS) and the quality perceived by the user. This perceived quality can in turn be a dominant dimension (or set of dimensions) of the QoE of that particular service.

From a more practical perspective, QoS plays an important role in how service providers can affect the users' QoE, in several ways. There is the obvious aspect of perceived quality as mentioned above, but also more subtle relations between QoS and QoE – for example through service pricing and users' expectations, as discussed further in Section 2.7. A somewhat reciprocal relation exists also when QoE, or at least some aspects of it, are used to manage QoS (e.g., quality-driven network management).

If we consider a layered approach to understanding QoE, such as that proposed in [11], then we can make a more clear distinction between QoS and QoE. In the proposed approach QoE is modeled in layers, as seen in Figure 2.1.[4] We can then see how QoE is dependent on QoS (but not only on it), and how QoS does not "reach high enough" in the stack to be sufficient to understand QoE.

[2] Albeit this impact can in some cases be important indeed.

[3] Of which some of the most commonly used ones, such as PSNR in the case of video, can safely be excluded, since they bear no meaningful correlation with subjective perception.

[4] We note that this classification of the human-related aspects of QoE differs slightly from the one provided in [4] in the way some aspects are classified as "human" or "user"-related. However, both approaches consider essentially the same aspects, and are fundamentally compatible. This is further discussed in Section 2.6.

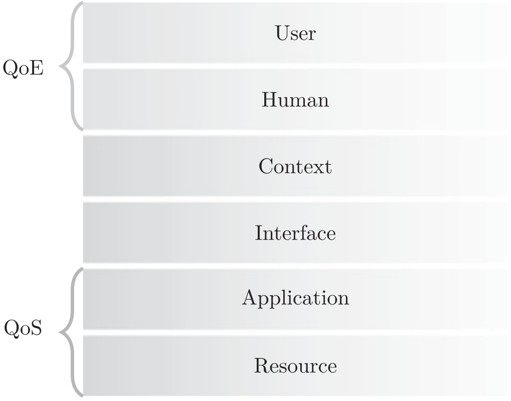


Figure 2.1 The layered QoE model proposed in [11]. QoE is about the user, and hence cannot be considered absent the physiological, psychological, social ("Human"), and role-related ("User") aspects of the user as a person, whereas QoS concerns the system, and hence it is considered at the Resource and Application layers. Incidentally, these layers can be mapped to the OSI network model

2.4 Factors Influencing QoE

A central question when evaluating and modeling QoE is the issue of what factors have a relevant influence on the quality of the end-user experience. A significant amount of past research has studied the relationships between technical QoS parameters and subjective quality perception (often reported as Mean Opinion Scores, MOS) [12–16]. However, such studies commonly neglect the impact of user-related or contextual factors on the outcome of the subjective evaluations. To give an example, two users evaluating QoE when using a Web-based service or watching a streaming video on their mobile phone under the same conditions (in terms of network and device performance) may provide significantly different QoE ratings due to the impact of their previous experiences in using such services, their expectations, their emotional state, etc. Other examples would be users' quality perceptions differing due to the different environments in which they are accessing a service (e.g., indoors, outdoors) or given different service costs (e.g., a person paying more for a service may expect a higher quality).

Building around the definition of QoE as previously cited from [4], the white paper states that

> *...in the context of communication services, QoE can be influenced by factors such as service, content, network, device, application, and context of use.*

While a number of works have addressed classification of the multitude of QoE Influencing Factors (IFs) [17–19], we highlight the three categories of possible influence factors (defined as: *Any characteristic of a user, system, service, application, or context whose actual state or setting may have influence on the Quality of Experience for the user)* as proposed in the white paper (and further elaborated in [20]) as follows:

Human IFs comprise any variant or invariant properties or characteristics related to a human user. Examples include socio-economic and demographic status, physical and mental constitution, affective state, etc. The impact of different IFs can also be considered at different perceptual levels. A further discussion of human factors is given in Section 2.6.

System IFs are factors that determine the technically produced quality of an application/service, further classified into four sub-categories, namely content-, network-, media-, and device-related IFs. Examples of content factors include video spatial and temporal resolution, and graphical design elements and semantic content in the case of websites. Media-related IFs refer to encoding, resolution, sampling rate, media synchronization, etc. Network-related factors describe the transmission system (e.g., in terms of delay, loss, throughput), while device-related factors specify the capabilities and characteristics of the systems/devices located at the end points of the communication path.

Context IFs cover a broad range of factors that identify any situational property to describe the user's environment in terms of physical (e.g., location and space), temporal (e.g., time of day), social (e.g., a user being alone or surrounded by other people), task (e.g., the purpose for which a user is using a service), technical (e.g., the presence of other surrounding technical systems), and economic characteristics (e.g., cost paid for a service) [21,22]. The importance of QoE from an economical point of view, as linked with economic pricing models, is discussed in further detail in Section 2.7.

Following this general classification, there are several points that need to be considered. Firstly, IFs may be interrelated (i.e., they should not be regarded as isolated). Drawing on a previous example, a user's expectations may be influenced by the context in which a given service is used (e.g., physical location, service cost) or the device being used to access a service. Hence, if we were to geometrically represent factors as belonging to multi-dimensional spaces (i.e., each factor representing a different dimension), the dimensions do not necessarily represent a basis of the IF space. A simple illustration is given in Figure 2.2, portraying a mapping from three multi-dimensional IF spaces to a QoE space (adapted from [19]). A point in each one of these spaces represents the corresponding state of the system (human state, system state, context state), with coordinates determined based on the values of the different factors considered. Certain constraints may be imposed to eliminate non-feasible points or regions in the factor space. We further consider a mapping from the different factor spaces to a multi-dimensional QoE space, referring to the decomposition of *overall QoE* (also referred to in the literature as *integral QoE* [23]) into multiple dimensions representing different QoE features, which in turn are not necessarily independent of each other. Hence, we can say that given a certain human, system, and context state (represented by points in the multi-dimensional spaces), a user's QoE is assessed (or estimated) in terms of multiple quality dimensions. These quality dimensions can then be further collapsed into a single, integral value of QoE (e.g., as a weighted combination of quality dimensions).

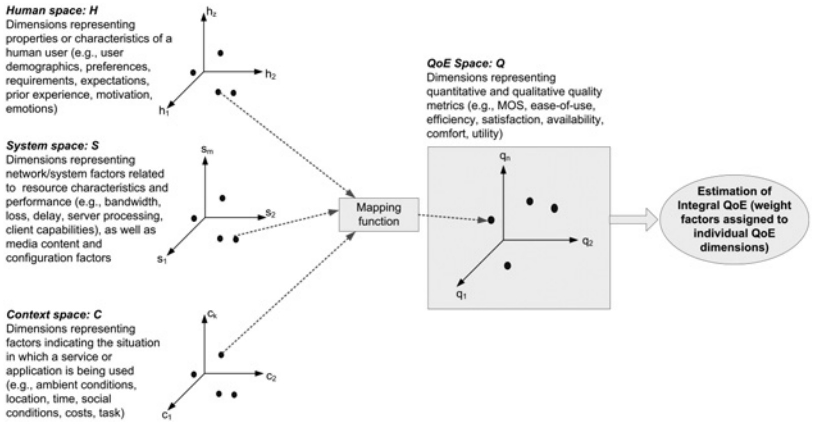

Human space: H
Dimensions representing properties or characteristics of a human user (e.g., user demographics, preferences, requirements, expectations, prior experience, motivation, emotions)
h_z
h_2
h_1
System space: S
Dimensions representing network/system factors related to resource characteristics and performance (e.g., bandwidth, loss, delay, server processing, client capabilities), as well as media content and configuration factors
s_m
s_2
s_1
Context space: C
Dimensions representing factors indicating the situation in which a service or application is being used (e.g., ambient conditions, location, time, social conditions, costs, task)
c_k
c_2
c_1
Mapping function
QoE Space: Q
Dimensions representing quantitative and qualitative quality metrics (e.g., MOS, ease-of-use, efficiency, satisfaction, availability, comfort, utility)
q_n
q_2
q_1
Estimation of Integral QoE (weight factors assigned to individual QoE dimensions)

Figure 2.2 Multi-dimensional view of QoE influence factors

The mapping to a QoE space represents different possible QoE assess-ment/estimation methods. In the case of objective quality assessment, such a function can feed relevant input parameters to existing models to obtain quality estimations. In the case of subjective quality assessment, a mapping function may correlate input parameters with user quality scores. Furthermore, regression techniques or other machine learning tools, such as neural networks, can be used to estimate QoE.

It should be noted that QoE can be estimated/measured at different points in time and in relation to different time scales. For example, QoE can be measured at a particular instant in time, following a certain period of service use (e.g., following a one-time use of a service or following multiple service uses), or over a longer period of time (e.g., a provider wishing to determine overall customer experience over a period of a few weeks). Hence, it is important to note that temporal considerations are needed when addressing QoE.

While in theory we can identify a wide scope of factors that will (likely) have an impact on QoE, challenges still remain in: (1) measuring certain factors, (2) determining their impact on QoE, and (3) utilizing the knowledge of such impacts in the context of optimizing QoE, taking into account the point of view of different actors involved in the service delivery chain. While these challenges have been widely addressed in the context of the QoS-to-QoE mapping (focusing on identifying relationships between technical network performance and user-perceived quality), the measurements and impact of human-related psychological factors have been much less studied.

Finally, specific IFs are relevant for different types of services and applications, as will be elaborated on further in the following section. For example, a user's gaming skills are very important when considering QoE for games, while a person's medical condition may directly impact the QoE when using an e-Health service. If one were to consider the impact of system parameters, network delay may prove critical for a conversational VoIP service, whereas delay will have much less impact in the case of background file transfer or sending an email.

2.5 Service QoE

In order to truly be useful, the definition of QoE needs to be made operational for the different services one wishes to work on. The wide range of multimedia services being delivered today differs vastly in their characteristics, such as the type of media content being transmitted, the service purpose, real-time constraints, tolerance to loss, and degree of interactivity (e.g., one-way or bi-directional). When considering non-media services, such as Web or Cloud services, the inter-service variation in terms of relevant IFs and user requirements in terms of quality is even more pronounced. It is clear then that the relevant QoE IFs and their impacts on perceptual features need to be considered separately for different types of services.

There are two main types of reasons for trying to understand and estimate the QoE of a given service. The first one is related to the technical workings of the service, and in this context, the sought objective is to provide the user with a certain level of quality, given a set of constraints. This gives place to a variety of service and network management mechanisms that are driven by QoE estimations. Examples of these might be QoE-based mobility management [24], where an estimation of perceived quality provides guidance as to which available network a client should connect to, or more typical access control and traffic shaping mechanisms, which can also benefit from an understanding of QoE, both for different service types (e.g., video [25], VoIP [26]) or different network environments (e.g., LTE-A [27]).

The second type of reason is related to the business aspects of the services. Operators and service providers try to understand the role of QoE in how well a service is accepted, how much users are willing to pay for it [28], and how they can avoid churn. QoE plays an important role in these questions, as it provides a good approximation of the utility of the service for the users. Both types of reason often go hand in hand, as the technical aspects can dictate for example how a certain level of quality can be achieved to keep customers satisfied, given a set of operational constraints.

The large number of new services being adopted, coupled with the increased interest in understanding their QoE, make a solid case for developing new mechanisms to assess them, both subjectively and objectively. Studies addressing the quality of media services date back to the early days of circuit-switched telephony speech and television system quality evaluation, moving on to digital media services delivered via packet switched (fixed and mobile) networks [29]. While today numerous ITU standards recommend various quality models and assessment methodologies [30], new and emerging service scenarios (e.g., 3D audiovisual services, Web and Cloud-based services, collaborative services, gaming) are driving the need for new quality metrics and models. In order to illustrate the dependency of QoE on the type of service, we will briefly summarize research and standardization activities addressing QoE assessment and prediction for different traditional and emerging types of services in the QoE domain.

2.5.1 Speech

Studies evaluating the quality of voice services have gone from considering circuit switched voice to addressing VoIP in packet switched networks [31]. Detailed guidelines have been provided for conducting subjective assessments, for example listening quality and conversational quality tests [32–34], most commonly quantifying user ratings along a 5-point MOS scale.

Listening quality for speech services has been an object of extensive study in both academic and industrial research. Several models have been developed, originally focusing on encoding and Plain Old Telephone Service (POTS) transmission aspects, and more recently incorporating IP transmission factors. The available models range from trivial measures of distance between an original and degraded signal (e.g., SNR) to very complex ones based on detailed models of the Human Auditory System (HAS) [35], to models providing a more detailed estimation of speech quality along several dimensions of quality [36]. Current models for Full-Reference (FR) assessment of speech quality – such as ITU-T P.862 [37] and its replacement P.863 [35] – can provide very accurate estimations of perceived quality for speech under a variety of conditions. No-Reference (NR) models, both signal-based [38] and parametric ones (e.g., [39]), can produce accurate enough estimations of speech quality for instance for monitoring purposes.

Objective models for conversational speech are not as abundant or accurate as the ones available for listening quality. The most widely referenced NR model is the ITU-T E-Model [40]. The E-Model is a planning model, designed for network dimensioning rather than quality estimation, and so it is not the ideal tool for this task. However, given the scarcity of available models, it is often used for this purpose in the literature. A promising approach was proposed in [41], based on the PSQA methodology [42], but it requires further work to be useful in the general case.

One of the challenging aspects of conversational quality is that it depends not only on the technical aspects of the transmission chain, but also on the dynamics of the conversation itself, which are not always easy to characterize. An indicator of "conversational temperature" was introduced in [43]. That, and similar approaches, can provide a means to incorporate the interactivity aspects of a call into the quality estimations. More recently, a promising indicator for conversational quality has been introduced in [44].

New lines of research regarding conversational quality assessment are emerging in the domain of multi-party systems (such as tele meetings and VoIP bridges). The ITU has recently released a recommendation for the subjective assessment of such systems [45].

2.5.2 Video

The assessment of video quality has been under study since shortly after television became a common service. Traditionally, subjective video assessment is carried out according to the ITU-R B.500 series of recommendations (the latest of which is described in [46]), designed for television services. For more modern multimedia services, video assessment can be done according to [47]. More recently, the combined assessment of audio and video within audiovisual services has given rise to newer ways of assessing its quality [48], which take into account the two modalities involved at the same time.

Beyond subjective assessment, there exist a large number of objective models for estimating video quality. A very comprehensive overview of the current state of the art in this domain can be found in [49]. Most of the best available video quality estimators are FR models, which compare a degraded signal to the original one, and produce an estimate of its quality. Of the large variety of FR models available in the literature, it is worth mentioning PEVQ [50] and VQuad HD [51], which represent the current standards in FR video assessment. PEVQ is designed to assess streaming video in the presence of packet losses, for a variety of resolutions ranging from QCIF to VGA, in several commonly used encodings (H.264, MPEG-4 Part 2, Windows Media 9, and Real Video 10). VQuad HD, on the contrary, aims to assess high-definition (1080p, 1080i, and 720p) streams, also considering lossy transmission scenarios and frame freezes up to 2 s in duration.

Other well-known methods for FR assessment include those based on the Structural Similarity Index (SSIM) [52, 53], related variants [54], and MOSS [55]. Singular-Value Decomposition (SVD)-based methods (cf. [56]) are also available in the literature.

For audiovisual assessment (in particular for IPTV systems), we can single out the recent ITU standardized parametric models for audiovisual quality assessment ITU-T P.1201 [57] (P.1201.1 [58] for SD content, P.1201.2 [59] for HD content), which provide very accurate estimates and are suitable for use in monitoring and control applications, as they do not require a reference in order to produce estimates.

2.5.2.1 HTTP Video

Over the past few years, video streaming sites such as YouTube, Netflix, Hulu, or Vimeo have exploded onto the scene, with video streaming representing today a very large fraction of the overall traffic of the Internet. These services, in contrast to older ones, use HTTP as their transport (as opposed to RTP/UDP or MPEG2-TS, as is commonly the case in legacy services). This has interesting implications for quality assessments, as HTTP provides a lossless transport, whereas for UDP-based streaming, losses are the single major source of artifacts and quality degradations. In contrast, HTTP-based streaming suffers from delay issues, but no actual image quality degradation during transport. When packets are lost, and the underlying TCP congestion avoidance algorithm kicks in, the throughput decreases and the playout buffer may be emptied, which results in freezes and rebuffering events. In this context, the presence, length, and distribution of these freezes is what determines the quality perceived by the users. The move from progressive download streaming toward adaptive HTTP streaming (such as MPEG-DASH, or Apple's Live Streaming) further complicates the development of good predictive models for QoE of HTTP video. Efforts in this domain are becoming more common of late (cf., e.g., [60–63]). It has been shown that in the context of HTTP-based streaming, freezes seem to dominate the perceived quality of the service [64], but trying to infer when the playout buffer suffers under-runs from observed network behavior is still an issue to solve in order to create predictive models for the quality of these types of systems.

2.5.3 Cloud and Web-Based Services

In terms of services and their QoE, speech (also audio) and video are considered "traditional" and have been, as discussed above, studied extensively. New service trends, however, make it interesting to expand our understanding of QoE to other services which are not necessarily media-oriented. Over the past few years, there has been an increasing trend of migrating services to "the Cloud." Cloud-based versions of previously locally hosted services are, by the very nature of the Cloud, bound to perform differently, and also be experienced differently by their users. It is therefore important to understand what QoE means in this context (that is, what does it take to make the definition of QoE operational for this type of service). There are several challenges in this regard, which have been described in [65]. Studies have already been conducted for several non-traditional types of service such as Cloud storage [66] and remote desktop service [67].

In a more general setting, many Cloud-based services are provided via the Web, or via mobile applications which often use Web services as their back-end. This has spurred a strong interest in understanding and modeling the QoE of Web applications and services. This is complicated by the fact that practically any type of service can be delivered via the Web, and thus Web QoE models have restrictions in their generalization capabilities. For the most part, studies in this domain have focused on the performance aspects [12, 68, 69]. It is the case, however, that performance is not the only factor influencing QoE in the Web, and other aspects such as aesthetics [70] and usability, and their interplay with performance on Web QoE, are the focus of currently ongoing efforts.

2.6 Human Factors and QoE

Earlier in this chapter, we pointed to a set of factors at the human, system, and context level, that – either independently or interlinked – may have an influence on QoE. In this section, we take a closer look at a number of psychological and social aspects that are situated at the human level and need to be considered when investigating QoE, either because of their possible influence on QoE, or as elements of QoE. Note that as in [4] we refer to "human influence factors," which go beyond those defining the "user" as a role. The underlying rationale is that there are human characteristics and properties beyond those that are linked to a person in the specific role of the "user" of a service, product, or application, that may have a strong impact on QoE. This distinction is to a major extent artificial; although a specific role is taken up when using a service or application, the person remains the same, so therefore we give a brief overview of human IFs and point to the implications and related challenges for the field of QoE. This is represented in the layered model in Figure 2.1 by the User-as-a-role layer being on top of the Human layer representing the user as a person. In this context we also underline the particular importance of affective states in the light of the aforementioned new Qualinet definition of QoE.

2.6.1 Human Influence Factors: Known unknowns or unknown knowns?

Several conceptual representations of QoE (see, e.g., [17, 22, 71–73]) that have been introduced in the literature over the last years underline the importance of human IFs and the need to better understand which factors influence QoE, in which circumstances, and how (e.g., in which order of magnitude and direction). In addition, it is often repeated that human IFs are very complex: most of them refer to abstract and intangible notions that exist – consciously or unconsciously – in the mind of a human user. As a consequence, many of these factors are invisible and black boxed, they may somehow be formalized in the mind of a human user (e.g., in the form of stories, feelings, etc.), but in many cases they are not, and their influence takes place at an unconscious and latent level. Moreover, given their inherent subjective nature, the conceptual boundaries of a range of human factors such as values, emotions, attitudes, expectations, etc. have been the subject of long – and in some cases still ongoing – debates in a range of scientific communities and fields. Even further adding to the complexity, some of these factors have a very transient and dynamic character: they evolve in often unpredictable ways over time in a time-frame that can be very short (e.g., emotions) or longer (e.g., moods, skills). Although some factors have a more stable character (e.g., attitudes), that does not necessarily make them easier to grasp. Partly this is further reinforced by the fact that human factors are in many cases strongly intertwined and correlated, either with other human factors (e.g., values and socio-cultural background) or with other influencing factors at the system or context level (e.g., affective state and social context). As a result, their possible impact may for instance be reinforced or reduced depending on the presence or absence of specific other factors. Prevalent challenges are therefore not only to define and operationalize human factors, but also to measure them, to disentangle them, and to understand their relative impact on QoE [20].

In practice, and without a doubt strongly linked to this inherent complexity, the focus on human IFs was (and is still) very often limited to the relatively easy to grasp “usual suspects” such as gender, age, visual acuity, expertise level, etc. Moreover, these aspects are usually included as something in the margin (e.g., to describe a sample population), but not as the main focus of the research or as independent variables of which the impact on QoE is thoroughly investigated. More recently, the interest in human IFs in general, and in aspects such as expectations and affective states in particular, has strongly increased. This is also reflected in the growing literature and number of studies focusing on the influence of one or more human factors on QoE, in different application domains (such as IP-based services, mobile applications and services, gaming, etc.). Nevertheless, despite this growing interest, the understanding of more complex human IFs and their possible impact on QoE is still relatively limited. This largely “uncharted territory” status may be partly due to the fact that a clear and broader picture of the wide range of human factors (next to system and context IFs) that may bear an influence on QoE was for a long time missing within the community.

2.6.2 Shedding Light on Human Influence Factors

As mentioned in Section 2.4, part of the recent community-driven efforts (e.g., within the COST Action 1003 – Qualinet) toward further conceptualizing the field of QoE have been oriented toward tackling exactly this challenge. In the following, we give a number of examples of possible influence factors, using the classification presented in [4] and further discussed in [20] as a basis. Note that this overview can and should not be considered exhaustive.

Since human perception is an essential part of QoE, a possible way to structure the multitude of human IFs is – as mentioned above – to consider them at different perceptual levels. More concretely, some human IFs are most relevant for “low-level sensory processing.” In the literature, this is also referred to as “bottom-up” or “data-based” processing, which draws on the processing of “incoming data” from the senses and their receptors [74]. Characteristics related to the physical, emotional, and mental state of the human user may bear an influence on QoE in this respect [4]. Aspects that are particularly salient at this level are, for example, visual and auditory sensitivity (and other properties of the Human Visual System, HVS and HAS). It can be argued that most of these factors have a rather constant or even dispositional character. Moreover, they can be strongly correlated to other human factors (e.g., men have a higher probability of being color blind than women). Other factors, such as for example lower-order emotions (i.e., spontaneous and uncontrollable emotional reactions [75] that may also influence QoE at this perceptual level) have a much more dynamic character.

Top-down or knowledge-based processing (in the framework of [22] this is called "high-level cognitive processing"), on the contrary, refers to how people understand, interpret, and make sense of the stimuli they perceive. This type of processing is nearly always involved in the perceptual process (albeit not always consciously) and "knowledge" is to be broadly understood as *any information that a perceiver brings to a situation* [74]. In the context of QoE – with its inherent subjective character and the importance of the associated evaluative and judgmental processes – understanding the impact of human IFs at this level is of crucial importance. Again, a distinction can be made between factors that are relatively stable and factors that have a more dynamic character. Examples of the former are the socio-cultural background (which is to a large extent linked to and determined by contextual factors) and the educational level and socio-economical position of a human user, which are often strongly linked. Both are relatively, yet not entirely, stable: depending on the life-stage, their impact on different features may be very different (e.g., the perceived fairness of the price of a service may be different for a student without a stable income than for a pensioner who is in a much more comfortable financial situation). Other examples of relatively stable factors that may influence QoE and its related evaluation and judgment processes are goals, which fuel people's motivations and correspond to underlying values. A common distinction made is the one between goals and values that are more pragmatic and instrumental (called "do-goals" in [76]) and others – situated at a higher abstraction level – that go beyond the instrumental and that refer to ultimate life goals and values ("be-goals" in [76]). The saliency of specific goals and values in the mind of a human user may strongly influence QoE and the importance attached to specific features.

Similarly, the motivations (which can differ in terms of, e.g., intensity and orientation) of a person to use a service or application may have a strong impact. For example, when the use of an application (e.g., a mobile game) is intrinsically motivated, its inherent characteristics and features (e.g., game-play, story, joy-of-use, etc.) will be most important. Extrinsically motivated behavior, in contrast, is driven by an external, separable outcome [77] (e.g., to earn money). Other examples of more dispositional or relatively stable human influence factors are personality (in the literature, reference is also made to emotional traits or personality traits in this respect), preferences, and attitudes. The latter two are particularly interesting in relation to QoE, as they are both intentional (i.e., oriented toward a specific stimulus, object, person) and contain evaluative judgments or beliefs (see also [20]). Moreover, attitudes are not only about cognition, they also have an affective and behavioral component and these three components influence each other.

Examples of human factors that have a more dynamic character are expectations, previous experiences, moods, and emotions. As discussed in Section 2.2, the ITU definition of QoE already pointed explicitly to the possible influence of "user expectations" on QoE. In the literature, different types of expectations are distinguished (see, e.g., [78]) and expectations can be based on a range of sources. For instance, they may be modified based on previous experiences, on experiences from others, on advertisements, etc. Despite their strong relevance for QoE and a number of recent studies on expectations and QoE (see, e.g., [79, 80]), the impact of implicit and explicit expectations and the way in which they interplay with or depend on other influence factors (such as economical context, physical context, device context, content, etc.) is still rather poorly understood. At the end of this chapter, the relation between user expectations, general QoE, and price is further discussed.

Finally, we mention two affective states that may have a strong impact on QoE and that – in recent years – have strongly gained importance, namely emotions and moods (see also [81]). Both are dynamic, yet there are important differences. Firstly, emotions are intentional whereas moods are not. Secondly, whereas moods have a longer duration (they can last for a couple of hours to a number of days), emotions are characterized by their highly dynamic character: they last for a number of seconds to a number of minutes. Emotions are thus much more "momentary." The growing interest in emotions and other affective states in the field of QoE is well reflected in the recent literature [82–86] and in ongoing research activities (e.g., within the scope of Qualinet). The literature on emotions in other disciplines (psychology, neurosciences, etc.), focusing the influence of emotions on other psychological functions and how they relate for instance to human perception and attention, cognitive processes and behavior, is overwhelming. In spite of this, major challenges still lie ahead for the field of QoE. Firstly, to bring this already available knowledge together and integrate it into the state of the art in the field of QoE. Secondly, to gain a better understanding of how emotions and other affective states can be measured and influence QoE, based on what is already known from other fields and empirical research. Finally, an additional challenge lies in the development of heuristics and guidelines to better account for the influence of human affect on QoE *in practice.*

2.6.3 Implications and Challenges

Referring back to what was already argued earlier in this chapter, the field of QoE has strongly evolved and shifted more toward the user perspective over the last years. However, as the above brief overview of relevant human factors and aspects already indicated: new challenges and frontiers have appeared and these are situated exactly at this level. First of all, it is clear that the involvement of researchers from disciplines and fields studying "human factors" is a necessary precondition to deepen the understanding of human factors, related psychological processes (e.g., decision making, perception, cognition, emotion), and what they mean for QoE (both in theory and in practice). The related disciplines can be framed under the umbrella of behavioral and social ("soft") sciences and include, for example, social sciences (communication science, sociology, etc.), cognitive psychology, and social psychology, and anthropology. This also implies that prevalent language barriers and epistemological differences that exist between these more human-oriented fields and those that are investigating QoE from a technical (e.g., networking) perspective need to be bridged in order to enable mutual understanding and to approach QoE from a genuine interdisciplinary perspective (see also [87]).

Secondly, stating that a range of factors at the human level may have a strong influence is not enough. Its implications at the methodological level are major, and they cannot simply be ignored. As a result, methods and measures that allow us to investigate the importance and possible impact of human factors (not only separately, but also in relation to other influence factors) need to be further explored, adopted, and embedded into the traditional "subjective testing" practices. As the latter are too limited and fail to take subjective, complex, dynamic factors into account, complementary methods, tools, and approaches (e.g., use of crowdsourcing techniques, physiological and behavioral tools, QoE research in the home environment, etc.) are currently fully being explored within the QoE community.

Thirdly, the wide range of human factors – of which we gave a brief overview – also imply that users may differ from each other in many ways. Future research should therefore seek to move beyond the notion of the "average user" and to explore approaches that allow us to take the diversity and heterogeneity of users into account. In practice this means moving more toward the smart use of segmentation approaches, which can provide valuable input for the business domain in the QoE ecosystem (e.g., in view of developing tailored charging schemes and oriented toward QoE-based differentiation).

Finally, defining QoE in terms of emotional states ("a *degree of delight or annoyance*") and pointing to the importance of the *enjoyment* next to or as opposed to the *utility* of an application or service implies that the traditional indicators of QoE – which are oriented toward the estimation of user satisfaction in relation to experienced (technical) quality and instrumental aspects – need to be reconsidered as well. The relation between traditional QoE indicators and affective states such as delight and annoyance needs to be thoroughly investigated, not only as such, but also in the presence of the identified QoE influence factors. Several recent studies have started to explore the relevance and use of alternative self-report, behavioral, and physiological measures of QoE (to gain a better insight into indicators of, e.g., engagement, delight, positive and negative affect, etc.) in this respect (see, e.g., [88, 89]). However, they represent only a small proportion of the literature, implying that in most cases QoE research is still a matter of *business as usual* and that crossing some of the identified new frontiers will inevitably require major changes, or turn out to be a just a fancy new wrapping, but fundamentally nothing new.

2.7 The Role of QoE in Communication Ecosystems

Summarizing what has been said so far, we have seen that the transition from QoS to QoE and the corresponding turn toward the user involves significant methodological implications, and thus represents a genuine paradigm change rather than a mere update from "QoS 1.0" to "QoS 2.0", especially if we take into account that QoS research for decades has been shaped by mainly studying network QoS parameter like delay, jitter, packet loss rate, etc. Instead, shifting the end-user/end-customer back to the center of our attention reminds us of a similar, however much more illustrious, paradigm change, which took place in the middle of the 16th century, when Polish-Prussian astronomer Nicolaus Copernicus put the transition from the geocentric to the heliocentric model of the universe into effect. In an analogous way, the QoE model of service quality assumes that the user and her needs are placed in the center of the technological universe (and not the other way round), and the related paradigm change has consequently been labeled an "Anti-Copernican Revolution" [6, 90].

To describe the resulting interplay between end-users, business entities, and the technological environment in more detail, Kilkki has proposed using the overarching metaphor of "Communication Ecosystems" as an analogy to the well-known concept of biological ecosystems [91]. Remember that in biology, ecosystems basically describe communities of organisms and their environment as a joint system of mutual interactions and inter dependencies. Similarly, a communication ecosystem is populated by communities of private and business users as well as further commercial entities, all of them interacting which each other using communication services offered from the technological environment. As Kilkki points out in his analysis, this approach allows us to obtain a profound understanding of the generation of expected novel products and services together with corresponding business models, based on a deep understanding of human needs and how to serve them in terms of technology and economics.

It is worth noting that the analogy between biological and communication ecosystems goes even further, especially due to the fundamental role of hierarchical structures which we will analyze in a bit more detail here. In biology, there are mainly two ways of interaction within an ecosystem: in a horizontal perspective, individuals and communities either compete or cooperate with each other, whereas from a vertical point of view, the main types of interaction are "eating" and "being eaten," which are reflected in concepts like food chains or ecological pyramids. In this context, we distinguish producers (e.g., plants or algae) from consumers (i.e., animals like herbivores or carnivores) and decomposers (e.g., bacteriae or funghi). Moreover, producers and consumers are ordered along the food chain into five trophic layers, starting from plants over primary (herbivores eating plants), secondary (carnivores eating herbivores), and tertiary (carnivores eating carnivores) consumers up to apex predators which form the top of the food chain and do not have further natural enemies.

With communication ecosystems, we can easily identify comparable hierarchical structures on different levels, for instance:

- **ISO/OSI model.** In the basic reference model for open systems interconnection, [ISO/IEC 7498-1] distinguishes seven different layers of a communication system – the (1) physical, (2) data link, (3) network, (4) transport, (5) session, (6) presentation, and (7) application layers. Each layer is assumed to use services offered from the layers below to serve the needs of the layers above.
- **Internet protocols.** To fulfill the tasks attributed to the various layers of the OSI model, a plethora of communication protocols have been specified, which may compete against each other on a single layer; for instance on the transport layer, UDP has been designed for real-time communication and offers unreliable but efficient transport, while TCP provides guarantees against packet loss but can incur additional delays due to re-transmission and rate adaptation. Overall, protocols from a lower layer offer services to protocols from higher layers, which in the case of the Internet has led to the notorious "hourglass model" underlining the specific role of IP.
- **Communication services value chains.** Starting from physical network resources (like fiber networks) owned by infrastructure providers, and used by network service providers to offer Internet access and connectivity. Based on this, application service providers are able to offer specific services and applications, while content providers and OTTs (over-the-top providers) use the underlying structures to publish their content to the end-customer. Again, within one layer, providers may either compete against each other for joint market shares and/or cooperate, for instance in guaranteeing end-to-end inter-carrier services.

With communication ecosystems, modeling cooperation and competition between the various entities and communities involved is significantly facilitated by employing methods and tools from cooperative and non-cooperative game theory as well as the application of related

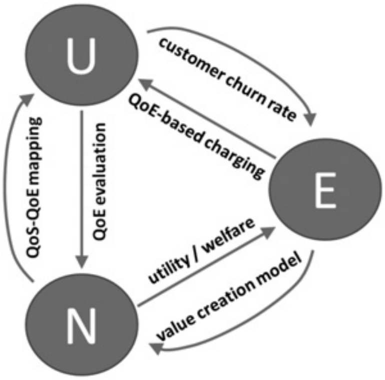


Figure 2.3 The triangle model for ecosystem quality

micro-economic concepts which altogether allow an in-depth understanding of the interplay between the different stakeholders and their individual interests. For instance, determining the value of a resource or service provided to an individual customer is not only one of the important issues within QoE research, but at the same time has traditionally played a key role in micro-economics, especially in the context of utility theory. Here, the so-called utility function is used to model user preferences over certain sets of goods and services, which allows us to derive indifference prices as well as maximize individual and social welfare in an efficient way.

Hence, research on communication ecosystems turns out to be an intrinsically interdisciplinary endeavor, joining methods and approaches from communication technology and user-centered research with micro-economics and a variety of further social sciences. Of course, in order to enable a methodologically consistent approach, we have to suppose – as mentioned earlier in this chapter – a certain degree of common language for formal description as an additional requirement. In fact, this highlights the pivotal position of service quality in our context: while the notion of quality serves as a key concept for the technological as well as the economical and the user side, respectively, and while the primary understanding of this concept may still be slightly different for all three perspectives (technology: QoS, economics: utility, user: QoE), we have, however, sufficient tools at hand to unify these approaches into a joint framework, as exemplified in Figure 2.3.

The resulting triangle model describing the close interplay between technological, economical, and user-centric forces may be summarized briefly as follows.

- **User:** QoE with a communication service is influenced by two main dimensions (i.e., the service quality as offered by the network and the tariff charged for this quality). While the former may be described in terms of a QoS-to-QoE mapping and is the focus of most of the ongoing QoE research, the latter is determined by corresponding economic pricing models. As a result, the user communicates his/her QoE evaluation (e.g., in the form of a MOS) as feedback into the system, while on a longer time scale, the quality experienced by the user is of fundamental importance to the customer churn rate.

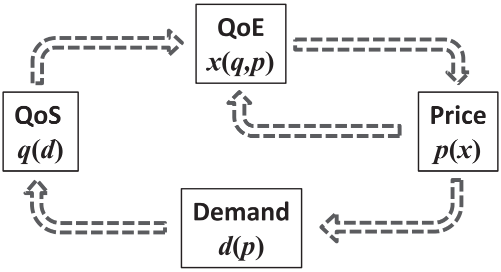


Figure 2.4 Model for QoE-based charging. *Source:* Reichl *et al.,* 2013 [28]

- **Network:** Network service providers, application service providers, and content providers cooperate along the value chain, according to business models which arrive as input from the economic side. As a result, a certain level of network QoS is delivered toward the user, which is assumed to satisfy user expectations and at the same time, aggregated over all customers, maximize social welfare in the system.
- **Economics:** Utility functions, which formally describe the value of technical resources to the user, are fundamental for deriving appropriate dynamic prices to be paid by the customer. At the same time, the customer churn rate serves as key input for confirming the chosen tariff structure as well as, again on a longer time scale, the underlying business and value-creation models.

Note that this triangle model comprises at the same time two different time/service granularities. On the one hand, the QoE evaluation by the user is very application- and context-specific and subject to continuous change over time, while the resulting maximization of social welfare as well as the corresponding dynamic pricing schemes provide direct reactions to the current state. On the other hand, customer churn rate and business models vary on a longer time scale, similar to the fundamental laws governing the relationship between network-centric QoS and user-centric QoE.

In the remainder of this section, our focus will be directed toward further analyzing the question of how to charge for QoE, which is intimately linked with the user perspective described above. To this end, Figure 2.4 depicts a simple feedback model for QoE-based charging as described in [28].

In this model, we analyze a system with limited overall resources, which is therefore prone to congestion depending on the total demand created by the users. Hence, the system may be described by the following four functions.

- **QoS function**: $q = q(d)$
- **Demand function:** $d = d(p)$
- **Price function:** $p = p(x)$
- **QoE function:** $x = x(q, p)$

Here, the mentioned relation between network QoS and overall demand is captured by the QoS function $q(d)$, while the demand essentially depends on the price charged to the customer, according to the demand function $d(p)$. Moreover, the level of network QoS delivered to the user is directly influencing his/her experience, hence the user's QoE function $x = x(q, .)$ depends on QoS as one input parameter.

The relationship between QoE and price is less trivial, and indeed forms the core of this model. Note that, in fact, prices act in an interesting dual role: on the one hand, the price to be paid for service quality depends on the delivered quality level (the better the quality, the more expensive the service); on the other hand, the price also has a direct impact on user expectations (the more expensive a service, the better it has to be) and, subsequently, also on the QoE evaluation (if customer expectations are high, the resulting experience is by definition lower than if prices and thus expectations are low from the beginning).

As a consequence, the price function $p = p(x)$ describes the dependency of the price on the service quality, while the price p serves at the same time as (another) input parameter to the QoE function $x(q, p)$. While, of course, this QoE function may additionally depend on further user context parameters, for the sake of clarity we restrict ourselves to the two-dimensional version denoted above.

Note that we may easily determine the marginal behavior of $x(q, p)$ for $q = 1$ and $p = 0$, respectively, as follows. If the price is kept constant at zero, this corresponds to a service delivery which is for free throughout. In this case, $x(q, p = 0)$ reflects the fundamental mapping between QoS and QoE, which for certain scenarios has already been captured in terms of fundamental laws for QoE (like, for instance, the IQX hypothesis [14] or logarithmic laws of QoE [13]). In contrast, keeping QoS constant at its maximal level allows us to express the impact of user expectations on QoE evaluation: if we assume that getting the best quality for free results in maximal QoE, then increasing the price should lead to a monotonically decreasing QoE function, whose shape, however, is largely unknown (in [28], linearity has been assumed, but also convex or sigmoid shapes might be possible). Finally, if either QoS is zero, or the price is infinitely high, then the QoE function will become zero. Altogether, a simple form for the QoE function which is sufficient for these marginal restrictions is provided by separating the impact of network QoS and user expectations and assuming a product form for QoE.

- **Separable QoE function:** $x = x(q, p) = x_{\mathrm{Q}}(q) \cdot x_E(p)$

Here x_Q refers to the QoS dimension, while x_E takes the impact of prices into sole account.

In [28] the resulting system of equations has been analyzed mathematically. As the main result, it turns out that under some rather general assumptions concerning the existence and sign of the second derivatives of the four mentioned functions, the system possesses a trivial non-stable fixed point where services are offered in minimal quality, but for free. In addition, the system possesses also a non-trivial, stable fixed point which forms a Nash equilibrium and is determined by balancing the trade-off between QoE and price to be charged.

Moreover, the existence of this equilibrium has been subject to extensive user trials that are also described in [28] and the references cited therein. For instance, a user study has been performed comprising more than 40 test subjects, who have been asked to choose arbitrarily between 20 quality/price levels of three short video-on-demand movies during a 5-minute free trial period. In order to achieve a realistic result, the test subjects have been given real money (10 euros in cash) at the beginning of the trial, which they could freely spend on improving the video quality (or take home afterwards).

As main results of this trial we can confirm that quality matters for the users, as more than 90% of the trial subjects decided to invest some of their (real) money in improving their QoE, and also the opportunity to test the various QoE/price levels in the initial 5 minutes of each movie was used extensively (some users with more than 80 level changes, i.e., one or more changes every 4 s!). The final fixed point has usually (i.e., in more than 80% of cases) been achieved according to a decaying oscillation pattern with different amplitudes and/or decay factors; note that a simple classification algorithm allows for a successful automatic classification of the user convergence behavior in almost all cases, see [28] for further details.

2.8 Conclusions

In this chapter we have done a "full-stack tour" of QoE, starting from its definitions, its relations to QoS, the factors that affect it, and how they relate to different services, its relation to human factors, and its role in communication ecosystems. The first conclusion to take away is that QoE is a complex concept, and that it lies at the junction of several, mostly unrelated, scientific, technical, and human disciplines.

It is clear that research in the QoE domain is evolving steadily toward the user. This poses conceptual and practical difficulties, as outlined in Section 2.6, but is a necessary step to take if QoE is to establish itself as a mature field of study. QoE is, after all, all about the user!

That is not to say, of course, that the technical aspects are to be left aside. As seen in Sections 2.4, and 2.5, the technical factors and services under consideration play a significant role in how quality is perceived by the users.

Finally, we have also considered the economic impact of QoE in future communications ecosystems. It seems clear that QoE will play an important part in the economy of Internet services in the near future, hence understanding it properly is key to developing successful business models and practices.

Acknowledgments

K. De Moor's work was carried out during the tenure of an ERCIM "Alain Bensoussan" Fellowship Programme and received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 246016.

M. Varela's work was partially supported by Tekes, the Finnish Funding Agency for Technology and Innovation, in the context of the CELTIC QuEEN project.

Additional support from the EU Marie Curie program in the framework of the RBUCE WEST International Research Chair NICE (Network-based Information and Communication Ecosystems) at Universite Europeenne de Bretagne/Telecom Bretagne is gratefully acknowledged.

References

[1] ITU-T. Recommendation P.10/G.100 – Vocabulary for Performance and Quality of Service — Amendment 1: New Appendix I – Definition of Quality of Experience (QoE). 2007.

[2] Fiedler, M., Kilkki, K., and Reichl, P., 'Executive Summary – From Quality of Service to Quality of Experience.' In Fiedler, M., Kilkki, K., and Reichl, P. (eds), *From Quality of Service to Quality of Experience*. Dagstuhl Seminar Proceedings 09192, Dagstuhl, Germany, 2009.

[3] Möller, S., *Quality Engineering – Qualitat kommunikationstechnischer Systeme.* Springer, Berlin, 2010.

[4] Le Callet, P., Möller, S., and Perkis, A. (eds), Qualinet White Paper on Definitions of Quality of Experience, Lausanne, Switzerland, June 2012.

[5] Crowcroft, J., *et al.*, 'QoS downfall: At the bottom, or not at all!' Proceedings of the ACM SIGCOMM Workshop on Revisiting IP QoS: What have we learned, why do we care? ACM, 2003, pp. 109–114.

[6] Reichl, P., 'Quality of experience in convergent communication ecosystems.' In Lugmayr, C.D.Z.A. and Lowe, G.F. (eds), *Convergent Divergence? Cross-Disciplinary Viewpoint on Media Convergence*. Springer, Berlin, 2014.

[7] ITU-T. Recommendation E.800 – Definitions of Terms Related to Quality of Service. 2008.

[8] ETSI. ETR 003 – Network Aspects (NA); General aspects of Quality of Service (QoS) and Network Performance (NP). 1994.

[9] Crawley, E., *et al.*, 'A framework for QoS-based routing in the Internet.' IETF RFC 2386, 1998.

[10] IETF Network Working Group. RFC 2475 – An Architecture for Differentiated Services. 1998.

[11] Guyard, F., *et al.*, 'Quality of experience estimators in networks.' In Mellouk, A. and Cuadra, A. (eds), *Quality of Experience Engineering for Customer Added Value Services: From Evaluation to Monitoring*. Iste/John Wiley & Sons, New York, 2014.

[12] Egger, S, *et al.*, 'Time is bandwidth? Narrowing the gap between subjective time perception and quality of experience.' 2012 IEEE International Conference on Communications (ICC 2012), Ottawa, Canada, June 2012.

[13] Reichl, P., *et al.*, 'The logarithmic nature of QoE and the role of the Weber–Fechner Law in QoE assessment.' ICC 2010, pp. 1–5.

[14] Fiedler, M., Hoßfeld, T., and Tran-Gia, P., 'A generic quantitative relationship between quality of experience and quality of service.' IEEE Network Special Issue on Improving QoE for Network Services, June 2010.

[15] Fiedler, M. and Hoßfeld, T., 'Quality of experience-related differential equations and provisioning-delivery hysteresis.' 21st ITC Specialist Seminar on Multimedia Applications – Traffic, Performance and QoE, Phoenix Seagaia Resort, Miyazaki, Japan, March 2010.

[16] Shaikh, J., Fiedler, M., and Collange, D., 'Quality of experience from user and network perspectives.' *Annals of Telecommunications*, **65**(1&2), 2010, 47–57.

[17] Möller, S., *et al.*, 'A taxonomy of quality of service and quality of experience of multimodal human–machine interaction.' International Workshop on Quality of Multimedia Experience (QoMEx), July 2009, pp. 7–12.

[18] Stankiewicz, R. and Jajszczyk, A., 'A survey of QoE assurance in converged networks.' *Computer Networks*, **55**, 2011, 1459–1473.

[19] Skorin-Kapov, L. and Varela, M., 'A multidimensional view of QoE: The ARCU model.' Proceedings of the 35th International Convention MIPRO, Opatija, Croatia, May 2012, pp. 662–666.

[20] Reiter, U., *et al.*, 'Factors influencing quality of experience.' In Möller, S. and Raake, A. (eds), *Quality of Experience: Advanced Concepts, Applications and Methods*. Springer, Berlin, 2014.

[21] Jumisko-Pyykkö, S. and Vainio, T., 'Framing the context of use for mobile HCI.' *International Journal of Mobile Human–Computer Interaction*, **2**(4), 2010, 1–28.

[22] Jumisko- Pyykkö, S. 'User-centered quality of experience and its evaluation methods for mobile television.' PhD thesis, Tampere University of Technology, Finland, 2011.

[23] Raake, A., 'Short- and long-term packet loss behavior: Towards speech quality prediction for arbitrary loss distributions.' *IEEE Transactions on Audio, Speech, and Language Processing*, **14**(6), 2006, 1957–1968.

[24] Varela, M. and Laulajainen, J.-P., 'QoE-driven mobility management – integrating the users' quality perception into network-level decision making.' 2011 Third International Workshop on Quality of Multimedia Experience (QoMEX), 2011, pp. 19–24.

[25] Seppänen, J. and Varela, M., 'QoE-driven network management for real-time over-the-top multimedia services.' IEEE Wireless Communications and Networking Conference 2013, Shanghai, China, April 2013.

[26] Fajardo, J.-O., *et al.*, 'QoE-driven dynamic management proposals for 3G VoIP services.' *Computer Communications*, **33**(14), 2010, 1707–1724.

[27] Gomez, G., *et al.*, 'Towards a QoE-driven resource control in LTE and LTE-A networks.' *Journal of Computer Networks and Communications*, 2013, **2013**, 1–15.

[28] Reichl, P., *et al.*, 'A fixed-point model for QoE-based charging.' Proceedings of the 2013 ACM SIGCOMM Workshop on Future Human-Centric Multimedia Networking (FhMN '13), Hong Kong, pp. 33–38.

[29] Raake, A., *et al.*, 'IP-based mobile and fixed network audiovisual media services.' *IEEE Signal Processing Magazine*, **28**(6), 2011, 68–79.

[30] ITU-T. Recommendation G.1011 – Reference Guide to Quality of Experience Assessment Methodologies. 2010.

[31] Raake, A., *Speech Quality of VoIP: Assessment and Prediction.* John Wiley & Sons, New York, 2006.

[32] ITU-T. Recommendation P.800 – Methods for Subjective Determination of Transmission Quality. 1996.

[33] ITU-T. Recommendation P.920 – Interactive Test Methods for Audiovisual Communications. 2000.

[34] Möller, S., *et al.*, 'Speech quality estimation: Models and trends.' *IEEE Signal Processing Magazine*, **28**(6), 2011, 18–28.

[35] ITU-T. Recommendation P.863 – Perceptual Objective Listening Quality Assessment. 2001.

[36] Möller, S. and Heute, U., 'Dimension-based diagnostic prediction of speech quality.' ITG Conference on Speech Communication, 2012, pp. 1–4.

[37] ITU-T. Recommendation P.862 – Perceptual Evaluation of Speech Quality (PESQ), an Objective Method for End-To-End Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs. 2001.

[38] ITU-T. Recommendation P.563 – Single Ended Method for Objective Speech Quality Assessment in Narrow-Band Telephony Applications. 2004.

[39] Rubino, G., Varela, M., and Mohamed, S., 'Performance evaluation of real-time speech through a packet network: A random neural networks-based approach.' *Performance Evaluation*, **57**(2), 2004, 141–162.

[40] ITU-T. Recommendation G.107 – The E-model: A Computational Model for Use in Transmission Planning. 2011.

[41] da Silva, A.C., *et al.*, 'Quality assessment of interactive real time voice applications.' *Computer Networks*, **52**, 2008, 1179–1192.

[42] Varela, M., 'Pseudo-subjective quality assessment of multimedia streams and its applications in control.' PhD thesis, INRIA/IRISA, Rennes, France, 2005.

[43] Hammer, F. and Reichl, P., 'Hot discussions and frosty dialogues: Towards a temperature metric for conversational interactivity.' 8th International Conference on Spoken Language Processing (ICSLP/INTERSPEECH 2004), Jeju Island, Korea, October 2004.

[44] Holub, J., *et al.*, 'Management conversational quality predictor.' Proceedings of PQS 2013, Vienna, Austria, September 2013.

[45] ITU-T. Recommendation P. 1301 – Subjective Quality Evaluation of Audio and Audiovisual Multiparty Telemeetings. 2012.

[46] ITU-R. Recommendation BT. 500-13 – Methodology for the Subjective Assessment of the Quality of Television Pictures. 2012.

[47] ITU-T. Recommendation P. 910 – Subjective Video Quality Assessment Methods for Multimedia Applications. 2008.

[48] ITU-T. Recommendation P.911 – Subjective Audiovisual Quality Assessment Methods for Multimedia Applications. 1998.

[49] Chikkerur, S., *et al.*, 'Objective video quality assessment methods: A classification, review, and performance comparison.' *IEEE Transactions on Broadcasting*, **57**(2), 2011, 165–182.

[50] ITU-T. Recommendation J.247 – Objective Perceptual Multimedia Video Quality Measurement In The Presence Of A Full Reference. 2008.

[51] ITU-T. Recommendation J.341 – Objective Perceptual Multimedia Video Quality Measurement Of HDTV For Digital Cable Television In The Presence Of A Full Reference. 2008.

[52] Lu, L., *et al.*, 'Full-reference video quality assessment considering structural distortion and no-reference quality evaluation of MPEG video.' ICME (1), IEEE, 2002, pp. 61–64.

[53] Wang, Z., Lu, L., and Bovik, A.C., 'Video quality assessment based on structural distortion measurement.' *Signal Processing: Image Communication*, **19**(2), 2004, 121–132.

[54] Moorthy, A.K. and Bovik, A.C., 'Efficient motion weighted spatio-temporal video SSIM index.' Human Vision and Electronic Imaging, 2010, p. 75271.

[55] Wang, Z., Simoncelli, E.P., and Bovik, A.C., 'Multi-scale structural similarity for image quality assessment.' Proceedings of IEEE Asilomar Conference on Signals, Systems, and Computers (Asilo3), 2003, pp. 1398–1402.

[56] Tao, P. and Eskicioglu, A.M., 'Video quality assessment using M-SVD.' Image Quality and System Performance IV, 2007.

[57] ITU-T. Recommendation P. 1201 – Parametric Non-Intrusive Assessment Of Audiovisual Media Streaming Quality. 2012.

[58] ITU-T. Recommendation P.1201.1 – Parametric Non-Intrusive Assessment Of Audiovisual Media Streaming Quality – Lower Resolution Application Area. 2012.

[59] ITU-T. Recommendation P.1201.2 – Parametric Non-Intrusive Assessment Of Audiovisual Media Streaming Quality – Higher Resolution Application Area. 2012.

[60] Oyman, O. and Singh, S., 'Quality of experience for HTTP adaptive streaming services.' *IEEE Communications Magazine*, **50**(4), 2012, 20–27.

[61] Alberti, C., *et al.*, 'Automated QoE evaluation of dynamic adaptive streaming over HTTP.' Fifth International Workshop on Quality of Multimedia Experience (QoMEX), 2013.

[62] Stockhammer, T., 'Dynamic adaptive streaming over HTTP – standards and design principles.' Proceedings of the Second Annual ACM Conference on Multimedia Systems (MMSys '11), San Jose, CA, 2011, pp. 133–144.

[63] Riiser, H., *et al.*, 'A comparison of quality scheduling in commercial adaptive HTTP streaming solutions on a 3G network.' Proceedings of the 4th Workshop on Mobile Video (MoVid '12), New York, 2012, pp. 25–30.

[64] Hoßfeld, T., *et al.*, 'Initial delay vs. interruptions: Between the devil and the deep blue sea.' Fourth International Workshop on Quality of Multimedia Experience (QoMEX), 2012, pp. 1–6.

[65] Hoßfeld, T., *et al.*, 'Challenges of QoE management for Cloud applications.' *IEEE Communications Magazine*, **50**(4), 2012, 28–36.

[66] Amrehn, P., *et al.*, 'Need for speed? On quality of experience for file storage services.' 4th International Workshop on Perceptual Quality of Systems (PQS), 2013.

[67] Casas, P., *et al.*, 'Quality of experience in remote virtual desktop services.' IFIP/IEEE International Symposium on Integrated Network Management (IM 2013), 2013, pp. 1352–1357.

[68] Strohmeier, D., Jumisko-Pyykkö, S., and Raake, A., 'Towards task-dependent evaluation of Web-QoE: Free exploration vs. "Who ate what?"' IEEE Globecom, Anaheim, CA, December 2012.

[69] Selvidge, P.R., Chaparro, B.S., and Bender, G.T., 'The world wide wait: Effects of delays on user performance.' *International Journal of Industrial Ergonomics*, **29**(1), 2002, 15–20.

[70] Varela, M., *et al.*, 'Towards an understanding of visual appeal in website design.' Fifth International Workshop on Quality of Multimedia Experience (QoMEX), 2013, pp. 70–75.

[71] Geerts, D., *et al.*, 'Linking an integrated framework with appropriate methods for measuring QoE.' Second International Workshop on Quality of Multimedia Experience (QoMEX), 2010, pp. 158–163.

[72] Wechsung, I., *et al.*, 'Measuring the quality of service and quality of experience of multimodal human–machine interaction.' *Journal on Multimodal User Interfaces*, **6**(1&2), 2012, 73–85.

[73] Laghari, K., Crespi, N., and Connelly, K., 'Toward total quality of experience: A QoE model in a communication ecosystem.' *IEEE Communications Magazine*, **50**(4), 2012, 58–65.

[74] Goldstein, E.B., *Sensation and Perception*, 6th edn. Wadsworth-Thomson Learning, Pacific Grove, CA, 2002.

[75] Zajonc, R., 'Feeling and thinking: Preferences need no inferences.' *American Psychologist*, **35**, 1980, 151–175.

[76] Hassenzahl, M., 'User experience (UX): Towards an experiential perspective on product quality.' Proceedings of the 20th International Conference of the Association Francophone d'Interaction Homme–Machine (IHM '08), Metz, France, 2008, pp. 11–15.

[77] Ryan, R. and Deci, E., 'Intrinsic and extrinsic motivations: Classic definitions and new directions.' *Contemporary Educational Psychology*, **25**(1), 2000, 54–67.

[78] Higgs, B., Polonsky, M., and Hollick, M., 'Measuring expectations: Forecast vs. ideal expectations. Does it really matter?' *Journal of Retailing and Consumer Services*, **12**(1), 2005, 49–64.

[79] Sackl, A., *et al.*, 'Wireless vs. wireline shootout: How user expectations influence quality of experience.' Fourth International Workshop on Quality of Multimedia Experience (QoMEX), 2012, pp. 148–149.

[80] Sackl, A. and Schatz, R., 'Evaluating the impact of expectations on end-user quality perception.' Fourth International Workshop on Perceptual Quality of Systems (PQS), 2013.

[81] Frijda, N., 'Varieties of affect: Emotions and episodes, moods, and sentiments.' In Ekman, P. and Davidson, R. (eds), *The Nature of Emotions: Fundamental questions.* Oxford University Press, New York, 1994, pp. 59–67.

[82] Reiter, U. and De Moor, K., ‘Content categorization based on implicit and explicit user feedback: Combining self-reports with EEG emotional state analysis.’ Fourth International Workshop on Quality of Multimedia Experience (QoMEX), 2012.

[83] Schleicher, R. and Antons, J.-N., ‘Evoking emotions and evaluating emotional impact.’ In Möller, S. and Raake, A. (eds), *Quality of Experience: Advanced Concepts, Applications and Methods.* Springer, Berlin, 2014.

[84] Antons, J.-N., *et al.*, ‘Brain activity correlates of quality of experience.’ In Möller, S. and Raake, A. (eds), *Quality of Experience: Advanced Concepts, Applications and Methods.* Springer, Berlin, 2014.

[85] Arndt, S., *et al.*, ‘Subjective quality ratings and physiological correlates of synthesized speech.’ Fifth International Workshop on Quality of Multimedia Experience (QoMEX), 2013, pp. 152–157.

[86] Arndt, S., *et al.*, ‘Perception of low-quality videos analyzed by means of electroencephalography.’ Fourth International Workshop on Quality of Multimedia Experience (QoMEX), 2012, pp. 284–289.

[87] De Moor, K., ‘Are engineers from Mars and users from Venus? Bridging gaps in quality of experience research: Reflections on and experiences from an interdisciplinary journey.’ PhD thesis, Ghent University, 2012.

[88] Antons, J., Arndt, S., and Schleicher, R., ‘Effect of questionnaire order on ratings of perceived quality and experienced affect.’ Fourth International Workshop on Perceptual Quality of Systems (PQS), 2013.

[89] De Moor, K., *et al.*, ‘Evaluating QoE by means of traditional and alternative measures: Results from an exploratory living room lab study on IPTV.’ Fourth International Workshop on Perceptual Quality of Systems (PQS), 2013.

[90] Reichl, P., ‘It’s the ecosystem, stupid: Lessons from an anti-Copernican revolution of user-centric service quality in telecommunications. 6th International Conference on Developments in e-Systems Engineering (DeSE-13), December 2013.

[91] Kilkki, K., *An Introduction to Communication Ecosystems*. CreateSpace Independent Publishing Platform, Helsinki, 2012.

Acronyms

ETSI	European Technical Standards Institute
HTTP	Hypertext Transfer Protocol
IETF	Internet Engineering Task Force
IF	Influencing Factor
IP	Internet Protocol
IPTV	IP Television
ITU	International Telecommunications Union
LTE-A	Long Term Evolution – Advanced
MOS	Mean Opinion Score
MPEG	Motion Picture Experts Group
MPEG-DASH	MPEG Dynamic Adaptive Streaming over HTTP
OSI	Open Systems Interconnection Model
OTT	Over The Top
POTS	Plain Old Telephone System
PSQA	Pseudo-Subjective Quality Assessment
QoE	Quality of Experience
QoS	Quality of Service
RTP	Real-time Transmission Protocol
SNR	Signal-to-Noise Ratio
TCP	Transmission Control Protocol
UDP	User Datagram Protocol
VGA	Video Graphics Array
VoIP	Voice over IP

3

Review of Existing Objective QoE Methodologies

Yuming Fang[1], Weisi Lin[1] and Stefan Winkler[2]
[1]*Nanyang Technological University, Singapore*
[2]*Advanced Digital Sciences Center (ADSC), Singapore*

3.1 Overview

Quality evaluation for multimedia content is a basic and challenging problem in the field of multimedia processing, as well as various practical applications such as process evaluation, implementation, optimization, testing, and monitoring. Generally, the quality of multimedia content is affected by various factors such as acquisition, processing, compression, transmission, output interface, decoding, and other systems [1–3]. The perceived quality of impaired multimedia content depends on various factors: the individual interests, quality expectations, and viewing experiences of the user; output interface type and properties; and so on [2–5].

Since the Human Visual System (HVS) and the Human Auditory System (HAS) are the ultimate receiver and interpreter of the content, subjective measurement represents the most accurate method and thus serves as the benchmark for objective quality assessment [2–4]. Subjective experiments require a number of subjects to watch and/or listen to the test material and rate its quality. The Mean Opinion Score (MOS) is used for the average rating over all subjects for each piece of multimedia content. A detailed discussion of subjective measurements can be found in Chapter 6. Although subjective experiments are accurate for the quality evaluation of multimedia content, they suffer from certain important drawbacks and limitations – they are time consuming, laborious, expensive, and so on [2]. Therefore, many objective metrics have been proposed to evaluate the quality of multimedia content in past decades. Objective metrics try to approximate human perceptions of multimedia quality. Compared with subjective viewing results, objective metrics are advantageous in terms of repeatability and scalability.

Multimedia Quality of Experience (QoE): Current Status and Future Requirements, First Edition.
Edited by Chang Wen Chen, Periklis Chatzimisios, Tasos Dagiuklas and Luigi Atzori.

Objective quality evaluation methods can be classified into two broad types: signal fidelity metrics and perceptual quality metrics [2]. Signal fidelity metrics evaluate the quality of the distorted signal by comparing it with the reference without considering the signal content type, while perceptual quality metrics take the signal properties into consideration together with the characteristics of the HVS (for image and video content) or HAS (for audio signals). Signal fidelity metrics include traditional objective quality assessment methods such as MAE (Mean Absolute Error), MSE (Mean Square Error), SNR (Signal-to-Noise Ratio), PSNR (Peak SNR), or one of their relatives [6]. These traditional objective metrics are widely accepted in the research community for several reasons: they are well defined, and their formulas are simple and easy to understand and implement. From a mathematical point of view, minimizing MSE is also well understood.

Although signal fidelity metrics are widely used to measure the quality of signals, they generally are poor predictors of perceived quality with non-additive noise distortions [7, 8]. They only have an approximate relationship with the quality perceived by human observers, since they are mainly based on byte-by-byte comparison without considering what each byte represents [3,4]. Signal fidelity metrics essentially ignore the spatial and temporal relationship in the content. It is well accepted that signal fidelity metrics do not align well with human perceptions of multimedia content for the following reasons [2, 3, 6, 9–11]:

1. Not every change in multimedia content is noticeable.
2. Not every region in multimedia content receives the same attention level.
3. Not every change yields the same extent of perceptual effect with the same magnitude of change.

To overcome the problems of signal fidelity metrics, a significant amount of effort has been spent trying to design more logical, economical, and user-oriented perceptual quality metrics [3, 5, 11–18]. In spite of the recent progress in related fields, objective evaluation of signal quality in line with human perceptions is still a long and difficult odyssey [3, 5, 12–16] due to the complex multidisciplinary nature of the problem (related to physiology, psychology, vision research, audio/speech research, and computer science), the limited understanding of human perceptions, and the diverse scope of applications and requirements. However, with proper modeling of major underlying physiological and psychological phenomena, it is now possible to develop better-quality metrics to replace signal fidelity metrics, starting with various specific practical situations.

This chapter is organized as follows. Section 3.2 provides an introduction to the quality metric taxonomy. In Section 3.3, the basic computational modules for perceptional quality metrics are given. Sections 3.4 and 3.5 introduce the existing quality metrics for images and video, respectively. Quality metrics for audio/speech are described in detail in Section 3.6. Section 3.7 presents joint audiovisual quality metrics. The final section concludes.

3.2 Quality Metric Taxonomy

Existing quality metrics can be classified according to different criteria, as depicted in Fig. 3.1. There are basically two categories of perceptual metrics [5], relying on a perception-based approach or a signal-driven approach. For the first category [19–22], objective metrics are

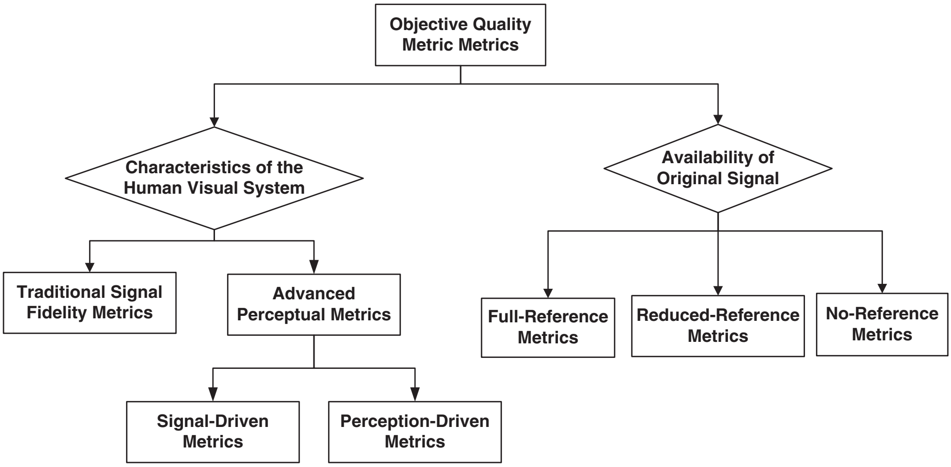

Objective Quality Metric Metrics
Characteristics of the Human Visual System
Availability of Original Signal
Traditional Signal Fidelity Metrics
Advanced Perceptual Metrics
Signal-Driven Metrics
Perception-Driven Metrics
Full-Reference Metrics
Reduced-Reference Metrics
No-Reference Metrics

Figure 3.1 The quality metric taxonomy

built upon relevant psychophysical properties and physiological knowledge of the HVS or HAS, while the signal-driven approach evaluates the quality of the signal from the aspect of signal extraction and analysis. Among the psychophysical properties and physiological knowledge used in perception-based approaches, the Contrast Sensitivity Function (CSF) models the HVS's sensitivity toward signal contrast with spatial frequencies and temporal motion velocities, and exhibits a parabola-like curve with increasing spatial and temporal frequencies, respectively; luminance adaptation refers to the noticeable luminance contrast as a function of background luminance; visual masking is usually the increase in HVS contrast threshold for visual content in the presence of another one, and can be divided into intra-channel masking by the visual content itself and inter-channel masking by visual content with different frequencies and orientations [1,2, 4,5]. For audio/speech signals, the commonly used psychophysical properties or physiological knowledge of the HAV include the effects of the outer and middle ear, simultaneous masking, forward and backward temporal masking, etc. [3].

Since the perception-based approaches involve high computational complexity and there are difficulties in bridging the gap between vision research and the requirements of engineering modeling, more recent research efforts have been directed at signal-driven perceptual quality metrics. Compared with perception-based approaches, signal-driven ones do not need to model human perception characteristics. Instead, signal-driven approaches attempt to evaluate quality from aspects of signal extraction and analysis, such as statistical features [23], structural similarity [24], luminance/color distortion [25], and common artifacts [26,27]. These metrics also consider the effects of human perception by content and distortion analysis, instead of fundamental bottom-up perception modeling.

Another classification of objective quality metrics is the availability of the original signal, which is considered to be distortion free or of perfect quality, and might be used as reference signal to evaluate the distorted signal. Based on the availability of the original signal, the quality metrics can be divided into three categories [1,2]: Full-Reference (FR) metrics, which require the processed signal and the complete reference signal [28–36], Reduced-Reference (RR) metrics, which require the processed signal and only part of the reference signal [23, 37], and No-Reference (NR) metrics, which require only the processed signal [38–41]. Traditional signal fidelity metrics for quality evaluation are FR metrics. Most perceptual quality metrics are of the FR type, including most perception-driven quality metrics and many signal-driven visual quality metrics. Most perceptual audio/speech quality metrics are FR. Generally, FR quality metrics can measure the quality of signals more accurately compared with RR or NR metrics, since they have more information available.

3.2.1 Full-Reference Quality Metrics

FR quality metrics evaluate the quality of the processed signal with respect to the reference signal. The traditional signal fidelity metrics – such as MSE, SNR, and PSNR – are early FR metrics. They have been the dominant quantitative performance metrics in the field of signal processing for decades. Although they exhibit poor accuracy when dealing with perceptual signals, they are still the standard criterion and widely used. Signal fidelity metrics in quality assessment try to provide a quantitative score that describes the level of error/distortion for the processed signal by comparing it with the reference signal. Suppose that $X = \{x_i | i = 1, 2, ..., N\}$ is a finite-length, discrete original signal (reference signal) and $\hat{X} = \{\hat{x}_i | i = 1, 2, ..., N\}$ is the corresponding distorted signal of the original signal, where N is the signal length, and x_i and $\hat{x}_i$ are the values of the ith samples in X and $\hat{X}$, respectively. The MSE between the distorted and the reference signals is calculated as

$$\mathrm{MSE}\left(\hat{X}, X\right) = \frac{1}{N} \sum_{i=1}^{N} \left(\hat{x}_i - x_i\right)^2 \tag{3.1}$$

where $\mathrm{MSE}\left(\hat{X}, X\right)$ is used as the quality measurement of the distorted signal $\hat{X}$. A more general form of the distortion $d_p\left(\hat{X}, X\right)$ is the l_p norm [11]:

$$d_p(\hat{X}, X) = \left(\sum_{i=1}^{N} \left(\hat{x}_i - x_i\right)^p \right)^{\frac{1}{p}} \tag{3.2}$$

The PSNR measure can be obtained from MSE as

$$\mathrm{PSNR} = 10\ \log_{10} \frac{\mathrm{MAX}^2}{\mathrm{MSE}} \tag{3.3}$$

where **MAX** is the maximum possible signal intensity value.

Aside from traditional signal fidelity metrics, many perceptual metrics are also FR metrics. As introduced previously, the perception-driven approach of quality assessment mainly measures the quality of images/video by modeling the characteristics of HVS. In the simplest perception-driven approaches, the HVS is considered as a single spatial filter modeling the CSF [42–45]. Many more sophisticated perception-driven approaches of FR metrics try to incorporate local contrast, spatiotemporal CSF, contrast/activity masking, and other advanced HVS functions to build quality metrics for images/video [19, 20–22, 46–50]. Compared with perception-driven approaches, signal-driven approaches for quality assessment are relatively less sophisticated and thus computationally inexpensive. Currently, there are many FR metrics for signal-driven approaches to visual quality assessment [35,36, 51–54]. Similarly, most perceptual audio/speech quality metrics are FR [3, 55–76]. Early audio quality metrics were designed for low-bit-rate speech and audio codecs. Perceptual-based models were used to optimize distortion for minimum audibility rather than traditional signal fidelity metrics, leading to an improvement in perceived quality [30]. Based on the characteristics of the HAS, various perceptual FR audio/speech quality metrics have been proposed [17, 63–76]. A detailed discussion of audio/speech quality metrics will be provided in Section 3.5.

Generally, FR metrics require the complete reference signal, usually in unimpaired and uncompressed form. This requirement is quite a heavy restriction for practical applications. Furthermore, FR metrics generally impose a precise alignment of the reference and distorted signals, so that each sample in the distorted signal can be matched with its corresponding reference sample. For video or audio signals, temporal registration in particular can be very difficult to achieve in practice due to the information loss, content repeats, or variable delays introduced by the system. Aside from the issue of spatial and temporal alignment, FR metrics usually do not respond well to global shifts in certain features (such as brightness, contrast, or color), and require a corresponding calibration of the signals. Therefore, FR metrics are most suitable for offline signal quality measurements such as codec tuning or lab testing.

3.2.2 Reduced-Reference Quality Metrics

RR quality metrics only require some information about the reference (e.g., in the form of a number of features extracted from the reference signal) for quality assessment tasks [77–108]. Normally, the more reference information is available to an RR metric, the more accurate the predictions it can make. In the extreme, when the rate is large enough to reconstruct the original signal, RR quality metrics converge to FR metrics.

Based on the underlying design philosophy, RR quality metrics can be classified into three approaches [77]: modeling the signal distortion, using characteristics or theories of the human perception system, and analysis of signal source. The last two types can be considered as general-purpose metrics, since the statistical and perceptual features are not related to any specific type of signal distortion.

The first type of RR quality metrics based on modeling signal distortion is developed mainly for specific applications. Straightforward solutions can be provided by these methods when there is sufficient knowledge about the processing of the content. With signal distortion from standard image or video compression, RR quality metrics can define the typical distortion artifacts of blurring and blockiness to measure the quality of the related visual content [79, 81]. Various RR quality metrics are proposed to measure the distortions occurring in standard compression systems such as MPEG2-coded video [82], JPEG images [83], H.264/AVC-coded video [84], etc. The drawback with this kind of RR quality metrics is the generalization capability, since they are always designed for certain specific kinds of signal distortion.

RR quality metrics based on the human perceptual system measure content quality by extracting perceptual features, where computational models of human perception may be employed. In [86], RR quality metrics are proposed by extracting perceptual features from JPEG and JPEG2000 images and obtain good evaluation performance. Many RR quality metrics are designed for video quality evaluation based on various characteristics of HVS such as color perception theory [87], CSF [88–90], structural information perception [91], texture masking [92, 93], etc. Aside from the features extracted in the spatial domain, there are also some RR quality metrics built based on features extracted by contourlet transform [90], wavelet transform [94], Fourier transform [95], etc.

The third type of RR quality metrics measure content quality based on models of the signal source. Since the reference signal is not available in a deterministic sense, these models are often based on statistical knowledge and capture certain statistical properties of the natural scenes [77, 85]. The distortions disturb the statistical properties of the natural scenes in unnatural ways, which can be measured by the statistical models of natural scenes. Many RR quality metrics are designed based on the difference calculation from feature distributions of color [99], motion [100], etc. This type of metrics can also be designed based on features in the transform domain such as divisive normalization transform [104], wavelet transform [105, 106], Discrete Cosine Transform (DCT) [102, 107, 108], etc.

RR approaches make it possible to avoid some of the assumptions and pitfalls of pure no-reference metrics while keeping the amount of reference information manageable. Similar to FR metrics, RR metrics also have alignment requirements. However, they are typically less stringent than full-reference metrics, as only the extracted features need to be aligned. Generally, RR metrics are better suited for monitoring in-service content at different points in the distribution system.

3.2.3 No-Reference Quality Metrics

Compared with FR and RR quality metrics, NR quality metrics do not require any reference information [109–156]. Thus, they are highly desirable in many practical applications where reference signals are not available.

A number of methods have been proposed to predict the MSE caused by certain specific compression schemes such as MPEG2 [112, 114], JPEG [116], or H.264 [117, 118, 120]. These methods use information from the bit stream directly, except the study [112] which adopts the decoded pixel information. The DCT coefficients in these techniques are usually modeled by Laplacian, Gaussian, or Cauchy distributions. The main drawback of these techniques is that they measure the distortion for each 8×8 block without considering the differences from neighboring blocks [110]. Another problem with these methods is that the performance decreases with lower bit rate due to more coefficients quantized to zero. Some NR visual quality metrics have tried to measure the MSE caused by packet loss errors [121, 124, 139], the difference between the processed signal and the smoothed signal [126], the variation within the smooth regions in the signal [129], etc.

Generally, NR quality metrics assume that the statistics of the processed signals are different from those of the original and extract features from the processed signals to evaluate model compliance [85, 110]. NR quality metrics can be designed based on features of various domains, such as the spatial domain [27, 132], Fourier domain [24], DCT domain [134, 135], or polynomial transform [136]. Additionally, many NR quality metrics are based on various features from visual content, such as sharpness [137], edge extent [138, 140], blurring [143, 144], phase coherence [145], ringing [146, 147], naturalness [148], or color [149]. In some NR quality metrics, blockiness, blurring, or ringing features are combined with other features, such as bit-stream features [150, 151], edge gradient [152], and so on. Compared with image content, temporal features have to be considered for quality assessment of video and audio signals. Various NR quality metrics for video content are proposed to measure flicker [153] or frame freezes [154, 156]. There are also some NR metrics proposed for speech quality evaluation [109, 113, 115, 119, 122, 123], which are mainly designed based on analysis of the audio spectrum.

NR metrics analyze the distorted signal without the need for an explicit reference signal. This makes them much more flexible than FR or RR metrics, as it can be difficult or impossible to get access to the reference in some cases (e.g., video coming out of a camera). They are also completely free from alignment issues. The main challenge of NR metrics lies in telling apart distortions from content, a distinction humans are usually able to make from experience. NR metrics always have to make assumptions about the signal content and/or the distortions of interest. This comes with a risk of confusing the actual content with distortions (e.g., a chessboard could be interpreted as a block of artifacts under certain conditions). Additionally, most NR quality metrics are designed for specific and limited types of distortion. They can face difficulties in modern communication systems, where distortions could be a combination of compression, adaptation, network delay, packet loss, and various types of process filtering. NR metrics are suited for monitoring in-service content at different points in the distribution system, since they do not require reference signals.

3.3 Basic Computational Modules for Perceptual Quality Metrics

Since the traditional signal fidelity metrics assess the quality of the distorted signal by simply comparing it with the reference one and they are well introduced in Section 3.2.1, here we just provide the basic computational modules for perceptual quality metrics. Generally, these include signal decomposition (e.g., decomposing an image or video into different color, spatial, and temporal channels), detection of common features (like contrast and motion) and artifacts (like blockiness and blurring), just-noticeable distortion (i.e., the maximum change in visual content that cannot be detected by the majority of viewers), Visual Attention (VA) (i.e., the HVS's selectivity in responding to the most interesting activities in the visual field), etc. First, many of these are based on related physiological and psychological knowledge. Second, most are independent research topics themselves, like just-noticeable distortion and VA modeling, and have other applications (image/video coding [157], watermarking [158], error resilience [159], computer graphics [160], to name just a few), in addition to perceptual quality metrics. Third, these modules can be simple perceptual quality metrics themselves in specific situations (e.g., blockiness and burring).

3.3.1 Signal Decomposition

Most perception-driven quality metrics use signal decomposition for feature extraction. Signal feature extraction and common artifact detection are at the core of many signal-driven quality metrics; the perceptual effect of common artifacts far exceeds the extent of their representation in MSE or PSNR. Just-noticeable distortion and VA models have been used either independently or jointly to evaluate the visibility and perceived extent of visual content differences. Therefore, all these techniques help to address the three basic problems (as mentioned at the beginning of this chapter) to be overcome relative to traditional signal fidelity metrics, since they enable the differentiation of various content changes for perceptual quality-evaluation purposes.

For images and video, the process of signal decomposition refers to the decomposition of visual content into different channels (spatial, frequency, and temporal) for further processing. It is well known that the HVS has separate processing for achromatic and chromatic components, different pathways for visual content with different motion, and special cells in the visual cortex for distinctive orientations [161]. Existing psychophysical studies also show that visual content is processed differently in the HVS by frequency [162] and orientation [163, 164]. Thus, decomposition of an image or video frame into different color, spatial, and temporal channels can evaluate content changes for unequal treatment of each channel to emulate the HVS response, which can address the third problem of traditional signal fidelity metrics mentioned at the beginning of this chapter.

Currently, there are various signal decomposition methods for color [165–168]. Two widely accepted color spaces in quality assessment are the opponent-color (black/white, red/green, blue/yellow) space [22, 166] based on physiological evidence of opponent cells in the parvocellular pathway and CIELAB space [167] based on human perceptions of color differences. With compressed visual content, YCbCr space is more convenient for feature extraction due to its wide use in image/video compression standards [53, 165, 168]. Other color spaces have also been used, such as YOZ [21]. In many metrics, only the luminance component of the visual content is used for efficiency [49, 169–171], since it is generally more important for human visual perception than chrominance components, especially in quality evaluation of compressed images (it is worthwhile pointing out that most coding decisions in current image/video compression algorithms are made based on luminance manipulation).

Temporal decomposition is implemented by a sustained (low-pass) filter and transient (bandpass) filters [46, 172] to stimulate two different visual pathways. Based on the fact that receptive fields in the primary visual cortex resemble Gabor patterns [173] that can be characterized by a particular spatial frequency and orientation, many types of spatial filter can be used to decompose each temporal channel, including Gabor filters, cortex filters [174], wavelets, Gaussian pyramid [175], and steerable pyramid filters [46, 176].

For audio/speech signals, signal decomposition is implemented based on the properties of the peripheral auditory system – such as the perception of loudness, frequency, masking, etc. [177]. In the Perceptual Evaluation of Audio Quality (PEAQ) ITU standard [178], two psychoacoustic models are adopted to transform the time-domain input signals into a basilar membrane representation for further processing: the FFT (Fast Fourier Transform)-based ear model and the filter-bank-based ear model [3]. For the FFT-based ear model, the input signal is first transformed into the frequency domain. The amplitude of the FFT is used for further processing. Then the characteristics of the effect in the outer and middle ear on audio signals are modeled based on Terhardt's approach for frequency components [38]. After that, the frequency components are grouped into critical frequency bands as perceived by the HAS. An internal ear-noise model is used to obtain the pitch patterns for audio signals [3]. These pitch patterns are smeared out over the frequencies by a spreading function modeling simultaneous masking. Finally, the forward masking characteristics of temporal masking effects are approximated by a simple first-order low-pass filter. In the filter-bank-based ear model, the audio signal is processed in the time domain [3]. Compared with the FFT-based ear model, the filter-bank-based ear model adopts a finer time resolution, which makes the modeling of backward masking possible and thus maintains the fine temporal structure of the signal. PEAQ is mainly based on the model in [179]. First, the input audio signal is decomposed into band-pass signals by a filter bank of equally spaced critical bands. Similar to the FFT-based ear model, the effects in the outer and middle ear on audio signals are modeled. Then the characteristics of simultaneous masking, backward masking, internal ear noise, and forward masking are modeled subsequently for the final feature extraction for audio signals [3].

Similar signal decomposition methods based on psychoacoustic models have been implemented in many studies [67, 75]. During the transformation, the input audio signals are decomposed into different band-pass signals by modeling various characteristics in the HAS, such as characteristics of effect of the outer and middle ear [38], simultaneous masking (frequency spreading) [180], forward and backward temporal masking effects [3, 122], etc. Following PEAQ, some other new psychoacoustic models have been proposed by incorporating recent findings into the design of perceptual audio quality metrics [17, 70].

3.3.2 Feature and Artifact Detection

The process of feature and artifact detection is common for visual quality evaluation in various scenarios. For example, meaningful visual information is conveyed by feature contrast such as luminance, color, orientation, texture, motion, etc. There is little or no information in a largely uniform image. The HVS perceives much more from signal contrast than from absolute signal strength, since there are specialized cells to process this information [181]. This is also the reason why contrast is central to CSF, luminance adaptation, contrast masking, visual attention, and so on.

For audio/speech quality metrics, various features are extracted for artifact detection in transform domains such as modulation, loudness, excitation, and slow-gain variation features [3, 178]. For NR metrics of speech quality assessment, the perceptual linear prediction coefficients are used for quality evaluation [115, 128]. In [155], the vocal tract and unnaturalness features of speech are extracted from speech signals for quality evaluation. In PEAQ, the cognitive model processes the parameters from the psychoacoustic model to obtain Model Output Variables (MOVs) and maps them to a single Overall Difference Grade (ODG) score [3]. The MOVs are extracted based on various parameters including loudness, amplitude modulation, adaption, masking, etc., and they also model concepts such as linear distortion, bandwidth, modulation difference, noise loudness, etc. The MOVs are used as input to the neural network and mapped to a distortion index. Then the ODG is calculated based on the distortion index to estimate the quality of the audio signal.

There are certain structural artifacts occurring in the prevalent signal compression and delivery process which result in annoying effects for the viewer. The common structural artifacts caused by coding include blockiness, blurring, edge damage, and ringing [171], whose perceptual effect is ignored in traditional signal fidelity metrics such as MSE and PSNR. In fact, uncompressed images/video usually include blurring artifacts due to the imperfect PSF (Point

Spread Function) and an out-of-focus imaging system as well as object motion during the signal capture process [182]. In video quality evaluation, the effects of motion and jerkiness have been investigated [156, 183]. Similarly, coding distortions in audio/speech signals have been well investigated [30, 63–65]. Some studies have investigated the quality evaluation of noise-suppressed audio/speech [125].

Another type of quality metrics is designed specifically to measure the impact of network losses on perceptual quality. This development is the result of increasing multimedia service delivery over IP networks, such as Internet streaming or IPTV. Since information loss directly affects the encoded bit stream, such metrics are often designed based on parameters extracted from the transport stream and the bit stream with no or little decoding. This has the added advantage of much lower data rates and thus lower bandwidth and processing requirements compared with metrics looking at the fully decoded video/audio. Using such metrics, it is thus possible to measure the quality of many video/audio streams or channels in parallel. At the same time, these metrics have to be adapted to specific codecs and network protocols. Due to the different types of features and artifacts, so-called "hybrid" metrics use a combination of different features or approaches for quality assessment [5]. Some studies explore the joint impact of packet loss rate and MPEG-2 bit rate on video quality [184], the influence of bit-stream parameters (such as motion vector length or number of slice losses) on the visibility of packet losses in MPEG-2 and H.264 videos [139], the joint impact of the low-bit-rate codec and packet loss on audio/speech quality [185], etc.

In some quality metrics, multiple features or quality evaluation approaches are combined. In [152], several structural features such as blocking, blurring, edge-based image activity, gradient-based image activity, and intensity masking are linearly combined for quality evaluation. The feature weights are determined by a multi-objective optimization method [152]. In [169], the input visual scene is decomposed into predicted and disorderly portions for quality assessment based on an internal generative mechanism in the human brain. Structure similarity and PSNR metrics are adopted for quality evaluation in these two portions respectively, and the overall score is obtained by combining these two results with an adaptive nonlinear procedure [169]. In [186], phase congruency and gradient magnitude are employed as two complementary roles for the quality assessment of images. After calculating the local quality map, the phase congruency is adopted again as a weighting function to derive the overall score [186]. The study in [187] presents a visual quality metric based on two strategies in the HVS: the detection-based strategy for high-quality images containing near-threshold distortions and the appearance-based strategy for low-quality images containing clearly supra-threshold distortions. Different measurement methods are designed for these two types of quality level, and the overall quality evaluation score is obtained by combining the results adaptively [187]. In [70], the linear and nonlinear distortions in the perceptual transform (excitation) domain are combined linearly for speech quality evaluation. In [76], the audio quality evaluation results from spectral and spatial features are multiplied to obtain the overall quality of audio signals.

Recently, a new fusion method for different features or quality evaluation approaches has emerged based on machine learning techniques [111, 125, 188, 189]. In [111], machine learning is adopted for the feature pooling process in visual quality assessment to address the limitations of existing pooling methods such as linear combination. A similar pooling process by support vector regression is introduced for speech quality assessment in [125]. Rather than using machine learning techniques for feature pooling, a multi-method fusion quality metric is introduced based on the nonlinear combination of scores from existing methods with suitable weights from a training process in [189]. In some NR quality metrics, machine learning techniques are also adopted to learn the mapping from feature space to quality scores [190].

3.3.2.1 Contrast

In [191], band-pass-filtered and low-pass-filtered images are used to evaluate the local image contrast. Following this methodology, image contrast is calculated as the ratio of the combined analytic oriented filter response to the low-pass filtered image in the wavelet domain [192] or the ratio of high-pass response in the Haar wavelet space [193]. Luminance contrast is estimated as the ratio of the noticeable pixel change to the average luminance in a neighborhood [53]. The contrast can also be computed as a local difference between the reference video frame and the processed one with the Gaussian pyramid decomposition [47], or the comparison between DCT amplitudes and the amplitude of the DC coefficient of the corresponding block [21]. The k-means clustering algorithm can be used to group image blocks for the calculation of color and texture contrast [194], where the largest cluster is considered as the image background. The contrast is then computed as the Euclidean distance from the means of the corresponding background cluster. Motion contrast can be obtained by relative motion, which is represented by object motion against the background [194].

3.3.2.2 Blockiness

Blockiness is a prevailing degradation caused by the block-based DCT coding technique, especially under low-bit-rate conditions, due to the different quantization sizes used in neighboring blocks and the lack of consideration for inter-block correlation. Given an image I with width W and height H, which is divided into $N \times N$ blocks, the horizontal and vertical difference at block boundaries can be computed as [27]

$$M_h = \left[\sum_{k=1}^{\frac{H}{N}-1} \sum_{x=0}^{W-1} \left(I\left(x, k * N - 1\right) - I\left(x, k * N\right) \right)^2 \right]^{\frac{1}{2}} \tag{3.4}$$

$$M_v = \left[\sum_{l=1}^{\frac{W}{N}-1} \sum_{y=0}^{H-1} \left(I\left(l * N - 1, y\right) - I\left(l * N, y\right) \right)^2 \right]^{\frac{1}{2}} \tag{3.5}$$

The method in [27] works only for a regular block structure with certain block size. It cannot work in modern video codecs (e.g., HEVC (High Efficiency Video Coding, H.265)) due to the different block sizes used. Other, similar calculation methods for blockiness can be found in [25, 195]. During the blockiness calculation, object edges at block boundaries can be excluded

[29]. Luminance adaptation and texture masking have recently been considered for blockiness evaluation [135]. Another method for gauging blockiness is based on harmonic analysis [196], which can be used in the case when block boundary positions are unknown beforehand (e.g., with video being cropped, retaken by a camera, or coded with variable block sizes).

3.3.2.3 Blurring

Blurring can be evaluated effectively around edges in images/video frames, since it is most noticeable there, and such detection is efficient due to only a small fraction of image pixels on edges. With an available reference signal, the extent of blurring can be estimated via contrast decrease on edges [53]. Various blind methods without a reference signal have been proposed for measuring the blurring/sharpness, such as edge spread detection [26, 40], kurtosis [146], frequency domain analysis [143, 197], PSF estimation [198], width/amplitude of lines and edges [39], and local contrast via 2D analytic filters [199].

3.3.2.4 Motion and Jerkiness

For visual quality evaluation of coded video, the major temporal distortion is jerkiness, which is mainly caused by frame dropping [200] and is very annoying to viewers, who prefer continuous and smooth temporal transitions. For decoded video without availability of the coding parameters, frame freeze can be simply detected by frame differences [201]; in the case when the frame rate is unavailable, the jerkiness effect can be evaluated using the frame rate [156, 183] or more comprehensively, both the frame rate and temporal activity such as motion [202]. In [203], inter-frame correlation analysis is used to estimate the location, number, and duration of lost frames. In [154, 200], lost frames are detected by inter-frame dissimilarity to measure fluidity; these studies conclude that, for the same level of frame loss, scattered fluidity breaks introduce less quality degradation than aggregated ones. The impact of the time interval between occurrences of significant visual artifacts has also been investigated [204].

3.3.3 Just-Noticeable Difference (JND) Modeling

As mentioned previously, not every signal change is noticeable. JND refers to a visibility or audibility threshold below which a change cannot be detected by the majority of viewers [201, 205–212]. Obviously, if a difference is below the JND value, it can be ignored in quality evaluation.

For images and video, DCT-based JND is the most investigated topic among all sub-band-based JND functions, since DCT has been used in all existing image/video compression standards such as JPEG, H.261/3/4, MPEG-1/2/4, and SVC. A general form of the DCT-sub-band luminance JND function is introduced in [52, 201, 207]. The widely used JND function developed by Ahumada and Peterson [211] for the base-line threshold fits spatial CSF curves with a parabola equation, which is a function of spatial frequencies and background luminance, and then compensates for the fact that the psychophysical experiments for determining CSF were conducted with a single signal at a time, and with spatial frequencies along just one direction. The luminance adaptation factor has been determined to represent the variation versus background luminance [205], to be more consistent with the findings of subjective viewing of digital images [206, 212]. The intra-band masking effect was investigated in [52, 208]. Inter-band masking effects can be assigned as low, medium, or high masking after classifying DCT blocks into smooth, edge, and texture ones [205, 207], according to energy distribution among sub-bands. For temporal CSF effects, the velocity perceived by the retina for an image block needs to be estimated [209]. A method for incorporating the effect of the velocity for temporal CSF in JND is given in [210].

The JND can also be defined in other frequency bands (e.g., Laplacian pyramid image decomposition [210], Discrete Wavelet Transform (DWT) [213]). In comparison with DCT-based JND, significantly more research is needed for DWT-based JND. DWT is a popular alternative transform, and more importantly, is similar to the HVS in its multiple sub-channel structure and frequency-varying resolution. Chrominance masking [214] still needs more convincing investigation for all sub-band domains.

There are situations where JND estimated from pixels is more convenient and efficient to use (e.g., motion search [168], video replenishment [215], filtering of motion-compensated residuals [165], and edge enhancement [48, 216]), since the operations are usually performed on pixels rather than sub-bands. For quality evaluation of images and video, pixel-domain JND models avoid unnecessary sub-band decomposition. Most pixel-based JND functions developed so far have used luminance adaptation and texture masking. A general pixel-based JND model can be found in [216]. The temporal effect was addressed in [217] by multiplying the spatial effect with an elevation parameter increasing with inter-frame changes. The major shortcoming of pixel-based JND modeling lies in the difficulty of incorporating CSF explicitly, except for the case with conversion from a sub-band domain [218].

For audio/speech signals, there are some types of JND based on different components of the signals – such as amplitude, frequency, etc. [219]. For the amplitude of audio/speech signals, the JND for the average listener is about 1 dB [130, 220], while the frequency JND for the average listener is approximately 1 Hz for frequencies below 500 Hz and about *f*/500 for frequencies *f* above 500 Hz [221]. The ability to discriminate audio/speech signals in the temporal dimension is another important characteristic of the acuity of the HAS. The duration of the gap between two successive audio/speech signals must be at least 4–6 ms in length to be detected correctly [222]. The ability of the HAS to detect changes over time in the amplitude and frequency of audio/speech signals is dependent on the rate of change and amount of change in amplitude and frequency [131].

3.3.4 Attention Modeling

Not every part of a multimedia presentation receives the same attention (the second problem of signal fidelity metrics mentioned in the introduction). This is due to the fact that human perception selects a part of the signal for detailed analysis and then responds. VA refers to the selective awareness/responsiveness to visual stimuli [223], as a consequence of human evolution.

There are two types of cue that direct attention to a particular point in an image [224]: bottom-up cues that are determined by external stimuli, and top-down cues that are caused by a voluntary shift in attention (e.g., when the subject is given prior information/instruction to direct attention to a specific location/object). The VA process can be regarded as two stages [225]: in the pre-attention stage, all information is processed across the entire visual field; in the attention stage, the features may be bound together (feature integration [226], especially for a bottom-up process), or the dominant feature is selected [227] (for a top-down process).

Most existing computational VA models are bottom-up (i.e., based on contrast evaluation of various low-level features in images, in order to determine which locations stand out from their surroundings). As to the top-down (or task-oriented) attention, there is still a need for more focused research, although some initial work has been done [228, 229].

An influential bottom-up VA computational model was proposed by Itti *et al.* [230] for still images. An image is first low-pass filtered and down-sampled progressively from scale 0 (the original image size) to scale 8 (1:256 along each dimension). This is to facilitate the calculation of feature contrast, which is defined as

$$\hat{F}(e,q) = |F(e) - F^{l}(q)| \tag{3.6}$$

where F represents the map for one of the image features as follows: intensity, color, and orientation; $e \in \{2,3,4\}$ and $F(e)$ denote the feature map at scale e; $q = e + \delta$, with $\delta \in \{3,4\}$, and $F^{l}(q)$ is the interpolation to the finer scale e from the coarser scale q. In essence, $\hat{F}(e,q)$ evaluates pixel-by-pixel contrast for a feature, since $F(e)$ represents the local information, while $F^{l}(q)$ approximates the surroundings.

With one intensity channel, two color channels, and four orientation channels (0°, 45°, 90°, 135°; detected by Gabor filters), 42 feature maps are computed: 6 for intensity, 12 for color, and 24 for orientation. After cross-scale combination and normalization, the winner-takes-all strategy identifies the most interesting location on the map. There are various other approaches for visual attention modeling [231–234].

The VA map along the temporal dimension (over multiple consecutive video frames) can also be estimated. In the scheme proposed in [194] for video, different features (such as color, texture, motion, human skin/face) were detected and integrated for the continuous (rather than the winner-takes-all) salience map. In [235], auditory attention was also considered and integrated with visual factors. This was done by evaluating sound loudness and its sudden change, and a Support Vector Machine (SVM) was employed to classify each audio segment into speech, music, silence, and other sounds; the ratio of speech/music to other sounds was measured for saliency detection. Recently, Fang *et al.* proposed saliency detection models for images and videos based on DCT coefficients (with motion vectors for video) in the compressed domain [233, 236]. These models can be combined with quality metrics obtained from DCT coefficients and motion vectors for visual quality evaluation. A detailed overview and discussion of visual attention models can be found in a recent survey paper [232].

Contrast sensitivity reaches its maximum at the fovea and decreases toward the peripheral retina. The JND model represents the visibility threshold when the attention is there. In other words, JND and VA account for the local and global responses of the HVS in appreciating an image, respectively. The overall visual sensitivity at a location in the image could be JND modulated by the VA map [194]. Alternatively, the overall visual sensitivity may be derived by modifying the JND at every location according to its eccentricity away from the foveal points, with the foveation model in [237].

VA modeling is generally easier for video than still images. If an observer has enough time to view an image, many points of the image will be attended to eventually. The perception of video is different: every video frame is displayed to an observer for a very short time interval. Furthermore, camera and/or object motion may guide the viewer's eye movements and attention.

Compared with visual content, there is much less research on auditory attention modeling. Currently, there are several studies proposing auditory attention models for audio signals [238–240]. Motivated by the formation of auditory streams, the study [239] designs a conceptual framework of auditory attention, which is implemented as a computational model composed of a network of neural oscillators. Inspired by the successful visual saliency detection model proposed by Itti *et al.* [230], some other auditory attention models are proposed using feature contrast – such as frequency contrast, temporal contrast, etc. [238, 240].

3.4 Quality Metrics for Images

The early image quality metrics are traditional signal fidelity metrics, which include MAE, MSE, SNR, PSNR, etc. As discussed previously, these metrics cannot predict image distortions as they are perceived. To address the drawback of traditional signal fidelity metrics, various perceptual-based image quality metrics have been proposed in the past decades [7, 11, 12, 24, 33–36]. These are classified and introduced in the following.

3.4.1 2D Image Quality Metrics

3.4.1.1 FR Metrics

Early perceptual image quality metrics were developed based on simple and systematic modeling of relevant psychophysical or physiological properties. Mannos and Sakrison [10] proposed a visual fidelity measure based on CSF for images. Faugeras [42] introduced a simple model of human color vision based on experimental evidence for image evaluation. Another early FR and multichannel model is the Visible Differences Predictor (VDP) of Daly [19], where the HVS model accounts for sensitivity variations due to luminance adaptation, spatial CSF, and contrast masking. The cortex transform is performed for signal decomposition, and different orientations are distinguished. Most existing schemes in this category follow a similar methodology, with differences in the color space adopted, the type of spatiotemporal decomposition, or the error pooling methods. In the JNDmetrix model [20], the Gaussian pyramid [175] was used for decomposition, with luminance and chrominance components in the image. Liu *et al.* [48] proposed a JND model to measure the visual quality of images. The perceptual effect can be derived by considering inter-channel masking [46]. Other similar algorithms using CSF and visual masking are described in [25, 241].

Recently, various perceptual image quality metrics have been proposed using signal modeling or processing of visual signals under consideration, which incorporate certain specific knowledge (such as the specific distortion [21]). This approach is relatively less sophisticated and therefore less computationally expensive. In [53], the image distortion is measured by the DCT coefficient differences weighted by JND. Similarly, the well-cited SSIM (Structural

SIMilarity) was proposed by Wang and Bovik based on the sensitivity of the HVS to image structure [36, 54, 242]. SSIM can be calculated as

$$Q = \frac{\sigma_{ab}}{\sigma_a \sigma_b} \frac{2\sigma_a \sigma_b}{(\sigma_a)^2 + (\sigma_b)^2} \frac{2\bar{a}\bar{b}}{(\bar{a})^2 + (\bar{b})^2} \tag{3.7}$$

where a and b represent the original and test images; $\bar{a}$ and $\bar{b}$ are their corresponding means, σ_a and σ_b are the corresponding standard deviations; σ_{ab} is the cross covariance. The three terms in equation (3.7) measure the loss of correlation, contrast distortion, and luminance distortion, respectively. The dynamic range of the SSIM value Q is $[-1, 1]$, with the best value of 1 when $a = b$.

Studies show that SSIM bears a certain relationship with MSE and PSNR [55, 243]. In [243], PSNR and SSIM are compared by their analytical formulas. The analysis shows that there is a simple logarithmic link between them for several common degradations, including Gaussian blur, additive Gaussian noise, JPEG and JPEG2000 compression [243]. PSNR and SSIM can be considered as closely related quality metrics, with differences in the degree of sensitivity to some image degradations.

Another method for feature detection with consideration of structural information is Singular Value Decomposition (SVD) [111]. With more theoretical background, the Visual Information Fidelity (VIF) [35] (an extension of the study [34]) is proposed based on the assumption that the Random Field (RF) from a sub-band of the test image, D, can be expressed as

$$D = GU + V \tag{3.8}$$

where U denotes the RF from the corresponding sub-band of the reference image, G is a deterministic scale gain field, and V is a stationary additive zero-mean Gaussian noise RF. The proposed model takes into account additive noise and blur distortion; it is argued that most distortion types prevalent in real-world systems can be roughly described locally by a combination of these two. The resultant metric measures the amount of information that can be extracted about the reference image from the test. In other words, the amount of information lost from a reference image as a result of distortion gives the loss of visual quality.

Another image quality metric with good theoretical foundations is the Visual Signal-to-Noise Ratio (VSNR) [33], which operates in two stages. In the first stage, the contrast threshold for distortion detection in the presence of the image is computed via wavelet-based models of visual masking and visual summation, in order to determine whether the distortion in the test image is visible. If the distortion is below the threshold of detection, the test image is deemed to be of perfect visual fidelity (VSNR $= \infty$). If the distortion is above the threshold, a second stage is applied, which operates based on the property of perceived contrast, and the mid-level visual property of global precedence. These two properties are modeled as Euclidean distances in distortion-contrast space of a multiscale wavelet decomposition, and VSNR is computed based on a simple linear sum of these distances.

Larson and Chandler [187] proposed a perceptual image quality metric called the “most apparent distortion” based on two separate strategies. Local luminance and contrast masking are adopted to estimate detection-based perceived distortion in high-quality images, while changes in the local statistics of spatial-frequency components are used to estimate appearance-based perceived distortion in low-quality images [187]. Recently, some new image quality metrics have been proposed using new concepts or methods [111, 169, 188, 189, 244]. Wu *et al.* [169] adopted the concept of Internal Generative Mechanism (IGM) theory to divide image regions into two different parts of predicted portion and disorderly portion, which are measured by the structural similarity and PNSR metrics, respectively. Liu *et al.* [244] used gradient similarity to measure the change in contrast and structure. A recent emerging scheme for an image quality metric is based on machine learning techniques [111, 188, 189].

3.4.1.2 RR Metrics

Some RR metrics for images are designed based on the properties of the HVS. In [88], several factors of the HVS – including CSF, psychophysical sub-band decomposition, and masking effect modeling – are adopted to design an RR quality metric for images. The study in [91] proposes an RR quality metric for wireless imaging based on the observation that HVS is trained to extract structural information from the viewing area. In [94], an RR quality metric is designed based on the wavelet transform, which is used for extracting features to simulate the psychological mechanisms of HVS. The study in [95] adopts the phase and magnitude of the 2D discrete Fourier transform to build an RR quality metric, which is motivated by the fact that the sensitivity of the HVS is frequency dependent. Recently, Zhai *et al.* [245] used the free-energy principle from cognitive processing to develop a psychovisual RR quality metric for images.

In RR metrics for images, various features can be extracted to measure the visual quality. The image distortion of some RR metrics is calculated based on features extracted from the spatial domain – such as color correlograms [99], image statistics in the gradient domain [101], structural information [77, 86], etc. Other RR metrics are proposed using features extracted in the transform domain – such as wavelet coefficients [98, 105, 106], coefficients from divisive normalization transform [104], DCT coefficients [102, 108], etc.

Some RR metrics are designed for specific distortion types. A hybrid image quality metric is designed by fusing several existing techniques to measure five specific artifacts in [79]. Other RR metrics are proposed to measure the distortion from JPEG compression [83], distributed source coding [96], etc.

3.4.1.3 NR Metrics

NR quality metrics for images are proposed based on various features or specific distortion types. Many studies compute edge-extent features of images to build their NR quality metrics [136, 138, 140]. The natural scene statistics of DCT coefficients is used to measure the visual quality of images in [116]. In that metric, a Laplace probability density function is adopted to model the distribution of DCT coefficients. DCT coefficients are also used to measure blur artifacts [143], blockiness artifacts [134, 135], and so on. Similarly, some NR quality metrics for images use the features extracted by Fourier transform to measure different types of artifact – such as blur artifact [145], blockiness artifact [24], etc. Besides, NR quality metrics can be designed based on other features – such as sharpness [137], ringing [146], naturalness [148], and color [149]. In some NR quality metrics, blockiness, blurring, or ringing features are combined with other features, such as the bit-stream feature [151], edge gradient [152], and so on. The noise-estimation-based NR image quality metrics calculate MSE based on the difference between the proposed signal and the smoothing signal [126], or the variation within certain smooth regions in visual signals [129].

3.4.2 3D Image Quality Metrics

Compared with visual quality metrics for 2D images, quality metrics for 3D images have to consider additionally the depth perception. The HVS uses a multitude of depth cues, which can be classified into oculomotor cues coming from the eye muscles, and visual cues from the scene content itself [163, 246, 247]. The oculomotor cues include the factors of accommodation and vergence [246]. Accommodation refers to the variation of the lens shape and thickness, which allows the eyes to focus on an object at a certain distance, while vergence refers to the muscular rotation of the eyeballs, which is used to converge both eyes on the same object. There are two types of visual cue, namely monocular and binocular [246]. Monocular visual cues include relative size, familiar size, texture gradients, perspective, occlusion, atmospheric blur, lighting, shading, and shadows, motion parallax, etc. The most important binocular visual cue is the retinal disparity between points of the same objects viewed from slightly different angles by the eyes, which is used in stereoscopic 3D systems such as 3DTV.

Although 3D image quality evaluation is a challenging problem due to the complexities of depth perception, a number of 3D image quality metrics have been proposed [163]. Most existing 3D image quality metrics evaluate the distortion of 3D images by combining the evaluation results of a 2D image pair and additional factors – such as depth perception, visual comfort, and other visual experiences. In [248], 2D image quality metrics are combined with disparity information to predict the visual quality of 3D compressed images with blurring distortion. Similarly, [249] integrates the disparity information with 2D quality metrics for quality evaluation for 3D images. In [250], a 3D image quality metric is designed based on absolute disparity information. Existing studies also explore the visual quality assessment for 3D images based on characteristics of the HVS – such as contrast sensitivity [251], viewing experience [252], binocular visual characteristics [253], etc.

Furthermore, there are several NR metrics proposed for 3D image quality evaluation. In [254], an NR 3D image quality metric is built for JPEG-coded stereoscopic images based on segmented local features of artifacts and disparity. Another NR quality metric for 3D image quality assessment is based on the nonlinear additive model, ocular dominance model, and saliency-based parallax compensation [141].

3.5 Quality Metrics for Video

The history and development of video quality metrics shares many similarities with image quality metrics, with the additional consideration of temporal effects.

3.5.1 2D Video Quality Metrics

3.5.1.1 FR Metrics

As stated previously, there are two types of perceptual visual quality metrics: vision-based and signal-driven [2, 255–269]. Vision-based approaches include for example [22, 255], where HVS-based visual quality metrics for coded video sequences are proposed based on contrast sensitivity and contrast masking. The study in [47] proposes a JND-based metric to measure the visual quality for video sequences. In [263], a MOtion-based Video Integrity Evaluation (MOVIE) metric is proposed based on characteristics of the Middle Temporal (MT) visual area of the human visual cortex for video quality evaluation. In [266], an FR quality metric is proposed to improve the video evaluation performance of the quality metrics in [264, 265]. Vision-based approaches typically measure the distortion of the processed video signal in the spatial domain [49], DCT domain [21], or wavelet domain [50, 262].

Signal-driven video quality metrics are based primarily on the analysis of specific features or artifacts in video sequences. In [256], Wang *et al.* propose a video structural similarity index based on SSIM to predict the visual quality of video sequences. In that study, an SSIM-based video quality metric evaluates the visual quality of video sequences from three levels: the local region level, the frame level, and the sequence level. A similar metric for visual quality evaluation is proposed in [257]. In [258], another SSIM-based video quality metric is designed based on a statistical model of human visual speed perception described in [259]. In [260], an FR video quality metric is proposed based on singular value decomposition. In [51], the video quality metric software tools provide standardized methods to measure the perceived quality of video systems. With the general model in that study, the main distortion types include blurring, blockiness, jerky/unnatural motion, noise in luminance and chrominance channels, and error blocks. In [261], a video quality metric is proposed based on the correlation between subjective (MOS) and objective (MSE) results. In [270], a low-complexity video quality metric is proposed based on temporal quality variations. Some existing metrics make use of both classes (vision-based and signal-driven). The metric proposed in [271] combines model-based and signal-driven methods based on the extent of blockiness in decoded video. A model-based metric was applied to blockiness-dominant areas in [29], with the help of a signal-driven measure.

Recently, various High-Definition Television (HDTV) videos have emerged, with large demand from users and increased development of high-speed wideband network services. Compared with Standard-Definition Television (SDTV) videos, HDTV content needs higher-resolution display screens. Although the viewing distance of HDTV systems is closer in terms of image height compared with SDTV systems, approximately the same number of pixels per degree of viewing angle exists in both these systems due to the higher spatial resolution in HDTV [267]. However, the higher total horizontal viewing angle for HDTV (about 30°) may influence the quality decisions compared with that for SDTV (about 12°) [267]. Additionally, with high-resolution display screens for HDTV, human eyes roam the picture in order to track specific objects and their motion, which causes the visual distortion outside the immediate area of attention to be perceived less compared with SDTV [267]. With emerging HDTV applications, some objective quality metrics have been proposed to evaluate the visual quality of HDTV specifically. The study in [267] conducts experiments to assess whether the NTIA general video quality metric [51] can be used to measure the visual quality of HDTV video. In [268], spatiotemporal features are extracted from visual signals to estimate the perceived quality of distortion caused by compression coding. In [269], an FR objective quality metric is proposed based on a fuzzy measure to evaluate coding distortion in HDTV content. Recently, some studies have investigated the visual quality for Ultra-High Definition (UHD) video sequences by subjective experiments [272–274]. The study in [274] conducts subjective experiments to analyze the performance of the popular objective quality metrics (PSNR, VSNR, SSIM, MS-SSIM, VIF, and VQM) on 4K UHD video sequences; experimental results show the content-dependent nature of most objective metrics, except VIF [274]. The subjective experiments in [273] demonstrate that the HEVC-encoded YUV420 4K-UHD video at a bit rate of 18 Mb/s has good visual quality in the usage of legacy DTV broadcasting systems with single-channel bandwidths of 6 MHz. The study in [272] presents a set of 15 4K UHD video sequences for the requirements of visual quality assessment in the research community.

3.5.1.2 RR Metrics

Video is a much more suitable application for RR metrics than images because of the streaming nature of the content and the much higher data rates involved. Typically, low-level spatiotemporal features from the original video are extracted as reference. Features from the reference video can then be compared with those from the processed video.

In the work performed by Wolf and Pinson [80], both spatial and temporal luminance gradients are computed to represent the contrast, motion, amount, and orientation of activity. Temporal gradients due to motion facilitate detecting and quantifying related impairments (e.g., jerkiness) using the time history of temporal features. The metric performs well in the VQEG FR-TV Phase II Test [275].

Some RR metrics for video are proposed based on specific features or properties of HVS in the spatial domain. In [87], an RR quality metric is designed based on a psychovisual color space from high-level human visual behavior for color video. The RR quality metric for video signals also takes advantage of contrast sensitivity [89]. The study in [92] incorporates the texture-masking property of the HVS. In [93], an RR quality metric is proposed based on features of SVD and HVS for wireless applications. An RR quality metric is proposed in [100], based on temporal motion smoothness. In [103], RR video quality metrics are proposed based on SSIM features.

Other RR metrics for video work directly in the encoded domain. In [81], blurring and blockiness from video compression are measured by a discriminative analysis of harmonic strength extracted from edge-detected images. In [80], RR quality metrics are proposed based on spatial and temporal features to measure the distortion occurring in standard video compression and communication systems. DCT coefficients are used to extract features for the perceptual quality evaluation of MPEG2-coded video in [82]. In [84], an RR quality metric is proposed based on multivariate data analysis to measure the artifacts of H.264/AVC video sequences. The RR quality metric in [107] also extracts features from DCT coefficients to measure the quality of distorted video. In [97], the differences between entropies of the wavelet coefficients of the reference and distorted video are calculated to measure the distortion of video signals.

3.5.1.3 NR Metrics

For NR video quality measurement, many studies build their metrics based on direct estimation of MSE or PSNR caused by specific block-based compression standards such as MPEG2, H.264, etc. [112, 114, 117, 118, 120]. In [112], the PSNR is calculated from the estimated quantization error caused by compression for visual quality evaluation. The study in [114] estimates PSNR based on DCT coefficients of MPEG2 video for visual quality evaluation. The transform coefficients are modeled by different distributions for visual quality evaluation such as a Gaussian model [118], Laplace model [117], and Cauchy distribution [120]. Some NR quality metrics have tried to measure the MSE caused by packet-loss errors [121, 124]. Bit-stream-based approaches predict the quality of video from the compressed video stream with packet losses [121]. The NR quality metric in [124] is designed to detect packet loss caused by specific compression of H.264 and motion-JPEG 2000, respectively. The noise-estimation-based NR quality metrics calculate the MSE based on the variation within certain smooth regions in visual signals [129]. Other NR quality metrics incorporate the characteristics of HVS for measuring the quality of visual content [139].

Beside the direct estimation of the MSE, some feature-based NR video quality metrics have been proposed for video quality assessment. The features in NR quality metrics for video signals can be extracted from different domains. In [27, 132], a blockiness artifact is measured based on features extracted from the spatial domain, while [134] evaluates visual quality for video based on features in the DCT domain. Additionally, in NR video quality metrics, various types of feature are used to calculate the distortion in video signals. In [152], the edge feature is extracted from visual signals to build NR quality metrics. In the NR quality metrics of [143, 144], the blurring feature is extracted from DCT coefficients. Some studies propose NR video quality metrics by combining blockiness, blurring, and ringing features together [150, 151]. Beside, some specific NR quality metrics for video are proposed to measure flicker [153] or frame freezes [154, 156]. In [281], an NR quality metric for HDTV is proposed to evaluate blockiness and blur distortions.

3.5.2 3D Video Quality Metrics

Recently, some studies have investigated quality metrics for the emerging applications of 3D video processing. The experimental results of the studies in [276,277] show that the 2D quality metrics can be used to evaluate the quality of 3D video content. The study in [190] discusses the importance of visual attention in 3DTV quality assessment. In [278], an FR stereo-video quality metric is proposed based on a monoscopic quality component and stereoscopic quality component. A 3D video quality metric is proposed based on the spatiotemporal structural information extracted from adjacent frames in [279]. Some studies also use characteristics of the HVS – including CSF, visual masking, and depth perception – to build perceptual 3D video quality metrics [28, 280]. Beside FR quality metrics, RR and NR quality metrics for 3D video quality evaluation have also been investigated in [78] and [133], respectively. However, 3D video quality measurement is still an open research area, because of the complexities of depth perception [246, 282].

3.6 Quality Metrics for Audio/Speech

Just like for images and video, traditional objective signal measures used for audio/speech quality assessment are built on basic mathematical measurements such as SNR, MSE, etc. They do not take psychoacoustic features of the HAS into consideration and thus cannot provide satisfactory performance compared with perceptual audio/speech quality assessment methods. Additionally, the shortcomings of traditional objective signal measures are evident in the non-linear and non-stationary codecs for audio/speech signals [3]. To overcome these drawbacks, various perceptual-based objective quality evaluation algorithms have been proposed based on characteristics of the HAS such as the perception of loudness, frequency, masking, etc. [3, 62, 177]. The amplitude of audio signals refers to the amplitude of the air pressure in the audio wave. The loudness is related to the amplitude of audio signals, which is perceived by listeners as the audio pressure level. The frequency of audio signals is measured in cycles per second (or Hz), and humans can perceive audio signals with frequencies in the range of 20 Hz to 20 kHz. Generally, the sensitivity of the HAS is frequency dependent. Auditory masking happens when the perception of one audio signal is affected by another audio signal. In the frequency domain, auditory masking is known as simultaneous masking, while it is known as temporal masking in the time domain.

Currently, most existing FR models (also called intrusive models) for audio/speech signals adopt perceptual models to transform both reference and distorted signals for feature extraction [63–66, 71, 178]. The quality of the distorted signal is estimated from the distance between features of the reference and distorted signals in transform domains. The NR models (also called non-intrusive models) estimate the quality of distorted speech signals without reference signals. Currently, there is no NR model for audio signals. Existing NR models for speech signals calculate the distortion results based on the signal production, signal likelihood, perception properties of noise loudness, etc. [62, 155, 283].

We are not aware of any RR metrics for audio/speech quality assessment in the literature.

3.6.1 FR Metrics

Studies of FR audio/speech quality metrics began with the requirement of low-bit-rate speech and audio codecs [62]. Since the 1970s, many studies have adopted perception-based models in speech/audio codecs to optimize the coding distortions for minimum audibility rather than MSE for improved perceived quality [30]. In [63], a noise-to-mask ratio measure was designed based on a perceptual masking model by comparing the level of the coding noise with the reference signal. Other similar waveform difference measures include the research work in [64,65]. The problem with these methods is that the estimated quality might be unreasonable for the distorted signals with substantial changes of signal waveform, which would result in large waveform differences [62]. To overcome this problem, researchers have tried to extract signal features in the transform domain, which is consistent with the hypothetical representation of the signal in the brain or peripheral auditory system. One successful approach is the auditory spectrum distance model [66], which is widely used in ITU standards [65, 71, 178]. In that model [66], the features of peripheral hearing in the time and frequency domain are extracted for quality evaluation based on psychoacoustic theory. The study in [67] adopts a model of HAS to calculate the internal representation of audio signals for quality evaluation based on the psychophysical domain. In these models, the time signals are first mapped into the time frequency domain, and then smeared and compressed to get two time-frequency loudness density functions [62]. These density functions are passed to a cognitive model interpreting their differences with possible substantial additional processing [62]. Generally, the cognitive model is trained by a large training database and should be validated by test data. These perceptual quality metrics show promising prediction performance for many aspects of psychoacoustic data due to the use of psychoacoustic theories [66,67]. Other studies try to improve the performance of existing metrics by using more detailed or advanced HAS models [17, 68–70, 73].

For audio quality assessment, the study in [284] calculates the probability of detected noise as a function of time for the coded audio signals. The study in [55] develops a model of the human ear in perceptual coding of audio signals. A frequency response equalization process is used in [179] for the quality assessment of audio signals. The study in [56] proposes an advanced quality metric based on a wide range of perceptual transformations. Some studies have tried to predict the perceived quality of audio signals based on the estimation of frontal spatial fidelity and surround spatial fidelity of multichannel audio [76], new distortion parameters and a cognitive model [57], and a multichannel expert system [58]. Recently, several perceptual objective metrics for audio signals have been proposed using an energy equalization approach [59,60] and mean structural similarity measure [18].

Speech quality assessment has an even longer history, with many metrics [61, 64, 66, 68, 69, 73–75, 285]. One early perceptual speech quality metric was proposed by Karjalainen based on the features of peripheral hearing in time and frequency known from psychoacoustic theory [66]. Later, a simple approach known as the Perceptual Speech Quality Measure (PSQM) was proposed for the standard ITU-T P.861 [65]. Different from earlier models, PSQM improved its salient interval processing, giving less emphasis to noise in silent periods than during speech, and its use of asymmetry weighting [62]. The drawback of PSQM and other early models is that they are trained on subjective tests of generic speech codecs, and thus their performance is poor with some types of telephone network [62]. To address this problem, some objective metrics for speech signals were proposed with specific telephone network conditions [74, 285, 286]. Several more recent FR speech quality metrics have been proposed based on Bayesian modeling [31], adaptive feedback canceller [32], etc.

3.6.2 NR Metrics

NR speech quality evaluation is more challenging due to the lack of reference signals. However, NR models are much more useful in practical applications such as wireless communications, voice over IP, and other in-service networks requiring speech quality monitoring, where the reference signal is unavailable. Currently, there is no NR quality metric for general audio signals. There are some studies trying to propose NR quality metrics for speech signals based on specific features.

Several NR speech quality metrics are designed based on specific distortions introduced by standard codecs or specific transmission networks. An early NR speech quality evaluation metric is built based on the spectrogram of the perceived signal for wireless communication [113]. The speech quality metric in [115] adopts Gaussian Mixture Models (GMMs) to create an artificial reference model to compare the degraded speech for quality evaluation; whereas in [119] speech quality is predicted by Bayesian inference, and Minimum Mean

Square Error (MMSE) estimation based on a trained set of GMMs. In [131], a perceptually motivated speech quality metric is presented based on a temporal envelope representation of speech. The study [123] proposes a low-complexity NR speech quality metric based on features extracted from commonly used speech coding parameters (e.g., spectral dynamics). The features are extracted globally and locally to design an NR speech quality metric in [109]. Machine learning techniques are adopted to predict the quality of distorted speech signals in [128].

There are also other studies developing NR quality metrics that assess the quality of noise-suppressed speech signals. An NR speech quality metric is proposed in [127] based on Kullback–Leibler distances for noise-suppressed speech signals. In [125], an NR speech quality metric is built for noise-suppressed speech signals based on mel-filtered energies and support vector regression.

3.7 Joint Audiovisual Quality Metrics

Generally, we watch video with an accompanying soundtrack. Therefore, comprehensive audiovisual quality metrics are required to analyze both modalities of multimedia content together. Audiovisual quality comprises two factors: synchronization between the two media signals (i.e., lip-sync) and interaction between audio and video quality [5, 44]. Currently, various research studies have been performed for audio/video synchronization. In actual lip-sync experiments, viewers perceive audio and video signals to be in sync up to about 80 ms of delay [287]. There is a consistently higher tolerance for video ahead of audio rather than vice versa, probably since this is also a more natural occurrence in the real world, where light travels faster than sound. Similar results were reported in experiments with non-speech clips showing a drummer [288]. The interaction between audio and video signals is another factor influencing the overall quality assessment of multimedia content, as shown by studies from neuroscience [289]. In [289], Lipscomb claims that at least two implicit judgments are made during the perceptual processing of the video experience: an association judgment and a mapping of accent structures. Based on the experimental results, the importance of synchronization decreases with more complicated audiovisual content for the interaction effect from audio and video signals [289].

Since most existing audiovisual quality metrics are proposed based on a combination of audio and video quality evaluation, the study in [44] analyzes the mutual influence between audio quality, video quality, and audiovisual quality. Based on the experimental analysis, the study obtains several general conclusions as follows. Firstly, both audio quality and video quality contribute to the overall audiovisual quality and their multiplication gets the highest correlation with the overall quality. Secondly, the overall quality is dominated by the video quality in general, whereas audio quality is more important than video quality in cases where the bit rates of both coded audio and video are low, or the video quality is larger than some certain threshold. With decreasing audio quality, the influence of audio quality increases in the overall quality. Additionally, with applications in which audio is obviously more important than video content (such as teleconference, news, music video, etc.), audio quality dominates the overall quality. Finally, audiovisual quality is also influenced by other factors, including motion information and complexity of the video content [44].

In [290], subjective experiments were carried out on audio, video, and audiovisual quality with results demonstrating that both audio and video quality contribute significantly to perceived audiovisual quality. The study also shows that the audiovisual quality can be evaluated with high accuracy by linear or bilinear combination from audio and video quality evaluation. Thus, many studies have adopted linear combination from audio and video quality evaluation to evaluate the quality of audio/video signals [291, 292].

Studies on audio/video quality metrics have focused mainly on low-bit-rate applications such as mobile communications, where the audio stream can use up a significance part of the total bit rate [293, 294]. Audio/video synchronization is incorporated, beside the fusion of audio and video quality in the audiovisual model proposed in [295]. Some studies focus on audiovisual quality evaluation for video conference applications [291, 292, 296]. The study in [297] presents a basic audiovisual quality metric based on subjective experiments on multimedia signals with simulated artifacts. The test data used in these studies is quite different in terms of content range and distortion, and these models obtain good prediction performance. In [142], an NR audiovisual quality metric is proposed to predict audiovisual quality and obtain good prediction performance. The study in [298] presents a graph-based perceptual audiovisual quality metric based on the contributions from modalities (audio and video) as well as the contribution of their relation. Some studies propose an audiovisual quality metric based on semantic analysis [299, 300].

Although there are some studies investigating audiovisual quality metrics, the progress of joint audiovisual quality assessment has been slow. The interaction between audio and video perception is complicated, and the perception of audiovisual content still lacks deep investigation. Currently, there are many quality metrics proposed based on the linear fusion of audio and video quality, but most studies choose fusion parameters empirically without theoretical support and little if any integration in the metric computation. However, audiovisual quality assessment is worthy of further investigation due to its wide application in signal coding, signal transmission, etc.

3.8 Concluding Remarks

Currently, traditional signal fidelity metrics are still widely used to evaluate the quality of multimedia content. However, perceptual quality metrics have shown promise in quality assessment, and a large number of perceptual quality assessment metrics have been proposed for various types of content, as introduced in this chapter. During the past ten years, some perceptual quality metrics have gained popularity and have been used in various signal-processing applications, such as SSIM. In the past, a lot of effort focused on designing FR metrics for audio or video. It is not easy to obtain good evaluation performance with RR or NR quality metrics. However, effective NR metrics are much desired, with more and more multimedia content (such as image, video, or music files) being distributed over the Internet today. The widely used Internet transmission and new compression standards bring many new challenges for multimedia quality evaluation, such as new types of transmission loss and compression distortions. Additionally, various emerging applications of 3D systems and displays require new quality metrics. Depth perception in particular should be investigated further for 3D quality evaluation. Other substantial quality evaluation topics include the quality assessment for super-resolution images/video and High Dynamic Range (HDR) images/video. All these emerging content types and their corresponding processing methods bring with them many challenges for multimedia quality evaluation.

References

[1] Chikkerur, S., Sundaram, V., Reisslein, M., and Karam, L.J., 'Objective video quality assessment methods: A classfication, review, and performance comparison.' *IEEE Transactions on Broadcasting*, **57**(2), 2011, 165–182.

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

[3] Campbell, D., Jones, E., and Glavin, M., 'Audio quality assessment techniques – a review and recent developments.' *Signal Processing*, **89**(8), 2009, 1489–1500.

[4] Winkler, S., *Digital Video Quality – Vision Models and Metrics*. John Wiley & Sons, Chichester, 2005.

[5] Winkler, S. and Mohandas, P., 'The evolution of video quality measurement: From PSNR to hybrid metrics.' *IEEE Transactions on Broadcasting*, **54**(3), 2008, 660–668.

[6] Eskicioglu, A.M. and Fisher, P.S., 'Image quality measures and their performance.' *IEEE Transactions on Communications*, **43**(12), 1995, 2959–2965.

[7] Karunasekera, S.A. and Kingsbury, N.G., 'A distortion measure for blocking artifacts in images based on human visual sensitivity.' *IEEE Transactions on Image Processing*, **4**(6), 1995, 713–724.

[8] Limb, J.O., 'Distortion criteria of the human viewer.' *IEEE Transactions on Systems, Man, and Cybernetics*, **9**(12), **1979**, 778–793.

[9] Girod, B., 'What's wrong with mean squared error?' In Watson, A.B. (ed.), *Digital Images and Human Vision*. MIT Press, Boston, MA, 1993, pp. 207–220.

[10] Mannos, J. and Sakrison, D., 'The effects of a visual fidelity criterion of the encoding of images.' *IEEE Transactions on Information Theory*, **20**(4), 1974, 525–536.

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

[12] Eckert, M.P. and Bradley, A.P., 'Perceptual quality metrics applied to still image compression.' *Signal Processing*, **70**, 1998, 177–200.

[13] Pappas, T.N. and Safranek, R.J., 'Perceptual criteria for image quality evaluation.' In Bovik, A.C. (ed.), *Handbook of Image and Video Processing*. Academic Press, New York, 2000, pp. 669–684.

[14] Video Quality Expert Group (VQEG), Final report from the video quality expert group on the validation of objective models of video quality assessment, March 2000. Available at: www.vqeg.org.

[15] Video Quality Expert Group (VQEG), Final report from the video quality expert group on the validation of objective models of video quality assessment, Phase II, August 2003. Available at: www.vqeg.org.

[16] Wang, Z., Bovik, A.C., and Lu, L., 'Why is image quality assessment so difficult?' IEEE International Conference on Acoustics, Speech, and Signal Processing, May 2002.

[17] Huber, R. and Kollmeier, B., 'PEMO-Q – A new method for objective audio quality assessment using a model of auditory perception.' *IEEE Transactions on Audio, Speech and Language Processing*, **14**(6), 2006, 1902–1911.

[18] Kandadai, S., Hardin, J., and Creusere, C.D., 'Audio quality assessment using the mean structural similarity measure.' IEEE International Conference on Acoustics, Speech, and Signal Processing, April 2008.

[19] Daly, S., 'The visible differences predictor: An algorithm for the assessment of image fidelity.' In Watson, A.B. (ed.), *Digital Images and Human Vision*. MIT Press, Cambridge, MA, 1993, pp. 179–206.

[20] Lubin, J., 'A visual discrimination model for imaging system design and evaluation.' In Peli, E. (ed.), *Vision Models for Target Detection and Recognition*. World Scientific, Singapore, 1995, pp. 245–283.

[21] Watson, A.B., Hu, J., and McGowan, J.F., 'DVQ: A digital video quality metric based on human vision.' *Journal of Electronic Imaging*, **10**(1), 2001, 20–29.

[22] Winkler, S., 'A perceptual distortion metric for digital color video.' Proceedings of SPIE no. 3644, 1999, pp. 175–184.

[23] Wolf, S., 'Measuring the end-to-end performance of digital video systems.' *IEEE Transactions on Broadcasting*, **43**(3), 1997, 320–328.

[24] Wang, Z., Bovik, A.C., and Evan, B.L., 'Blind measurement of blocking artifacts in images.' IEEE International Conference on Image Processing, September 2002.

[25] Miyahara, M., Kotani, K., and Algazi, V.R., 'Objective picture quality scale (PQS) for image coding.' *IEEE Transactions on Commununications*, **46**(9), 1998, 1215–1225.

[26] Marziliano, P., Dufaux, F., Winkler, S., and Ebrahimi, T., 'A no-reference perceptual blur metric.' IEEE International Conference on Image Processing, September 2002.

[27] Wu, H.R. and Yuen, M., 'A generalized block-edge impairment metric (GBIM) for video coding.' *IEEE Signal Processing Letters*, **4**(11), 1997, 317–320.

[28] Jin, L., Boev, A., Gotchev, A., and Egiazarian, K., '3D-DCT based perceptual quality assessment of stereo video.' IEEE International Conference on Image Processing, September 2011.

[29] Yu, Z., Wu, H.R., Winkler, S., and Chen, T., 'Vision-model-based impairment metric to evaluate blocking artifacts in digital video.' *Proceedings of the IEEE*, **90**, 2002, 154–169.

[30] Schroeder, M.R., Atal, B.S., and Hall, J.L., 'Optimizing digital speech coders by exploiting masking properties of the human ear.' *Journal of the Acoustical Society of America*, **66**(6), 1979, 1647–1652.

[31] Chen, G. and Parsa, V., 'Loudness pattern-based speech quality evaluation using Bayesian modeling and Markov chain Monte Carlo methods.' *Journal of the Acoustical Society of America*, **121**(2), 2007, EL77–EL83.

[32] Manders, A.J., Simpson, D.M., and Bell, S.L., 'Objective prediction of the sound quality of music processed by an adaptive feedback canceller.' *IEEE Transactions on Audio, Speech, and Language Processing*, **20**(6), 2012, 1734–1745.

[33] Chandler, D.M. and Hemami, S.S., 'VSNR: A wavelet-based visual signal-to-noise ratio for natural images.' *IEEE Transactions on Image Processing*, **16**(9), 2007, 2284–2298.

[34] Sheikh, H.R., Bovik, A.C., and de Veciana, G., 'An information fidelity criterion for image quality assessment using natural scene statistics.' *IEEE Transactions on Image Processing*, **14**(12), 2005, 2117–2128.

[35] Sheikh, H.R. and Bovik, A.C., 'Image information and visual quality.' *IEEE Transactions on Image Processing*, **15**(2), 2006, 430–444.

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

[37] Horita, Y., Miyata, T., Gunawan, I.P., Murai, T., and Ghanbari, M., 'Evaluation model considering static-temporal quality degradation and human memory for SSCQE video quality.' *Proceedings of SPIE: Visual Communications and Image Processing*, **5150**(11), 2003, 1601–1611.

[38] Terhardt, E., 'Calculating virtual pitch.' *Hearing Research*, **1**, 1979, 155–182.

[39] Dijk, J., van Grinkel, M., van Asselt, R.J., van Vliet, L.J., and Verbeek, P.W., 'A new sharpness measure based on Gaussian lines and edges.' Proceedings of the International Conference on Computational Analysis of Images and Patterns (CAIP). Lecture Notes in Computer Science, Vol. 2756. Springer-Verlag, Berlin, 2003, pp. 149–156.

[40] Ong, E., Lin, W., Lu, Z., Yao, S., Yang, X., and Jiang, L., 'No reference JPEG-2000 image quality metric.' Proceedings of IEEE International Conference Multimedia and Expo (ICME), 2003, pp. 545–548.

[41] Muijs, R. and Kirenko, I., 'A no-reference block artifact measure for adaptive video processing.' EUSIPCO 2005.

[42] Faugeras, O.D., 'Digital color image processing within the framework of a human visual model.' *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **27**, 1979, 380–393.

[43] Lukas, F., and Budrikis, Z., 'Picture quality prediction based on a visual model.' *IEEE Transactions on Communications*, **30**, 1982, 1679–1692.

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

[45] Tong, X., Heeger, D., and Lambrecht, C.V.D.B., 'Video quality evaluation using STCIELAB.' *Proceedings of SPIE: Human Vision, Visual Processing and Digital Display*, **3644**, 1999, 185–196.

[46] Winkler, S., 'Vision models and quality metrics for image processing applications.' Swiss Federal Institute of Technology, Thesis 2313, December 2000, Lausanne, Switzerland.

[47] Sarnoff Corporation. 'Sarnoff JND vision model.' In Lubin, J. (ed.), Contribution to IEEE G-2.1.6 Compression and Processing Subcommittee, 1997.

[48] Liu, A., Lin, W., Paul, M., Deng, C., and Zhang, F., 'Just noticeable difference for images with decomposition model for separating edge and textured regions.' *IEEE Transactions on Circuits and Systems for Video Technology*, **20**(11), 2010, 1648–1652.

[49] Ong, E., Lin, W., Lu, Z., Yao, S., and Etoh, M., 'Visual distortion assessment with emphasis on spatially transitional regions.' *IEEE Transactions on Circuits and Systems for Video Technology*, **14**(4), 2004, 559–566.

[50] Masry, M.A., Hemami, S.S., and Sermadevi, Y., 'A scalable wavelet-based video distortion metric and applications.' *IEEE Transactions on Circuits and Systems for Video Technology*, **16**(2), 2006, 260–273.

[51] Pinson, M.H. and Wolf, S., 'A new standardized method for objectively measuring video quality.' *IEEE Transactions on Broadcasting*, **50**(3), 2004, 312–322.

[52] Watson, A.B., 'DCTune: A technique for visual optimization of DCT quantization matrices for individual images.' Society for Information Display Digest of Technical Papers, Vol. XXIV, 1993, pp. 946–949.

[53] Lin, W., Dong, L., and Xue, P., 'Visual distortion gauge based on discrimination of noticeable contrast changes.' *IEEE Transactions on Circuits and Systems for Video Technology*, **15**(7), 2005, 900–909.

[54] Wang, Z., and Bovik, A.C., 'A universal image quality index.' *IEEE Signal Processing Letters*, **9**(3), 2002, 81–84.

[55] Colomes, C., Lever, M., Rault, J.B., and Dehery, Y.F., 'A perceptual model applied to audio bit-rate reduction.' *Journal of the Audio Engineering Society*, **43**(4), 1995, 233–240.

[56] Thiede, T., Treurniet, W.C., Bitto, R., *et al.*, 'PEAQ – The ITU standard for objective measurement of perceived audio quality.' *Journal of the Audio Engineering Society*, **48**(1/2), 2000, 3–29.

[57] Barbedo, J. and Lopes, A., 'A new cognitive model for objective assessment of audio quality.' *Journal of the Audio Engineering Society*, **53**(1/2), 2005, 22–31.

[58] Zielinski, S., Rumsey, F., Kassier, R., and Bech, S., 'Development and initial validation of a multichannel audio quality expert system.' *Journal of the Audio Engineering Society*, **53**(1/2), 2005, 4–21.

[59] Vanam, R., and Creusere, C., 'Evaluating low bitrate scalable audio quality using advanced version of PEAQ and energy equalization approach.' Proceedings of IEEE ICASSP, Vol. 3, 2005, pp. 189–192.

[60] Creusere, C., Kallakuri, K., and Vanam, R., 'An objective metric of human subjective audio quality optimized for a wide range of audio fidelities.' *IEEE Transactions on Audio, Speech, and Language Processing*, **16**(1), 2008, 129–136.

[61] Novorita, B., 'Incorporation of temporal masking effects into Bark spectral distortion measure.' Proceedings of IEEE ICASSP, Vol. 2, 1999, pp. 665–668.

[62] Rix, A.W., Beerends, J.G., Kim, D., Kroon, P., and Ghitza, O., 'Objective assessment of speech and audio quality-technology and applications.' *IEEE Transactions on Audio, Speech, and Language Processing*, **14**(6), 2006, 1890–1901.

[63] Brandenburg, K., 'Evaluation of quality for audio encoding at low bit rates.' Proceedings of 82nd Audio Engineering Society Convention, 1987, preprint 2433.

[64] Quackenbush, S.R., Barnwell, T.P., and Clements, M.A., *Objective Measures of Speech Quality*. Prentice-Hall, Englewood Cliffs, NJ, 1988.

[65] 'Objective quality measurement of telephone-band (300–3400 Hz) speech codecs.' ITU-T P.861, 1998.

[66] Karjalainen, M., 'A new auditory model for the evaluation of sound quality of audio system.' Proceedings of IEEE ICASSP, 1985, pp. 608–611.

[67] Beerends, J.G. and Stemerdink, J.A., 'A perceptual audio quality measure based on a psychoacoustic sound representation. *Journal of the Audio Engineering Society*, **40**(12), 1992, 963–974.

[68] Hansen, M. and Kollmeier, B., 'Using a quantitative psycho-acoustical signal representation for objective speech quality measurement.' Proceedings of ICASSP, 1997, pp. 1387–1390.

[69] Hauenstein, M., 'Application of Meddis' inner hair-cell model to the prediction of subjective speech quality.' Proceedings of IEEE ICASSP, 1998, pp. 545–548.

[70] Moore, B.C.J., Tan, C.-T., Zacharov, N., and Mattila, V.-V., 'Measuring and predicting the perceived quality of music and speech subjected to combined linear and nonlinear distortion.' *Journal of the Audio Engineering Society*, **52**(12), 2004, 1228–1244.

[71] 'Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs.' ITU-T P.862, 2001.

[72] Beerends, J.G. and Stemerdink, J.A., 'The optimal time-frequency smearing and amplitude compression in measuring the quality of audio devices.' Proceedings of 94th Audio Engineering Society Convention, 1993.

[73] Ghitza, O., ‘Auditory models and human performance in tasks related to speech coding and speech recognition.’ *IEEE Transactions on Speech and Audio Processing*, **2**(1), 1994, 115–132.

[74] Rix, A.W. and Hollier, M.P., ‘The perceptual analysis measurement system for robust end-to-end speech quality assessment.’ Proceedings of IEEE ICASSP, Vol. 3, 2000, pp. 1515–1518.

[75] Beerends, J.G. and Stemerdink, J.A., ‘A perceptual speech quality measure based on a psychoacoustic sound representation.’ *Journal of the Audio Engineering Society*, **42**(3), 1994, 115–123.

[76] George, S., Zielinski, S., and Rumsey, F., ‘Feature extraction for the prediction of multichannel spatial audio fidelity.’ *IEEE Transactions on Audio, Speech, and Language Processing*, **14**(6), 2006, 1994–2005.

[77] Rehman, A., and Wang, Z., ‘Reduced-reference image quality assessment by structural similarity estimation.’ *IEEE Transactions on Image Processing*, **21**(8), 2012, 3378–3389.

[78] Hewage, C.T.E.R. and Martini, M.G., ‘Reduced-reference quality assessment for 3D video compression and transmission.’ *IEEE Transactions on Consumer Electronics*, **57**(3), 2011, 1185–1193.

[79] Kusuma, T.M. and Zepernick, H.-J., ‘A reduced-reference perceptual quality metric for in-service image quality assessment.’ Proceedings of 1st Workshop on Mobile Future and Symposium on Trends in Communications, October 2003, pp. 71–74.

[80] Wolf, S. and Pinson, M.H., ‘Spatio-temporal distortion metrics for in-service quality monitoring of any digital video system.’ Proceedings of SPIE, Vol. 3845, 1999, pp. 266–277.

[81] Gunawan, I. and Ghanbari, M., ‘Reduced-reference video quality assessment using discriminative local harmonic strength with motion consideration.’ *IEEE Transactions on Circuits and Systems for Video Technology*, **18**(1), 2008, 71–83.

[82] Yang, S., ‘Reduced reference MPEG-2 picture quality measure based on ratio of DCT coefficients.’ *Electronics Letters*, **47**(6), 2011, 382–383.

[83] Altous, S., Samee, M.K., and Gotze, J., ‘Reduced reference image quality assessment for JPEG distortion.’ ELMAR Proceedings, September 2011, pp. 97–100.

[84] Oelbaum, T. and Diepold, K., ‘Building a reduced reference video quality metric with very low overhead using multivariate data analysis.’ *Journal of Systemics, Cybernetics, and Informatics*, **6**(5), 2008, 81–86.

[85] Wang, Z. and Bovik, A.C., ‘Reduced and no reference visual quality assessment – the natural scene statistic model approach.’ *IEEE Signal Processing Magazine, Special Issue on Multimedia Quality Assessment*, **29**(6), 2011, 29–40.

[86] Carnec, M., Le Callet, P., and Barba, D., ‘Visual features for image quality assessment with reduced reference.’ Proceedings of IEEE International Conference on Image Processing, Vol. 1, September 2005, pp. 421–424.

[87] Le Callet, P., Viard-Gaudin, C., and Barba, D., ‘Continuous quality assessment of MPEG2 video with reduced reference.’ Proceedings of International Workshop on Video Processing Quality Metrics for Consumer Electronics, Scottsdale, AZ, January 2005.

[88] Carnec, M., Le Callet, P., and Barba, D., ‘Objective quality assessment of color images based on a generic perceptual reduced reference.’ *Signal Processing: Image Communication*, **23**(4), 2008, 239–256.

[89] Amirshahi, S.A. and Larabi, M., ‘Spatial-temporal video quality metric based on an estimation of QoE.’ Third International Workshop on Quality of Multimedia Experience (QoMEX), September 2011, pp. 84–89.

[90] Tao, D., Li, X., Lu, W., and Gao, X., ‘Reduced-reference IQA in contourlet domain.’ *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, **39**(6), 2009, 1623–1627.

[91] Engelke, U., Kusuma, M., Zepernick, H.-J., and Caldera, M., ‘Reduced-reference metric design for objective perceptual quality assessment in wireless imaging.’ *Signal Processing: Image Communication*, **24**(7), 2009, 525–547.

[92] Ma, L., Li, S., and Ngan, K.N., ‘Reduced-reference video quality assessment of compressed video sequences.’ *IEEE Transactions on Circuits and Systems for Video Technology*, **22**(10), 2012, 1441–1456.

[93] Yuan, F. and Cheng, E., ‘Reduced-reference metric design for video quality measurement in wireless application.’ 11th IEEE International Conference on Communication Technology (ICCT), November 2008, pp. 641–644.

[94] Zhai, G., Zhang, W., Yang, X., and Xu, Y., ‘Image quality assessment metrics based on multi-scale edge presentation.’ IEEE Workshop on Signal Processing Systems Design and Implementation, November 2005, pp. 331–336.

[95] Narwaria, M., Lin, W., McLoughlin, I.V., Emmanuel, S., and Chia, L.-T., ‘Fourier transform-based scalable image quality measure.’ *IEEE Transactions on Image Processing*, **21**(8), 2012, 3364–3377.

[96] Chono, K., Lin, Y.-C., Varodayan, D., Miyamoto, Y., and Girod, B., ‘Reduced-reference image quality assessment using distributed source coding.’ IEEE International Conference on Multimedia and Expo, April 2008, pp. 609–612.

[97] Soundararajan, R. and Bovik, A.C., ‘Video quality assessment by reduced reference spatio-temporal entropic differencing.’ *IEEE Transactions on Circuits and Systems for Video Technology*, **23**(4), 2013, 684–694.

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

[99] Redi, J.A., Gastaldo, P., Heynderickx, I., and Zunino, R., ‘Color distribution information for the reduced-reference assessment of perceived image quality.’ *IEEE Transactions on Circuits and Systems for Video Technology*, **20**(12), 2010, 1757–1769.

[100] Zeng, K. and Wang, Z., ‘Temporal motion smoothness measurement for reduced-reference video quality assessment.’ IEEE International Conference on Acoustics, Speech, and Signal Processing, March 2010, pp. 1010–1013.

[101] Cheng, G. and Cheng, L., ‘Reduced reference image quality assessment based on dual derivative priors.’ *Electronics Letters*, **45**(18), 2009, 937–939.

[102] Ma, L., Li, S., Zhang, F., and Ngan, K.N., ‘Reduced-reference image quality assessment using reorganized DCT-based image representation.’ *IEEE Transactions on Multimedia*, **13**(4), 2011, 824–829.

[103] Albonico, A., Valenzise, G., Naccari, M., Tagliasacchi, M., and Tubaro, S., ‘A reduced-reference video structural similarity metric based on no-reference estimation of channel-induced distortion.’ IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), April 2009, pp. 1857–1860.

[104] Wang, X., Jiang, G., and Yu, M., ‘Reduced reference image quality assessment based on contourlet domain and natural image statistics.’ 5th International Conference on Image and Graphics (ICIG), September 2009, pp. 45–50.

[105] Wang, Z. and Simoncelli, E.P., ‘Reduced-reference image quality assessment using a wavelet-domain natural image statistic model.’ Proceedings of SPIE: Human Vision and Electronic Imaging X, Vol. 5666, January 2005.

[106] Li, Q. and Wang, Z., ‘Reduced-reference image quality assessment using divisive normalization-based image representation.’ *IEEE Journal on Selected Topics in Signal Processing*, **3**(2), 2009, 202–211.

[107] Atzori, L., Ginesu, G., Giusto, D.D., and Floris, A., ‘Streaming video over wireless channels: Exploiting reduced-reference quality estimation at the user-side.’ *Signal Processing: Image Communication*, **27**(10), 2012, 1049–1065.

[108] Atzori, L., Ginesu, G., Giusto, D.D., and Floris, A., ‘Rate control based on reduced-reference image quality estimation for streaming video over wireless channels.’ IEEE International Conference on Communications (ICC), June 2012, pp. 2021–2025.

[109] Audhkhasi, K. and Kumar, A., ‘Two scale auditory feature based nonintrusive speech quality evaluation.’ *IETE Journal of Research*, **56**(2), 2010, 111–118.

[110] Hemami, S. and Reibman, A., ‘No-reference image and video quality estimation: Applications and human-motivated design.’ *Signal Processing: Image Communication*, **25**(7), 2010, 469–481.

[111] Narwaria, M. and Lin, W., ‘SVD-based quality metric for image and video using machine learning.’ *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, **42**(2), 2012, 347–364.

[112] Turaga, D.S., Chen, Y., and Caviedes, J., ‘No reference PSNR estimation for compressed pictures.’ *Signal Processing: Image Communication*, **19**, 2004, 173–184.

[113] Au, O.L. and Lam, K., ‘A novel output-based objective speech quality measure for wireless communication.’ Proceedings of 4th International Conference on Signal Processing, Vol. 1, 1998, pp. 666–669.

[114] Ichigaya, A., Nishida, Y., and Nakasu, E., ‘Non reference method for estimating PSNR of MPEG-2 coded video by using DCT coefficients and picture energy.’ *IEEE Transactions on Circuits and Systems for Video Technology*, **18**(6), 2008, 817–826.

[115] Falk, T., Xu, Q., and Chan, W.Y., ‘Non-intrusive GMM-based speech quality measurement.’ Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005, pp. 125–128.

[116] Brandao, T. and Queluz, M.P., ‘No-reference image quality assessment based on DCT domain statistics.’ *Signal Processing*, **88**, 2008, 822–833.

[117] Eden, A., ‘No-reference estimation of the coding PSNR for H.264-coded sequences.’ *IEEE Transactions on Consumer Electronics*, **53**(2), 2007, 667–674.

[118] Choe, J. and Lee, C., 'Estimation of the peak signal-to-noise ratio for compressed video based on generalized Gaussian modelling.' *Optical Engineering*, **46**(10), 2007, 107401.

[119] Chen, G. and Parsa, V., 'Bayesian model based non-intrusive speech quality evaluation.' Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005, pp. 385–388.

[120] Shim, S.-Y., Moon, J.-H., and Han, J.-K., 'PSNR estimation scheme using coefficient distribution of frequency domain in H.264 decoder.' *Electronics Letters*, **44**(2), 2008, 108–109.

[121] Reibman, A.R., Vaishampayan, V.A., and Sermadevi, Y., 'Quality monitoring of video over a packet network.' *IEEE Transactions on Multimedia*, **6**(2), 2004, 327–334.

[122] Tobias, J., *Foundations of Modern Auditory Theory*. Academic Press, New York, 1970.

[123] Grancharov, V., David, Y., Jonas, L., and Bastiaan, W., 'Low complexity nonintrusive speech quality assessment.' *IEEE Transactions on Speech and Audio Processing*, **14**(6), 2006, 1948–1956.

[124] Nishikawa, K., Munadi, K., and Kiya, H., 'No-reference PSNR estimation for quality monitoring of motion JPEG2000 video over lossy packet networks.' *IEEE Transactions on Multimedia*, **10**(4), 2008, 637–645.

[125] Narwaria, M., Lin, W., McLoughlin, I.V., Emmanuel, S., and Chia, L.-T., 'Nonintrusive quality assessment of noise suppressed speech with mel-filtered energies and support vector regression.' *IEEE Transactions on Audio, Speech, and Language Processing*, **20**(4), 2012, 1217–1232.

[126] Li, X., 'Blind image quality assessment.' IEEE International Conference on Image Processing, 2002.

[127] Falk, T., Yuan, H., and Chan, W.Y., 'Single-ended quality measurement of noise suppressed speech based on Kullback–Leibler distances.' *Journal of Multimedia*, **2**(5), 2007, 19–26.

[128] Falk, T. and Chan, W.Y., 'Single-ended speech quality measurement using machine learning methods.' *IEEE Transactions on Audio, Speech, and Language Processing*, **14**(6), 2006, 1935–1947.

[129] Kayargadde, V. and Martens, J.-B., 'An objective measure for perceived noise.' *Signal Processing*, **49**(3), 1996, 187–206.

[130] Jesteadt, W., Wier, C., and Green, D., 'Intensity discrimination as a function of frequency and sensation level.' *Journal of the Acoustical Society of America*, **61**(1), 1977, 169–177.

[131] Painter, T. and Spanias, A., 'Perceptual coding of digital audio.' *Proceedings of the IEEE*, **88**(4), 2000, 451–513.

[132] Suthaharan, S., 'A perceptually significant block-edge impairment metric for digital video coding.' Proceedings of International Conference on Acoustics, Speech, and Signal Processing, 2003, pp. III-681–III-684.

[133] Ha, K. and Kim, M., 'A perceptual quality assessment metric using temporal complexity and disparity information for stereoscopic video.' IEEE International Conference on Image Processing, September 2011, pp. 2525–2528.

[134] Liu, S. and Bovik, A.C., 'Efficient DCT-domain blind measurement and reduction of blocking artifacts.' *IEEE Transactions on Circuits and Systems for Video Technology*, **12**(12), 2002, 1139–1149.

[135] Zhai, G., Zhang, W., Yang, X., Lin, W., and Xu, Y., 'No-reference noticeable blockiness estimation in images.' *Signal Processing: Image Communication*, **23**, 2008, 417–432.

[136] Meesters, L. and Martens, J.-B., 'A single-ended blockiness measure for JPEG-coded images.' *Signal Processing*, **82**, 2002, 369–387.

[137] Ferzli, R. and Karam, L.J., 'A no-reference objective image sharpness metric based on the notion of just noticeable blur JNB.' *IEEE Transactions on Image Processing*, **18**(4), 2009, 717–728.

[138] Marziliano, P., Dufaux, F., Winkler, S., and Ebrahimi, T., 'Perceptual blur and ringing metrics: Application to JPEG2000.' *Signal Processing: Image Communication*, **19**, 2004, 163–172.

[139] Kanumuri, S., Cosman, P.C., Reibman, A.R., and Vaishampayan, V.A., 'Modeling packet-loss visibility in MPEG-2 video.' *IEEE Transactions on Multimedia*, **8**(2), 2006, 341–355.

[140] Ong, E., Lin, W., Lu, Z., Yang, X., Yao, S., Jiang, L., and Moschetti, F., 'A no-reference quality metric for measuring image blur.' IEEE International Symposium on Signal Processing and its Applications, 2003, pp. 469–472.

[141] Gu, K., Zhai, G., Yang, X., and Zhang, W., 'No-reference stereoscopic IQA approach: From nonlinear effect to parallax compensation.' *Journal of Electrical and Computer Engineering*, 2012, 1.

[142] Winkler, S. and Faller, C., 'Audiovisual quality evaluation of low-bitrate video.' Proceedings of SPIE Human Vision and Electronic Imaging, Vol. 5666, January 2005, pp. 139–148.

[143] Marichal, X., Ma, W.-Y., and Zhang, H.-J., 'Blur determination in the compressed domain using DCT information.' IEEE International Conference on Image Processing, 1999, pp. 386–390.

[144] Yang, K.-C., Guest, C.C., and Das, P.K., ‘Perceptual sharpness metric (PSM) for compressed video.’ IEEE International Conference on Multimedia and Expo, 2006.

[145] Blanchet, G., Moisan, L., and Rouge, B., ‘Measuring the global phase coherence of an image.’ IEEE International Conference on Image Processing, 2008, pp. 1176–1179.

[146] Feng, X. and Allebach, J.P., ‘Measurement of ringing artifacts in JPEG images.’ SPIE, Vol. 6076, 2006.

[147] Liu, H., Klomp, N., and Heynderickx, I., ‘A no-reference metric for perceived ringing.’ International Workshop on Video Processing and Quality Metrics, 2009.

[148] Sheikh, H.R., Bovik, A.C., and Cormak, L., ‘No-reference quality assessment using natural scene statistics: JPEG 2000.’ *IEEE Transactions on Image Processing*, **14**(11), 2005, 1918–1927.

[149] Susstrunk, S.E. and Winkler, S., ‘Color image quality on the Internet.’ SPIE, Vol. 5304, 2004.

[150] Davis, A.G., Bayart, D., and Hands, D.S., ‘Hybrid no-reference video quality prediction.’ IEEE International Symposium on Broadband Multimedia Systems, 2009.

[151] Hands, D., Bayart, D., Davis, A., and Bourret, A., ‘No reference perceptual quality metrics: Approaches and limitations.’ Human Vision and Electronic Imaging XIV, 2009.

[152] Engelke, U. and Zepernick, H.-J., ‘Pareto optimal weighting of structural impairments for wireless imaging quality assessment.’ IEEE International Conference on Image Processing, 2008, pp. 373–376.

[153] Kuszpet, Y., Kletsel, D., Moshe, Y., and Levy, A., ‘Post-processing for flicker reduction in H.264/AVC.’ Picture Coding Symposium, 2007.

[154] Pastrana-Vidal, R.R. and Gicquel, J.-C., ‘Automatic quality assessment of video fluidity impairments using a no-reference metric.’ International Workshop on Video Processing and Quality Metrics, 2006.

[155] ‘Single-ended method for objective speech quality assessment in narrow-band telephony applications.’ ITU-T P.563, 2004.

[156] Yang, K.-C., Guest, C.C., El-Maleh, K., and Das, P.K., ‘Perceptual temporal quality metric for compressed video.’ *IEEE Transactions on Multimedia*, **9**(7), 2007, 1528–1535.

[157] Bradley, A.P. and Stentiford, F.W.M., ‘Visual attention for region of interest coding in JPEG 2000.’ *Journal of Visual Communication and Image Representation*, **14**(3), 2003, 232–250.

[158] Wolfgang, R.B., Podilchuk, C.I., and Delp, E.J., ‘Perceptual watermarks for digital images and video.’ *Proceedings of IEEE*, **87**(7), 1999, 1108–1126.

[159] Frossard, P. and Verscheure, O., ‘Joint source/FEC rate selection for quality-optimal MPEG-2 video delivery.’ *IEEE Transactions on Image Processing*, **10**(12), 2001, 1815–1825.

[160] Ramasubramanian, M., Pattanaik, S.N., and Greenberg, D.P., ‘A perceptual based physical error metric for realistic image synthesis.’ *Computer Graphics* (SIGGRAPH ‘99 Conference Proceedings), 33(4), 1999, 73–82.

[161] Wandell, B., *Foundations of Vision*. Sinauer Associates, Sunderland, MA, 1995.

[162] Kelly, D.H., ‘Motion and vision II: Stabilized spatiotemporal threshold surface.’ *Journal of the Optical Society of America*, **69**(10), 1979, 1340–1349.

[163] Moorthy, A.K. and Bovik, A.C., ‘A survey on 3D quality of experience and 3D quality assessment.’ SPIE Proceedings: Human Vision and Electronic Imaging, 2013.

[164] Legge, G.E. and Foley, J.M., ‘Contrast masking in human vision.’ *Journal of the Optical Society of America*, **70**, 1980, 1458–1471.

[165] Yang, X., Lin, W., Lu, Z., Ong, E., and Yao, S., ‘Motion-compensated residue preprocessing in video coding based on just-noticeable-distortion profile.’ *IEEE Transactions on Circuits and Systems for Video Technology*, **15**(6), 2005, 742–750.

[166] Poirson, A.B. and Wandell, B.A., ‘Pattern-color separable pathways predict sensitivity to simple colored patterns.’ *Visible Research*, **36**(4), 1996, 515–526.

[167] Zhang, X. and Wandell, B.A., ‘Color image fidelity metrics evaluated using image distortion maps.’ *Signal Processing*, **70**(3), 1998, 201–214.

[168] Yang, X., Lin, W., Lu, Z., Ong, E., and Yao, S., ‘Just noticeable distortion model and its applications in video coding.’ *Signal Processing: Image Communication*, **20**(7), 2005, 662–680.

[169] Wu, J., Lin, W., Shi, G., and Liu, A., ‘Perceptual quality metric with internal generative mechanism.’ *IEEE Transactions on Image Processing*, **22**(1), 2013, 43–54.

[170] Winkler, S., ‘Quality metric design: A closer look.’ SPIE Proceedings: Human Vision and Electronic Imaging Conference, Vol. 3959, 2000, pp. 37–44.

[171] Yuen, M. and Wu, H.R., 'A survey of MC/DPCM/DCT video coding distortions.' *Signal Processing*, **70**(3), 1998, 247–278.

[172] Frederickson, R.E. and Hess, R.F., 'Estimating multiple temporal mechanisms in human vison.' *Vision Research*, **38**(7), 1998, 1023–1040.

[173] Daugman, J.G., 'Two-dimensional spectral analysis of cortical receptive field profiles.' *Vision Research*, **20**(10), 1980, 847–856.

[174] Watson, A.B., 'The cortex transform: Rapid computation of simulated neural images.' *Computer Visual Graphics and Imaging Processes*, **39**(3), 1987, 311–327.

[175] Burt, P.J. and Adelson, E.H., 'The Laplacian pyramid as a compact image code.' *IEEE Transactions on Communications*, **31**(4), 1983, 532–540.

[176] Simoncelli, E.P., Freeman, W.T., Adelson, E.H., and Heeger, D.J., 'Shiftable multi-scale transforms.' *IEEE Transactions on Information Theory*, **38**(2), 1992, 587–607.

[177] Moore, B.C.J., *An Introduction to the Psychology of Hearing*, 4th edn. Academic Press, Norwell, MA, 1997.

[178] 'Method for objective measurements of perceived audio quality.' ITU-R BS.1387, 1999.

[179] Thiede, T. and Kabot, E., 'A new perceptual quality measure for bit rate reduced audio.' Proceedings of 100th Audio Engineering Society Convention, 1996.

[180] Thiede, T., 'Perceptual audio quality assessment using a non-linear filter bank.' PhD Thesis, Fachbereich Electrotechnik, Technical University of Berlin, 1999.

[181] Hubel, D.H., *Eye, Brain, and Vision*. W.H. Freeman, New York, 1988.

[182] Elder, J.H. and Zucker, S.W., 'Local scale control for edge detection and blur estimation.' *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **20**(7), 1998, 699–716.

[183] Quan, H.-T. and Ghanbari, M., 'Temporal aspect of perceived quality of mobile video broadcasting.' *IEEE Transactions on Broadcasting*, **54**(3), 2008, 641–651.

[184] Verscheure, O., Frossard, P., and Hamdi, M., 'User-oriented QoS analysis in MPEG-2 delivery.' *Real-Time Imaging*, **5**(5), 1999, 305–314.

[185] Liang, J. and Kubichek, R., 'Output-based objective speech quality.' Proceedings of IEEE Vehicular Technology Conference, Stockholm, Sweden, 1994, pp. 1719–1723.

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

[187] Larson, E.C. and Chandler, D.M., 'Most apparent distortion: Full reference image quality assessment and the role of strategy.' *Journal of Electronic Imaging*, **19**(1), 2010, 011006-1–011006-21.

[188] Narwaria, M., Lin, W., and Çetin, A.E., 'Scalable image quality assessment with 2D mel-cepstrum and machine learning approach.' *Pattern Recognition*, **45**(1), 2012, 299–313.

[189] Liu, T.-J., Lin, W., and Kuo, C.-C.J., 'Image quality assessment using multi-method fusion.' *IEEE Transactions on Image Processing*, **22**(5), 2013, 1793–1807.

[190] Huynh-Tuh, Q., Barkowsky, M., and Le Callet, P., 'The importance of visual attention in improving the 3D-TV viewing experience: Overview and new perspectives.' *IEEE Transactions on Broadcasting*, **57**(2), 2011, 421–431.

[191] Peli, E., 'Contrast in complex images.' *Journal of the Optical Society of America*, **7**(10), 1990, 2032–2040.

[192] Winkler, S. and Vandergheynst, P., 'Computing isotropic local contrast from oriented pyramid decompositions.' Proceedings of International Conference on Image Processing, 1999, pp. 420–424.

[193] Lai, Y.-K. and Kuo, C.-C.J., 'A Haar wavelet approach to compressed image quality measurement.' *Journal of Visual Communication and Image Representation*, **11**(1), 2000, 17–40.

[194] Lu, Z., Lin, W., Yang, X., Ong, E., and Yao, S., 'Modeling visual attention's modulatory aftereffects on visual sensitivity and quality evaluation.' *IEEE Transactions on Image Processing*, **14**(11), 2005, 1928–1942.

[195] Ong, E., Yang, X., Lin, W., *et al.*, 'Perceptual quality and objective quality measurements of compressed videos.' *Journal of Visual Communication and Image Representation*, **17**(4), 2006, 717–737.

[196] Tan, K.T. and Ghanbari, M., 'Blockiness detection for MPEG2-coded video.' *IEEE Signal Processing Letters*, **7**(8), 2000, 213–215.

[197] Kundur, D. and Hatzinakos, D., 'Blind image deconvolutions.' *IEEE Signal Processing Magazine*, **13**, 1996, 43–63.

[198] Wu, S., Lin, W., Xie, S., Lu, Z., Ong, E., and Yao, S., 'Blind blur assessment for vision based applications.' *Journal of Visual Communication and Image Representation*, **20**(4), 2009, 231–241.

[199] Winkler, S., 'Visual fidelity and perceived quality: Towards comprehensive metrics.' *Proceedings of SPIE*, **4299**, 2001, 114–125.

[200] Pastrana-Vidal, R., Gicquel, J., Colomes, C., and Cherifi, H., 'Sporadic frame dropping impact on quality perception.' SPIE Proceedings: The International Society for Optical Engineering, Vol. 5292, 2004.

[201] Lin, W., 'Computational models for just-noticeable difference.' In Wu, H.R. and Rao, K.R. (eds), *Digital Video Image Quality and Perceptual Coding*. CRC Press, Boca Raton, FL, 2006.

[202] Lu, Z., Lin, W., Boon, C.S., Kato, S., Ong, E., and Yao, S., 'Perceptual quality evaluation on periodic frame-dropping video.' IEEE International Conference on Image Processing (ICIP), 2007.

[203] Montenovo, M., Perot, A., Carli, M., Cicchetti, P., and Neri, A., 'Objective quality evaluation of video services.' Third International Workshop on Video Processing and Quality Metrics for Consumer Electronics, January 2006.

[204] Suresh, N., Jayant, N., and Yang, O., 'Mean time between failures: A subjectively meaningful quality metric for consumer video.' Third International Workshop on Video Processing and Quality Metrics for Consumer Electronics, January 2006.

[205] Zhang, X., Lin, W., and Xue, P., 'Improved estimation for just-noticeable visual distortion.' *Signal Processing*, **85**(4), 2005, 795–808.

[206] Chou, C.H. and Li, Y.C., 'A perceptually tuned subband image coder based on the measure of just-noticeable-distortion profile.' *IEEE Transactions on Circuits and Systems for Video Technology*, **5**(6), 1995, 467–476.

[207] Tong, H.Y. and Venetsanopoulos, A.N., 'A perceptual model for jpeg applications based on block classification, texture masking, and luminance masking.' Proceedings of the IEEE International Conference on Image Processing (ICIP), Vol. 3, 1998.

[208] Hontsch, I. and Karam, L.J., 'Adaptive image coding with perceptual distortion control.' *IEEE Transactions on Image Processing*, **11**(3), 2002, 213–222.

[209] Daly, S., 'Engineering observations from spatiovelocity and spatiotemporal visual models.' In van den Branden Lambrecht, C.J. (ed.), *Vision Models and Applications to Image and Video Processing*. Kluwer Academic, Norwell, MA, 2001.

[210] Jia, Y., Lin, W., and Kassim, A.A., 'Estimating just-noticeable distortion for video.' *IEEE Transactions on Circuits and Systems for Video Technology*, **16**(7), 2006, 820–829.

[211] Ahumada, A.J. and Peterson, H.A., 'Luminance-model-based DCT quantization for color image compression.' SPIE Proceedings: Human Vision, Visual Processing, and Digital Display III, 1992, pp. 365–374.

[212] Jayant, N., Johnston, J., and Safranek, R., 'Signal compression based on models of human perception.' *Proceedings of IEEE*, **81**, 1993, 1385–1422.

[213] Wang, Z., Bovik, A.C., and Lu, L., 'Wavelet-based foveated image quality measurement for region of interest image coding.' Proceedings of International Conference on Image Processing, Vol. 2, 2001, pp. 89–92.

[214] Ahumada, A.J. and Krebs, W.K., 'Masking in color images.' SPIE Proceedings: Human Vision and Electronic Imaging VI, 2001, p. 4299.

[215] Chiu, Y.J. and Berger, T., 'A software-only videocodec using pixelwise conditional differential replenishment and perceptual enhancements.' *IEEE Transactions on Circuits and Systems for Video Technology*, **9**(3), 1999, 438–450.

[216] Lin, W., Gai, Y., and Kassim, A.A., 'A study on perceptual impact of edge sharpness in images.' *IEE Proceedings on Vision, Image, and Signal Processing*, **153**(2), 2006, 215–223.

[217] Chou, C.H. and Chen, C.W., 'A perceptually optimized 3-D subband image codec for video communication over wireless channels.' *IEEE Transactions on Circuits and Systems for Video Technology*, **6**(2), 1996, 143–156.

[218] Zhang, X., Lin, W., and Xue, P., 'Just-noticeable difference estimation with pixels in images.' *Journal of Visual Communication and Image Representation*, **19**(1), 2008, 30–41.

[219] Kollmeier, B., Brand, T., and Meyer, B., 'Perception of speech and sound.' In Benesty, J., Mohan Sondhi, M., and Huang, Y. (eds), *Springer Handbook of Speech Processing.* Springer-Verlag, Berlin, 2008, p. 65.

[220] Riesz, R., 'Differential intensity sensitivity of the ear for pure tones.' *Physical Review*, **31**(5), 1928, 867–875.

[221] Zwicker, E. and Fastl, H., *Psycho-Acoustics, Facts and Models*. Springer-Verlag, Berlin, 1999.

[222] Plomp, R., 'Rate of decay of auditory sensation.' *Journal of the Acoustical Society of America*, **36**(2), 1964, 277–282.

[223] Chun, M.M. and Wolfe, J.M., 'Visual attention.' In Goldstein, B. (ed.), *Blackwell Handbook of Perception.* Blackwell, Oxford, 2001, pp. 272–310.

[224] Posner, M.I., 'Orienting of attention.' *Quarterly Journal of Experimental Psychology*, **32**, 1980, 2–25.

[225] Pashler, H.E., *The Psychology of Attention.* MIT Press, Boston, MA, 1998.

[226] Treisman, A.M. and Gelade, G., 'A feature integration theory of attention.' *Cognitive Psychology*, **12**(1), 1980, 97–136.

[227] Desimone, R. and Duncan, J., 'Neural mechanisms of selective visual attention.' *Annual Review of Neuroscience*, **18**, 1995, 193–222.

[228] Hopfinger, J.B., Buonocore, M.H., and Mangun, G.R., 'The neural mechanisms of top-down attentional control.' *Nature Neuroscience*, **3**, 2000, 284–291.

[229] Navalpakkam, V. and Itti, L., 'Top-down attention selection is fine-grained.' *Journal of Vision*, **6**(11), 2006, 1180–1193.

[230] 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), 1998, 1254–1259.

[231] Hou, X. and Zhang, L., 'Saliency detection: A spectral residual approach.' IEEE Conference on Computer Visual Pattern Recognition, 2007.

[232] Borji, A. and Itti, L., 'State-of-the-art in visual attention modeling.' *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **35**(1), 2013, 185–207.

[233] Fang, Y., Chen, Z., Lin, W., and Lin, C.-W., 'Saliency detection in the compressed domain for adaptive image retargeting.' *IEEE Transactions on Image Processing*, **21**(9), 2012, 3888–3901.

[234] Fang, Y., Lin, W., Lee, B.-S., Lau, C.T., Chen, Z., and Lin, C.-W., 'Bottom-up saliency detection model based on human visual sensitivity and amplitude spectrum.' *IEEE Transactions on Multimedia*, **14**(1), 2012, 187–198.

[235] Ma, Y.-F., Hua, X.-S., Lu, L., and Zhang, H.-J., 'A generic framework of user attention model and its application in video summarization.' *IEEE Transactions on Multimedia*, **7**(5), 2005, 907–919.

[236] Fang, Y., Lin, W., Chen, Z., Tsai, C.-M., and Lin, C.-W., 'Video saliency detection in compressed domain.' *IEEE Transactions on Circuits and Systems for Video Technology*, **24**(1), 2014, 27–38.

[237] Wang, Z., Lu, L., and Bovik, A.C., 'Foveation scalable video coding with automatic fixation selection.' *IEEE Transactions on Image Processing*, **12**, 2003, 1703–1705.

[238] Kalinli, O. and Narayanan, S., 'A saliency-based auditory attention model with applications to unsupervised prominent syllable detection in speech.' Proceedings of Interspeech, 2007.

[239] Wrigley, S.N. and Brown, G.J., 'A computational model of auditory selective attention.' *IEEE Transactions on Neural Networks*, **15**(5), 2004, 1151–1163.

[240] Kayser, C., Petkov, C.I., Lippert, M., and Logothetis, N.K., 'Mechanisms for allocating auditory attention: An auditory saliency map.' *Current Biology*, **15**(21), 2005, 1943–1947.

[241] Damera-Venkata, N., Kite, T.D., Geisler, W.S., Evans, B.L., and Bovik, A.C., 'Image quality assessment based on a degradation model.' *IEEE Transactions on Image Processing*, **9**(4), 2000, 636–650.

[242] Wang, Z., Simoncelli, E.P., and Bovik, A.C., 'Multi-scale structural similarity for image quality assessment.' Proceedings of Asilomar Conference on Signals, Systems and Computers, Vol. 2, 2003.

[243] Horé, A. and Ziou, D., 'Is there a relationship between peak-signal-to-noise ratio and structural similarity index measure?' *IET Image Processing*, **7**(1), 2013, 12–24.

[244] Liu, A., Lin, W., and Narwaria, M., 'Image quality assessment based on gradient similarity.' *IEEE Transactions on Image Processing*, **21**(4), 2012, 1500–1512.

[245] Zhai, G., Wu, X., Yang, X., Lin, W., and Zhang, W., 'A psychovisual quality metric in free-energy principle.' *IEEE Transactions on Image Processing*, **21**(1), 2012, 41–52.

[246] Winkler, S. and Min, D., 'Stereo/multiview picture quality: Overview and recent advances.' *Signal Processing: Image Communication*, **28**(10), 2013, 1358–1373.

[247] Huynh-Thu, Q., Le Callet, P., and Barkowsky, M., 'Video quality assessment: From 2D to 3D – challenges and future trends.' IEEE International Conference on Image Processing, 2010.

[248] Benoit, A., Le Callet, P., Campisi, P., and Cousseau, R., 'Quality assessment of stereoscopic images.' *EURASIP Journal on Image and Video Processing*, 2008, 2009, 1–13.

[249] You, J., Xing, L., Perkis, A., and Wang, X., 'Perceptual quality assessment for stereoscopic images based on 2D image quality metrics and disparity analysis.' Proceedings of International Workshop on Video Processing and Quality Metrics, 2010.

[250] Yang, J., Hou, C., Zhou, Y., Zhang, Z., and Guo, J., 'Objective quality assessment method of stereo images.' 3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video, 2009, pp. 1–4.

[251] Shen, L., Yang, J., and Zhang, Z., 'Stereo picture quality estimation based on a multiple channel HVS model.' IEEE International Congress on Image and Signal Processing, 2009.

[252] Lambooij, M., IJsselsteijn, W., Bouwhuis, D., and Heynderickx, I., 'Evaluation of stereoscopic images: Beyond 2D quality.' *IEEE Transactions on Broadcasting*, **57**(2), 2011, 432–444.

[253] Shao, F., Lin, W., Gu, S., Jiang, G., and Srikanthan, T., 'Perceptual full-reference quality assessment of stereoscopic images by considering binocular visual characteristics.' *IEEE Transactions on Image Processing*, **22**(5), 2013, 1940–1953.

[254] Sazzad, Z.M.P., Yamanaka, S., Kawayoke, Y., and Horita, Y., 'Stereoscopic image quality prediction.' IEEE Quality of Media Experience, 2009.

[255] van den Branden Lambrecht, C.J. and Verscheure, O., 'Perceptual quality measure using a spatio-temporal model of the human visual system.' Proceedings of SPIE Digital Video Compression: Algorithms and Technologies, Vol. 2668, 1996, pp. 450–461.

[256] Wang, Z., Lu, L., and Bovik, A., 'Video quality assessment based on structural distortion measurement.' *Signal Processing: Image Communication*, **19**(2), 2004, 121–132.

[257] Lu, L., Wang, Z., Bovik, A., and Kouloheris, J., 'Full-reference video quality assessment considering structural distortion and no-reference quality evaluation of MPEG video.' Proceedings of IEEE International Conference and Multimedia Expo, 2002.

[258] Wang, Z. and Li, Q., 'Video quality assessment using a statistical model of human visual speed perception.' *Journal of the Optical Society of America A: Optics, Image Science, and Vision*, **24**(12), 2007, B61–B69.

[259] Stocker, A.A. and Simoncelli, E.P., 'Noise characteristics and prior expectations in human visual speed perception.' *Nature Neuroscience*, **9**, 2006, 578–585.

[260] Tao, P. and Eskicioglu, A.M., 'Video quality assessment using M-SVD.' Proceedings of the International Society of Optical Engineers, 2007.

[261] Bhat, A., Richardson, I., and Kannangara, S., 'A new perceptual quality metric for compressed video.' IEEE Conference on Acoustics, Speech, and Signal Processing, 2009.

[262] Lee, C. and Kwon, O., 'Objective measurements of video quality using the wavelet transform.' *Optical Engineering*, **42**(1), 2003, 265–272.

[263] Seshadrinathan, K. and Bovik, A.C., 'Motion tuned spatio-temporal quality assessment of natural videos.' *IEEE Transactions on Image Processing*, **19**(2), 2010, 335–350.

[264] Ong, E., Yang, X., Lin, W., Lu, Z., and Yao, S., 'Video quality metric for low bitrate compressed video.' Proceedings of International Conference on Image Processing, 2004.

[265] Ong, E., Lin, W., Lu, Z., and Yao, S., 'Colour perceptual video quality metric.' Proceedings of International Conference on Image Processing, 2006.

[266] Ndjiki-Nya, P., Barrado, M., and Wiegand, T., 'Efficient full-reference assessment of image and video quality.' Proceedings of International Conference on Image Processing, 2007.

[267] Pinson, M. and Wolf, S., 'Application of the NTIA general video quality metric VQM to HDTV quality monitoring.' 3rd International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM2007), 2007.

[268] Sugimoto, O., Naito, S., Sakazawa, S., and Koike, A., 'Objective perceptual picture quality measurement method for high-definition video based on full reference framework.' Proceedings of the International Society of Optical Engineers, Vol. 7242, 2009.

[269] Okamoto, J., Watanabe, K., Hondaii, A., Uchida, M., and Hangai, S., 'HDTV objective video quality assessment method applying fuzzy measure.' Proceedings of International Workshop on Quality Multimedia Experience (QoMEX), 2009.

[270] Narwaria, M., Lin, W., and Liu, A., 'Low-complexity video quality assessment using temporal quality variations.' *IEEE Transactions on Multimedia*, **14**(3–1), 2012, 525–535.

[271] Tan, K.T. and Ghanbari, M., 'A multimetric objective picture-quality measurement model for MPEG video.' *IEEE Transactions on Circuits and Systems for Video Technology*, **10**(7), 2000, 1208–1213.

[272] Song, L., Tang, X., Zhang, W., Yang, X., and Xia, P., 'The SHTU 4K video sequence dataset.' Proceedings of International Workshop on Quality of Multimedia Experience (QoMEX), 2013.

[273] Bae, S.-H., Kim, J., Kim, M., Cho, S., and Choi, J.S., 'Assessments of subjective video quality on HEVC-encoded 4K-UHD video for beyond-HDTV broadcasting services.' *IEEE Transactions on Broadcasting*, **59**(2), 2013, 209–222.

[274] Hanhart, P., Korshunov, P., and Ebrahimi, T., Benchmarking of quality metrics on ultra-high definition video sequences. International Conference on Digital Signal Processing, 2013.

[275] Babu, R.V., Bopardikar, A.S., Perkis, A., and Hillestad, O.I., 'No-reference metrics for video streaming applications.' International Workshop on Packet Video, 2004.

[276] Yasakethu, S.L.P., Hewage, C.T.E.R., Fernando, W.A.C., and Kondoz, A.M., 'Quality analysis for 3D video using 2D video quality models.' *IEEE Transactions on Consumer Electronics*, **54**(4), 2008, 1969–1976.

[277] Bosc, E., Pepion, R., Le Callet, P., *et al.*, 'Towards a new quality metric for 3-D synthesized view assessment.' *IEEE Journal of Selected Topics in Signal Processing*, **5**(7), 2011, 1332–1343.

[278] Boev, A., Gotchev, A., Egiazarian, K., Aksay, A., and Akar, G.B., 'Towards compound stereo-video quality metric: A specific encoder-based framework.' IEEE Southwest Symposium on Image Analysis and Interpretation, 2006.

[279] Han, J., Jiang, T., and Ma, S., 'Stereoscopic video quality assessment model based on spatial-temporal structural information.' VCIP, 2012.

[280] Zhu, Z. and Wan, Y., 'Perceptual distortion metric for stereo video quality evaluation.' *WSEAS Transactions on Signal Processing*, **5**(7), 2009, 241–250.

[281] Keimel, C., Oelbaum, T., and Diepold, K.J., 'No-reference video quality evaluation for high-definition video.' IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2009.

[282] Chen, W., Fournier, J., Barkowsky, M., and Le Challet, P., 'New requirements of subjective video quality assessment methodologies for 3DTV.' Proceedings of VPQM, 2010.

[283] Kim, D.-S., 'ANIQUE: An auditory model for single-ended speech quality estimation.' *IEEE Transactions on Speech and Audio Processing*, **13**(5), 2005, 821–831.

[284] Paillard, B., Mabilleau, P., Morisette, S., and Soumagne, J., 'PERCEVAL: Perceptual evaluation of the quality of audio signals.' *Journal of the Audio Engineering Society*, **40**(1/2), 1992, 21–31.

[285] Rix, A.W., Hollier, M.P., Hekstra, A.P., and Beerends, J.G., 'Perceptual evaluation of speech quality (PESQ), the new ITU standard for end-to-end speech quality assessment, Part I – Time-delay compensation.' *Journal of the Audio Engineering Society*, **50**(10), 2002, 755–764.

[286] Salmela, J. and Mattila, V.-V., 'New intrusive method for the objective quality evaluation of acoustic noise suppression in mobile communications.' Proceedings of the 116th Audio Engineering Society Convention, 2004.

[287] Steinmetz, R., 'Human perception of jitter and media synchronization.' *IEEE Journal on Selected Areas in Communications*, **14**(1), 1996, 61–72.

[288] Arrighi, R., Alais, D., and Burr, D., 'Perceptual synchrony of audiovisual streams for natural and artificial motion sequences.' *Journal of Vision*, **6**(3), 2006, 260–268.

[289] Lipscomb, S.D., 'Cross-modal integration: Synchronization of auditory and visual components in simple and complex media.' Proceedings of the Forum Acusticum, Berlin, 1999.

[290] Winkler, S. and Faller, C., 'Perceived audiovisual quality of low-bitrate multimedia content.' *IEEE Transactions on Multimedia*, **8**(5), 2006, 973–980.

[291] Beerends, J.G. and de Caluwe, F.E., 'The influence of video quality on perceived audio quality and vice versa.' *Journal of the Audio Engineering Society*, **47**(5), 1999, 355–362.

[292] Jones, C. and Atkinson, D.J., 'Development of opinion-based audiovisual quality models for desktop video-teleconferencing.' Proceedings of International Workshop on Quality of Service, Napa, CA, May 18–20, 1998, pp. 196–203.

[293] Ries, M., Puglia, R., Tebaldi, T., Nemethova, O., and Rupp, M., 'Audiovisual quality estimation for mobile streaming services.' Proceedings of International Symposium on Wireless Communication Systems, Siena, Italy, September 5–7, 2005.

[294] Jumisko-Pyykko, S., 'I would like to see the subtitles and the face or at least hear the voice: Effects of picture ratio and audiovideo bitrate ratio on perception of quality in mobile television.' *Multimedia Tools and Applications*, **36**(1&2), 2008, 167–184.

[295] Hayashi, T., Yamagishi, K., Tominaga, T., and Takahashi, A., 'Multimedia quality integration function for videophone services.' Proceedings of the IEEE International Conference on Global Telecommunications, 2007, pp. 2735–2739.

[296] Goudarzi, M., Sun, L., and Ifeachor, E., ‘Audiovisual quality estimation for video calls in wireless applications.’ IEEE GLOBECOM, 2010.

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

[298] Thang, T.C., Kang, J.W., and Ro, Y.M., ‘Graph-based perceptual quality model for audiovisual contents.’ Proceedings of the IEEE International Conference on Multimedia and Expo (ICME’07), Beijing, China, July 2007, pp. 312–315.

[299] Thang, T.C., Kim, Y.S., Kim, C.S., and Ro, Y.M., ‘Quality models for audiovisual streaming.’ Proceedings of SPIE: Electronic Imaging, Vol. 6059, 2006, pp. 1–10

[300] Thang, T.C. and Ro, Y.M., ‘Multimedia quality evaluation across different modalities.’ Proceedings of SPIE: Electronic Imaging, Vol. 5668, 2005, pp. 270–279.

Acronyms

CSF	Contrast Sensitivity Function
DCT	Discrete Cosine Transform
DWT	Discrete Wavelet Transform
FFT	Fast Fourier Transform
FR	Full Reference
GMM	Gaussian Mixture Model
HAS	Human Auditory System
HDR	High Dynamic Range
HDTV	High-Definition Television
HEVC	High-Efficiency Video Coding
HVS	Human Visual System
IGM	Internal Generative Mechanism
JND	Just-Noticeable Difference
MAE	Mean Absolute Error
MMSE	Minimum Mean Square Error
MOS	Mean Opinion Score
MOV	Model Output Variable
MOVIE	Motion-Based Video Integrity Evaluation
MSE	Mean Square Error
MT	Middle Temporal
NR	No Reference
ODG	Overall Difference Grade
PEAQ	Perceptual Evaluation of Audio Quality
PSF	Point Spread Function
PSNR	Peak SNR
PSQM	Perceptual Speech Quality Measure
RR	Reduced Reference
SDTV	Standard-Definition Television
SNR	Signal-to-Noise Ratio
SSIM	Structural Similarity

SVD	Singular Value Decomposition
UHD	Ultra-High Definition
VA	Visual Attention
VDP	Visible Differences Predictor
VIF	Visual Information Fidelity
VSNR	Visual Signal-to-Noise Ratio

4

Quality of Experience for HTTP Adaptive Streaming Services

Ozgur Oyman, Vishwanath Ramamurthi, Utsaw Kumar, Mohamed Rehan and Rana Morsi
Intel Corporation, USA

4.1 Introduction

With the introduction of smartphones like the iPhone$^{\text{TM}}$ and Android$^{\text{TM}}$-based platforms, the emergence of new tablets like the iPad$^{\text{TM}}$, and the continued growth of netbooks, ultrabooks, and laptops, there is an explosion of powerful mobile devices in the market which are capable of displaying high-quality video content. In addition, these devices are capable of supporting various video-streaming applications, interactive video applications like video conferencing, and can capture video for video-sharing, video-blogging, video-Twitter$^{\text{TM}}$, and video-broadcasting applications. Cisco predicts that mobile traffic will grow by a factor of 11 until 2018, and that this traffic will be dominated by video (so, by 2018, over 66% of the world's mobile traffic will be video).[1] As a result, future wireless networks will need to be optimized for the delivery of a range of video content and video-based applications.

Yet, video communication over mobile broadband networks today is challenging due to limitations in bandwidth and difficulties in maintaining the high reliability, quality, and latency demands imposed by rich multimedia applications. Even with the migration from 3G to 4G networks – or Radio Access Networks (RANs) and backhaul upgrades to 3G networks – the demand on capacity for multimedia traffic will continue to increase. As subscribers take

[1] See the following white papers from Cisco Visual Networking Index:
http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.html and http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-481360_ns827_Networking_Solutions_White_Paper.html.

Multimedia Quality of Experience (QoE): Current Status and Future Requirements, First Edition.
Edited by Chang Wen Chen, Periklis Chatzimisios, Tasos Dagiuklas and Luigi Atzori.

advantage of new multimedia content, applications, and devices, they will consume all available bandwidth and still expect the same quality of service that came with their original service plans – if not better. Such consumer demand requires exploration of new ways to optimize future wireless networks for video services toward delivering higher user capacity to serve more users and also deliver enhanced Quality of Experience (QoE) for a rich set of video applications.

One of the key video-enhancing solutions is adaptive streaming, which is an increasingly promising method to deliver video to end-users, allowing enhancements in QoE and network bandwidth efficiency. Adaptive streaming aims to optimize and adapt the video configurations over time in order to deliver the best possible quality video to the user at any given time, considering changing link or network conditions, device capabilities, and content characteristics. Adaptive streaming is especially effective in better tackling the bandwidth limitations of wireless networks, but also allows for more intelligent video streaming that is device-aware and content-aware.

Most of the expected broad adoption of adaptive streaming will be driven by new deployments over the existing web infrastructure based on the HyperText Transfer Protocol (HTTP) [1], and this kind of streaming is referred to here as HTTP Adaptive Streaming (HAS). HAS follows the pull-based streaming paradigm, rather than the traditional push-based streaming based on stateful protocols such as the Real-Time Streaming Protocol (RTSP) [2], where the server keeps track of the client state and drives the streaming. In contrast, in pull-based streaming such as HAS, the client plays the central role by carrying the intelligence that drives the video adaptation (i.e., since HTTP is a stateless protocol). Several important factors have influenced this paradigm shift from traditional push-based streaming to HTTP streaming, including: (i) broad market adoption of HTTP and TCP/IP protocols to support the majority of Internet services offered today; (ii) HTTP-based delivery avoids Network Address Translation (NAT) and firewall traversal issues; (iii) a broad deployment of HTTP-based (non-adaptive) progressive download solutions already exists today, which can conveniently be upgraded to support HAS; and (iv) the ability to use standard/existing HTTP servers and caches instead of specialized streaming servers, allowing for reuse of the existing infrastructure and thereby providing better scalability and cost-effectiveness. Accordingly, the broad deployment of HAS technologies will serve as a major enhancement to (non-adaptive) progressive download methods, allowing for enhanced QoE enabled by intelligent adaptation to different link conditions, device capabilities, and content characteristics.

As a relatively new technology in comparison with traditional push-based adaptive streaming techniques, deployment of HAS techniques presents new challenges and opportunities for content developers, service providers, network operators, and device manufacturers. One such important challenge is developing evaluation methodologies and performance metrics to accurately assess user QoE for HAS services, and effectively utilizing these metrics for service provisioning and optimizing network adaptation. In that vein, this chapter provides an overview of HAS concepts and recent Dynamic Adaptive Streaming over HTTP (DASH) standardization, and reviews the recently adopted QoE metrics and reporting framework in Third-Generation Partnership Project (3GPP) standards. Furthermore, we present an end-to-end QoE evaluation study on HAS conducted over 3GPP LTE networks and conclude with a discussion of future directions and challenges in QoE optimization for HAS services.

4.2 HAS Concepts and Standardization Overview

HAS has already been spreading as a form of Internet video delivery, with the recent deployment of proprietary solutions such as Apple HTTP Live Streaming, Microsoft Smooth Streaming, and Adobe HTTP Dynamic Streaming.[2] In the meantime, the standardization of HAS has also made great progress, with the recent completion of technical specifications by various standards bodies. In particular, DASH has recently been standardized by Moving Picture Experts Group (MPEG) and 3GPP as a converged format for video streaming [1, 2], and the standard has been adopted by other organizations including Digital Living Network Alliance (DLNA), Open IPTV Forum (OIPF), Digital Entertainment Content Ecosystem (DECE), World-Wide Web Consortium (W3C), and Hybrid Broadcast Broadband TV (HbbTV). DASH today is endorsed by an ecosystem of over 50 member companies at the DASH Industry Forum. Going forward, future deployments of HAS are expected to converge through broad adoption of these standardized solutions.

The scope of both MPEG and 3GPP DASH specifications [1,2] includes a normative definition of a media presentation or manifest format (for DASH access client), a normative definition of the segment formats (for media engine), a normative definition of the delivery protocol used for the delivery of segments, namely HTTP/1.1, and an informative description of how a DASH client may use the provided information to establish a streaming service. This section will provide a technical overview of the key parts of the DASH-based server–client interfaces, which are part of MPEG and 3GPP DASH standards. More comprehensive tutorials on various MPEG and 3GPP DASH features can be found in [3–5].

The DASH framework between a client and web/media server is depicted in Figure 4.1. The media preparation process generates segments that contain different encoded versions of one or several media components of the media content. The segments are then hosted on one or several media origin servers, along with the Media Presentation Description (MPD) that characterizes the structure and features of the media presentation, and provides sufficient information to a client for adaptive streaming of the content by downloading the media segments from the server over HTTP. The MPD describes the various representations of the media components (e.g., bit rates, resolutions, codecs, etc.) and HTTP URLs of the corresponding media segments, timing relationships across the segments, and how they are mapped into media presentations.

The MPD is an XML-based document containing information on the content, based on a hierarchical data model as depicted in Figure 4.2. Each period consists of one or more adaptation sets. An adaptation set contains interchangeable/alternate encodings of one or more media content components encapsulated in representations (e.g., an adaptation set for video, one for primary audio, one for secondary audio, one for captions, etc.). In other words, representations encapsulate media streams that are considered to be perceptually equivalent. Typically, dynamic switching happens across representations within one adaptation set. Segment

[2] Related white papers can be found at the following links:

- Microsoft Smooth Streaming: http://www.microsoft.com/download/en/details.aspx?id=17678.
- Adobe HTTP Dynamic Streaming: http://www.adobe.com/products/httpdynamicstreaming/.
- Apple HTTP Live Streaming: http://developer.apple.com/library/ios/documentation/networkinginternet/conceptual/streamingmediaguide/StreamingMediaGuide.pdf.

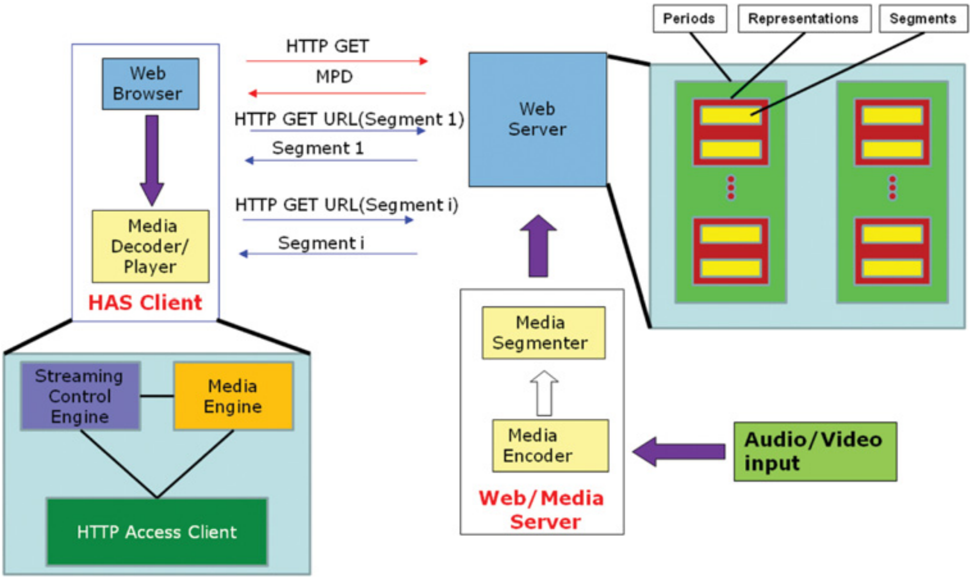


Figure 4.1 HAS framework between the client and web/media server

alignment permits non-overlapping decoding and presentation of segments from different representations. Stream Access Points (SAPs) indicate presentation times and positions in segments at which random access and switching can occur. DASH also uses a simplified version of XLink in order to allow loading parts of the MPD (e.g., periods) in real time from a remote

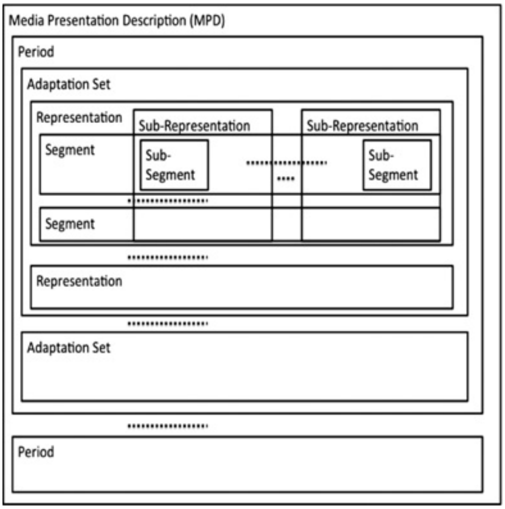


Figure 4.2 DASH MPD hierarchical data model

location. The MPD can be static or dynamic: a dynamic MPD (e.g., for live presentations) also provides segment availability start time and end time, approximate media start time, and the fixed or variable duration of segments. It can change and will be periodically reloaded by the client, while a static MPD is valid for the whole presentation. Static MPDs are a good fit for video-on-demand applications, whereas dynamic MPDs are used for live and Personal Video Recorder (PVR) applications.

A DASH segment constitutes the entity body of the response when issuing a HTTP GET or a partial HTTP GET request, and is the minimal individually addressable unit of data. DASH segment formats are defined for the ISO Base Media File Format (BMFF) and the MPEG2 Transport Stream format. A media segment contains media components and is assigned an MPD URL element and a start time in the media presentation. Segment URLs can be provided in the MPD in the form of exact URLs (segment list) or in the form of templates constructed via temporal or numerical indexing of segments. Dynamic construction of URLs is also possible, by combining parts of the URL (base URLs) that appear at different levels of the hierarchy. Each media segment also contains at least one SAP, which is a random access or switch-to point in the media stream where decoding can start using only data from that point forward. An initialization segment contains initialization information for accessing media segments contained in a representation and does not itself contain media data. Index segments, which may appear either as side files or within the media segments, contain timing and random access information, including media time vs. byte range relationships of sub-segments.

DASH provides the ability to the client to fully control the streaming session (i.e., it can intelligently manage the on-time request and smooth playout of the sequence of segments), potentially adjusting bit rates or other attributes in a seamless manner. The client can automatically choose the initial content rate to match the initial available bandwidth and dynamically switch between different bit-rate representations of the media content as the available bandwidth changes. Hence, DASH allows fast adaptation to changing network and link conditions, user preferences, and device states (e.g., display resolution, CPU, memory resources, etc.). Such dynamic adaptation provides better user QoE, with higher video quality, shorter startup delays, fewer rebuffering events, etc.

At MPEG, DASH was standardized by the Systems Sub-Group, with the activity beginning in 2010, becoming a Draft International Standard in January 2011, and an International Standard in November 2011. The MPEG DASH standard [1] was published as ISO/IEC 23009-1:2012 in April 2012. In addition to the definition of media presentation and segment formats standardized in [1], MPEG has also developed additional specifications [6–8] on aspects of implementation guidelines, conformance and reference software, and segment encryption and authentication. Toward enabling interoperability and conformance, DASH also includes profiles as a set of restrictions on the offered MPD and segments based on the ISO BMFF [9] and MPEG2 Transport Streams [10], as depicted in Figure 4.3. In the meantime, MPEG DASH is codec agnostic and supports both multiplexed and non-multiplexed encoded content. Currently, MPEG is also pursuing several core experiments toward identifying further DASH enhancements, such as signaling of quality information, DASH authentication, server and network-assisted DASH operation, controlling DASH client behavior, and spatial relationship descriptions.

At 3GPP, DASH was standardized by the 3GPP SA4 Working Group, with the activity beginning in April 2009 and Release 9 work with updates to Technical Specification (TS)

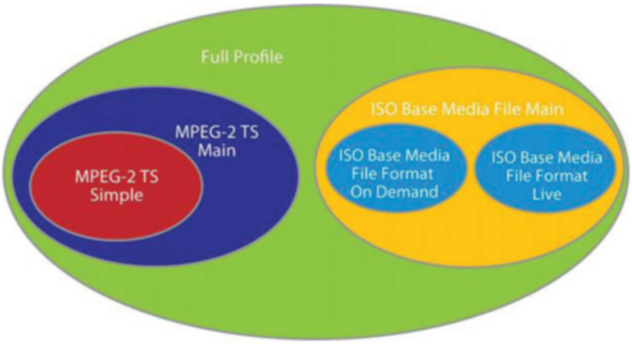
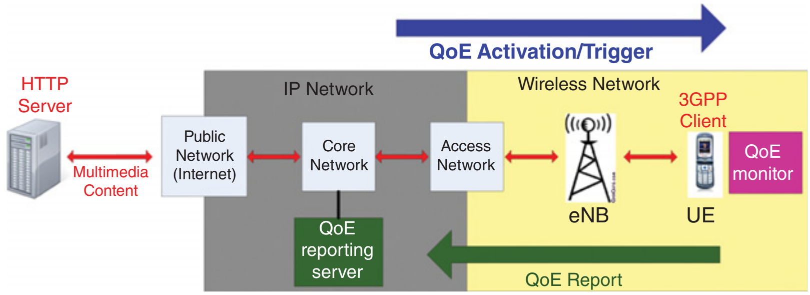

Full Profile
MPEG-2 TS Main
MPEG-2 TS Simple
ISO Base Media File Main
ISO Base Media File Format On Demand
ISO Base Media File Format Live

Figure 4.3 MPEG DASH profiles

26.234 on the Packet Switched Streaming Service (PSS) [11] and TS 26.244 on the 3GPP file format [12] completed in March 2010. During Release 10 development, a new specification TS 26.247 on 3GPP DASH [2] was finalized in June 2011, in which ISO BMFF-based DASH profiles were adopted. In conjunction with a core DASH specification, 3GPP DASH also includes additional system-level aspects, such as codec and Digital Rights Management (DRM) profiles, device capability exchange signaling, and QoE reporting. Since Release 11, 3GPP has been studying further enhancements to DASH and toward this purpose collecting new use cases and requirements, as well as operational and deployment guidelines. Some of the documented use cases in the related Technical Report (TR) 26.938 [13] include: operator control for DASH (e.g., for QoE/QoS handling), advanced support for live services, DASH as a download format for push-based delivery services, enhanced ad insertion support, enhancements for fast startup and advanced trick play modes, improved operation with proxy caches, Multimedia Broadcast and Multicast Service (MBMS)-assisted DASH services with content caching at the User Equipment (UE) [8], handling special content over DASH and enforcing specific client behaviors, and use cases on DASH authentication.

4.3 QoE in 3GPP DASH

The development of QoE evaluation methodologies, performance metrics, and reporting protocols plays a key role in optimizing the delivery of HAS services. In particular, QoE monitoring and feedback are beneficial for detecting and debugging failures, managing streaming performance, enabling intelligent client adaptation (useful for device manufacturer), and allowing for QoE-aware network adaptation and service provisioning (useful for the network operator and content/service provider). Having recognized these benefits, both 3GPP and MPEG bodies have adopted QoE metrics for HAS services as part of their DASH specifications. Moreover, the 3GPP DASH specification also provides mechanisms for triggering QoE measurements at the client device as well as protocols and formats for delivery of QoE reports to the network servers. Here, we shall describe in detail the QoE metrics and reporting framework for 3GPP DASH, while it should be understood that MPEG has also standardized similar QoE metrics in MPEG DASH.



3GPP QoE Metric	Progressive Download	DASH/HAS
HTTP Request/Response Transactions	Yes	Yes
Representation Switch Events	No	Yes
Average Throughput	Yes	Yes
Initial Playout Delay	Yes	Yes
Buffer Level	Yes	Yes
Play List	Yes	Yes
MPD Information	No	Yes

Figure 4.4 QoE metrics and reporting framework for 3GPP DASH and progressive download

In the 3GPP DASH specification TS 26.247, QoE measurement and reporting capability is defined as an optional feature for client devices. However, if a client supports the QoE reporting feature, the DASH standard also mandates the reporting of all the requested metrics at any given time (i.e., the client should be capable of measuring and reporting all of the QoE metrics specified in the standard). It should also be noted here that 3GPP TS 26.247 also specifies QoE measurement and reporting for HTTP-based progressive download services, where the set of QoE metrics in this case is a subset of those provided for DASH.

Figure 4.4 depicts the QoE monitoring and reporting framework specified in 3GPP TS 26.247, summarizes the list of QoE metrics standardized by 3GPP in the specification TS 26.247, and indicates the list of metrics applicable for DASH/HAS (adaptive streaming) and HTTP-based progressive download (non-adaptive). At a high level, the QoE monitoring and reporting framework is composed of the following phases: (1) server activates/triggers QoE reporting, requests a set of QoE metrics to be reported, and configures the QoE reporting framework; (2) client monitors or measures the requested QoE metrics according to the QoE configuration; (3) client reports the measured parameters to a network server. We now discuss each of these phases in the following sub-sections.

4.3.1 Activation and Configuration of QoE Reporting

3GPP TS 26.247 specifies two options for the activation or triggering of QoE reporting. The first option is via the QualityMetrics element in the MPD and the second option is via the OMA

Device Management (DM) QoE management object. In both cases, the trigger message from the server would include reporting configuration information such as the set of QoE metrics to be reported, the URIs for the server(s) to which the QoE reports should be sent, the format of the QoE reports (e.g., uncompressed or gzip), information on QoE reporting frequency and measurement interval, percentage of sessions for which QoE metrics will be reported, and Access Point Name (APN) to be used for establishing the Packet Data Protocol (PDP) context for sending the QoE reports.

4.3.2 QoE Metrics for DASH

The following QoE metrics have been defined in 3GPP DASH specification TS 26.247, to be measured and reported by the client upon activation by the server. It should be noted that these metrics are specific to HAS and content streaming over the HTTP/TCP/IP stack, and therefore differ considerably from QoE metrics for traditional push-based streaming protocols.

- **HTTP request/response transactions.** This metric essentially logs the outcome of each HTTP request and corresponding HTTP response. For every HTTP request/response transaction, the client measures and reports (i) the type of request (e.g., MPD, initialization segment, media segment, etc.), (ii) times for when the HTTP request was made and the corresponding HTTP response was received (in wall clock time), (iii) the HTTP response code, (iv) contents in the byte-range-spec part of the HTTP range header, (v) the TCP connection identifier, and (vi) throughput trace values for successful requests. From the HTTP request/response transactions, it is also possible to derive more specific performance metrics such as the fetch durations of the MPD, initialization segment, and media segments.
- **Representation switch events.** This metric is used to report a list of representation switch events that took place during the measurement interval. A representation switch event signals the client's decision to perform a representation switch from the currently presented representation to a new representation that is later presented. As part of each representation switch event, the client reports the identifier for the new representation, the time of the switch event (in wall clock time) when the client sent the first HTTP request for the new representation, and the media time of the earliest media sample played out from the new representation.
- **Average throughput.** This metric indicates the average throughput that is observed by the client during the measurement interval. As part of the average throughput metric, the client measures and reports (i) the total number of content bytes (i.e., the total number of bytes in the body of the HTTP responses) received during the measurement interval, (ii) the activity time during the measurement interval, defined as the time during which at least one GET request is still not completed, (iii) the wall clock time and duration of the measurement interval, (iv) the access bearer for the TCP connection for which the average throughput is reported, and (v) the type of inactivity (e.g., pause in presentation, etc.).
- **Initial playout delay.** This metric signals the initial playout delay at the start of the streaming of the presentation. It is measured as the time from when the client requests the fetch of the first media segment (or sub-segment) to the time at which media is retrieved from the client buffer.

- **Buffer level.** This metric provides a list of buffer occupancy-level measurements carried out during playout. As part of the buffer-level metric, the client measures and reports the buffer level that indicates the playout duration for which media data is available, starting from the current playout time along with the time of the measurement of the buffer level.
- **Play list.** This metric is used to log a list of playback periods in the measurement interval, where each playback period is the time interval between a user action and whichever occurs soonest of the next user action, the end of playback, or a failure that stops playback. The type of user actions that trigger playout may include a new playout request, resume playout from pause, or user-requested quality change. For each playback period, the client measures and reports the identifiers of the representations that were rendered and their rendering times (in media time) and durations, playback speed relative to normal playback speed (e.g., to track trick modes such as fast forward or rewind), and reasons for why continuous playback of this representation was interrupted (e.g., due to representation switch events, rebuffering, user request, or end of period, media content or a metrics collection period).
- **MPD information.** This metric allows for reporting information on the media presentations from the MPD so that servers without direct access to the MPD can learn the media characteristics. Media representation attributes on bit rate, resolution, quality ranking, and codec-related media information – including profile and level – can be reported by the client via this metric.

4.3.3 QoE Reporting Protocol

In 3GPP DASH, QoE reports are formatted as an eXtensible Markup Language (XML)[3] document complying with the XML schema provided in specification TS 26.247. The client uses HTTP POST request signaling (RFC 2616) carrying XML-formatted metadata in its body to send the QoE report to the server.

4.4 Link-Aware Adaptive Streaming

The central intelligence in HAS resides in the client rather than the server. The requested representation levels of video chunks (forming the HAS segments) are determined by the client and communicated to the server. Based on the frame levels, the operation of the client in a link-aware adaptive streaming framework can be characterized into four modes or states: (i) startup mode, (ii) transient state, (iii) steady state, and (iv) rebuffering state (see Figure 4.5).

Startup mode is the initial buffering mode, during which the client buffers video frames to a certain limit before beginning to play back the video (i.e., the client is in the startup mode as long as $A_i \leq A_{\text{thresh}}^{\text{StartUp}}$, where A_i represents the total number of video frames received until frame slot i. Steady state represents the state in which the UE buffer level is above a certain threshold (i.e., $B_i \leq B_{\text{thresh}}^{\text{Steady}}$), where B_i tracks the number of frames in the client buffer that are available for playback in frame slot i. The transient state is the state in which the UE buffer level falls below a certain limit after beginning to play back (i.e., $B_i < B_{\text{thresh}}^{\text{Steady}}$). The rebuffering

[3] For further information on XML, see http://www.w3.org/XML/.

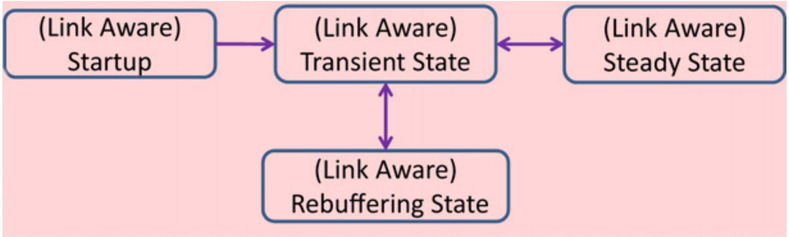

(Link Aware) Startup
(Link Aware) Transient State
(Link Aware) Steady State
(Link Aware) Rebuffering State

Figure 4.5 Adaptive streaming client player states

state is the state that the client enters when the buffer level becomes zero after beginning to play back. Once it enters the rebuffering state, it remains in that state until it rebuilds its buffer level to a satisfactory level to begin playback (i.e., until $B_i \leq B_{\text{thresh}}^{\text{Rebuff}}$).

One of the key aspects of adaptive streaming is the estimate of available link bandwidth. A typical throughput estimate is the average segment or HAS throughput, which is defined as the average ratio of segment size to download time of HAS segments:

$$R_j^j = \frac{1}{F} \sum_{s=S_j^i-F+1}^{S_j^i} \frac{S_j\,(s)}{T_j^{\text{dwld}}\,(s) - T_j^{\text{fetch}}\,(s)} \tag{4.1}$$

where $S_j(s)$, $T_j^{\text{fetch}}(s)$, and $T_j^{\text{dwld}}(s)$ are the size, fetch time, and download time of the sth video segment of client j, S_j^i the number of segments downloaded by client j until frame slot i, and F the number of video segments over which the average is computed. Based on this estimate, the best video representation level possible for the next video segment request is determined as follows:

$$\begin{aligned} Q_{i,j}^{\text{sup}} &= \arg\max_k b_k; \\ \text{s.t.}\quad b_k &\leq R_j^i;\quad k = 1, 2, \ldots, N \end{aligned} \tag{4.2}$$

The key QoE metrics of interest are: (i) startup delay, (ii) startup video quality, (iii) overall average video quality, and (iv) rebuffering percentage. Startup delay refers to the amount of time it takes to download the initial frames necessary to begin playback. Average video quality is the average video quality experienced by a user. Startup video quality refers to the average video quality in the startup phase. Rebuffering percentage is the percentage of time the client spends in the rebuffering state. It has been observed that rebuffering is the most annoying to video-streaming users, and hence it is important to keep the rebuffering percentage low by judicious rate adaptation.

Typical HAS algorithms use either application or transport-layer throughputs (as in Eq. (4.1)) for video rate adaptation [14]. We refer to these approaches as PHY Link Unaware (PLU). However, using higher layer throughputs alone can potentially have adverse effects on user QoE when the estimated value is different from what is provided by the wireless link conditions – a lower estimate results in lower quality and a higher estimate can result in rebuffering. These situations typically occur in wireless links due to changes in environmental and/or loading conditions. In [15], a Physical Link-Aware (PLA) approach to adaptive streaming was proposed to improve video QoE in changing wireless conditions. Physical-layer (PHY) goodput, used as a complement to higher layer-throughput estimates, allows us to track radio-link variations at a finer time scale. This opens up the possibility for opportunistic link-aware video-rate adaptation that is to improve the QoE of the user. Average PHY-layer goodput at time t is defined as the ratio of the number of bits received during the time period $(t - T, t)$ to the averaging duration T as follows:

$$R_t^{\text{phy}} = \frac{X(t) - X(t-T)}{T} \tag{4.3}$$

Using PHY goodput for HAS requires collaboration between the application and the physical layers, but it can provide ways to improve various QoE metrics for streaming over wireless using even simple enhancements. Here we describe two simple enhancements for the startup and steady states.

Typical HAS startup algorithms request one video segment every frame slot at the lowest representation level to build the playback buffer quickly. This compromises the playback video quality during the startup phase. Link-aware startup can be used to optimize video quality based on wireless link conditions right from the beginning of the streaming session. An incremental quality approach could be used so that startup delay does not increase beyond satisfactory limits due to quality optimization. The next available video adaptation rate is chosen if enough bandwidth is available to support such a rate. For this purpose, the ratio δ_i is defined as follows:

$$\delta_i = R_i^{\text{phy}} \Big/ b_{Q_{i-1}+1} \tag{4.4}$$

This ratio represents the ratio of the average PHY goodput to the next video representation level that is possible. Q_0 is initialized based on historical PHY goodput information before the start of the streaming session:

$$\begin{aligned} &Q_0 = \arg\max_k b_k; \\ &\text{s.t.} \quad b_k \le R_t^{\text{phy}}/(1+\alpha); \quad k = 1, 2, ..., N \end{aligned} \tag{4.5}$$

The representation level for the segment request in frame slot i is then selected as follows:

$$\begin{aligned} Q_i &= \min\left(Q_{i-1} + 1, Q_i^{\text{sup}}\right) \quad && \text{if } \left(\delta_i \ge (1+\alpha)\right) \\ Q_i &= \min\left(Q_{i-1}, Q_i^{\text{sup}}\right) && \text{otherwise} \end{aligned} \tag{4.6}$$

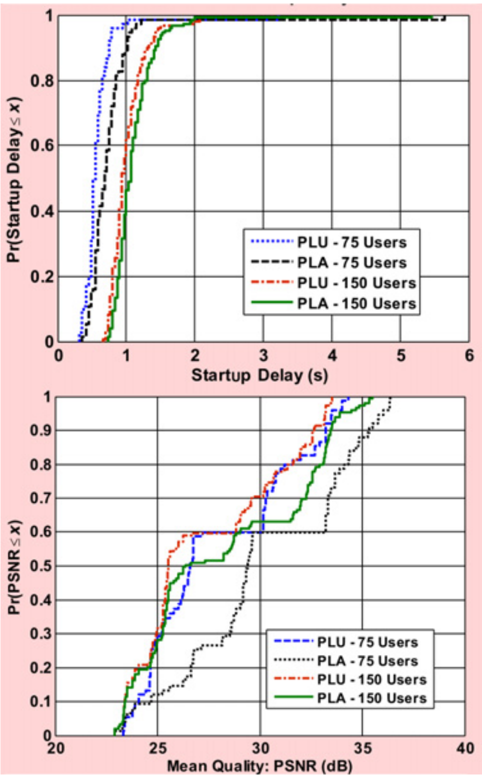


Figure 4.6 Startup delay and startup quality comparison for PLA and PLU approaches

The next representation level is chosen only when δ_i is greater than $(1 + \alpha)$. $\alpha > 0$ is a parameter that can be chosen depending on how aggressively or conservatively we would like to optimize quality during the startup phase. The condition $\delta_i \geq (1 + \alpha)$ ensures that the rate adaptation does not fluctuate with small-scale fluctuations of wireless link conditions.

For the following evaluation results, we use Peak Signal-to-Noise Ratio (PSNR) for video quality, although our approach is not restricted to this and other metrics such as Structural Similarity (SSIM) could also be used.

Figure 4.6 shows a comparison of the Cumulative Distribution Functions (CDF) of startup delay and average video quality during the startup phase for PLA and PLU approaches. For the 75-user scenario, PHY link awareness can improve the average startup quality by 2 to 3 dB for more than 90% of users, at the cost of only a slight (tolerable) increase in startup delay. In the 150-user scenario, we see a slightly lower 1 to 2 dB improvement in average video quality for more than 50% of users, with less than 0.25 s degradation in startup delay. These results demonstrate that PLA can enhance the QoE during the startup phase by improving video quality with an almost unnoticeable increase in startup delay.

In the steady-state mode, the buffer level at the client is above a certain level. In traditional steady-state algorithms, the objective is to maintain the buffer level without compromising video quality. This is typically done by periodically requesting one segment worth of frames for each segment duration in the steady state. However, this might result in rebuffering in wireless links that fluctuate. PHY goodput responds quickly to wireless link variations, while segment throughput responds more slowly to link variations. So, PHY goodput could be used as a complement to fragment throughput to aid in rate adaptation. When link conditions are good, $R_t^{\mathrm{phy}} > R_i^{\mathrm{seg}}$ and when link conditions are bad, $R_t^{\mathrm{phy}} < R_i^{\mathrm{seg}}$. A conservative estimate of maximum throughput is determined based on PHY goodput and segment throughput, which could help avoid rebuffering in the steady state. Such a conservative estimate may be achieved as follows:

$$R_i^{\mathrm{con}} = \begin{cases} R_i^{\mathrm{phy}} & \text{if } R_i^{\mathrm{seg}} > (1+\beta) R_i^{\mathrm{phy}} \\ R_i^{\mathrm{seg}} & \text{otherwise} \end{cases} \tag{4.7}$$

This approach ensures that (i) when the link conditions are bad and segment throughput is unable to follow the variation in link conditions, we use PHY goodput to lower the estimate of the link bandwidth that is used for video rate adaptation and (ii) when the link conditions are good in steady state, we get as good video quality as using the PLU approach. The constant β in Eq. (4.7) prevents short-term variations in link conditions from changing the rate adaptation. The best video representation level possible in frame slot i, $Q_i^{\mathbf{sup}}$, is determined conservatively based on R_i^{con}:

$$\begin{aligned} Q_i^{\mathrm{sup}} &= \arg\max_k b_k; \\ \text{s.t.}\quad b_k &\le R_i^{\mathrm{con}};\ \ k = 1, 2, ..., N \end{aligned} \tag{4.8}$$

Figure 4.7 compares the CDFs of rebuffering percentage and average playback video quality performance using PLA and PLU approaches for 100 and 150 users. In the 100-user scenario, the number of users not experiencing rebuffering improves from around 75% to 92% (a 17% improvement) and the peak rebuffering percentage experienced by any user reduces from around 30% to 13% using the PLA approach. This improvement in rebuffering performance is at the cost of only a slight degradation in video quality (0.6 dB average) compared with the PLU approach for some users. In the highly loaded 150-user scenario, we observe that using the PLA approach we can obtain around a 20% improvement in number of users not experiencing rebuffering (from around 56% to 76%) at the cost of minimal degradation in average video quality by less than 0.5 dB on average for 50% of users. Thus, PLA can enhance the user QoE during video playback by reducing the rebuffering percentage significantly at the cost of a very minor reduction in video quality.

4.5 Video-Aware Radio Resource Allocation

Wireless links are fluctuating by nature. In most cellular wireless networks, the UEs send to the Base Station (BS) periodic feedback regarding the quality of wireless link that they are experiencing in the form of Channel Quality Information (CQI). The CQI sent by the UEs is

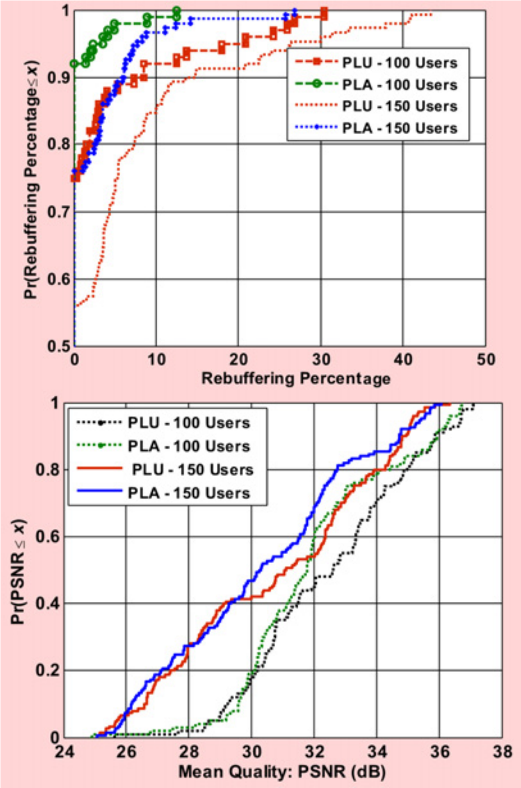


Figure 4.7 Rebuffering and average quality comparison for PLA and PLU approaches

discretized, thus making the overall channel state m discrete. The BS translates the CQI into a peak rate vector $\mu^{\mathbf{m}} = (\mu_1^m, \mu_2^m, ..., \mu_J^m)$, with μ_j^m representing the peak achievable rate by user j in channel state m. For every scheduling resource, the BS has to make a decision as to which user to schedule in that resource. Scheduling the best user would always result in maximum cell throughput but may result in poor fairness. Scheduling resources in a round-robin fashion might result in an inability to take advantage of the wireless link quality information that is available. So, typical resource allocation algorithms in wireless networks seek to optimize the average service rates $\mathbf{R} = \left(R_1, R_2, R_3, \ldots, R_J\right)$ to users such that a concave utility function $H(\mathbf{R})$ is maximized subject to the capacity (resource) limits in the wireless scenario under consideration, i.e.

$$\begin{aligned} \text{Basic RA} : \max\ & H(\mathbf{R}) \\ \text{s.t.}\ \ & \mathbf{R} \in \mathbf{V} \end{aligned} \tag{4.9}$$

where **V** represents the capacity region of the system. Utility functions of the sum form have attracted the most interest:

$$H(\mathbf{R}) = \sum_j H_j\left(R_j\right) \tag{4.10}$$

where each $H_j(R_j)$ is a strictly concave, continuously differentiable function defined for $R_j > 0$. The Proportional Fair (PF) and Maximum Throughput (MT) scheduling algorithms are special cases of objective functions of the form, with $H_j\left(R_j\right) = \log\left(R_j\right)$ and $H\left(R_j\right) = R_j$, respectively.

The key objective of a video-aware optimization framework for multi-user resource allocation is to reduce the possibility of rebuffering without interfering with the rate-adaptation decisions taken by the HAS client. To this end, a buffer-level feedback-based scheduling algorithm in the context of HAS was proposed in [10] by modifying the utility function of the PF algorithm to give priority to users with buffer levels lower than a threshold. However, this emergency-type response penalizes other users into rebuffering, especially at high loading conditions, thus decreasing the effectiveness of the algorithm. To overcome this limitation, a video-aware optimization framework that constrains rebuffering was proposed in [16]. In order to avoid rebuffering at a video client, video segments need to be downloaded at a rate that is faster than the playback rate of the video segments. Let $T_j(s)$ be the duration of time taken by user j to download a video segment s and $\tau_j(s)$ be the media duration of the segment. Then, to avoid rebuffering, the following constraint is introduced:

$$T_j(s) \leq \tau_j(s) / (1+\delta) \quad \forall j,\ s \tag{4.11}$$

where $\delta > 0$ is a small design parameter to account for variability in wireless network conditions. Segment download time $T_j(s)$ depends on the size of the video segment $S_j(s)$ and the data rates experienced by user j. $S_j(s)$ in turn depends on the video content and representation (adaptation) level that is chosen by the HAS client. The HAS client chooses the representation level for each video segment based on its state and its estimate of the available link bandwidth. Based on all this, we propose a Rebuffering Constrained Resource Allocation (RCRA) framework as follows:

$$\text{RCRA}: \quad \begin{array}{l} \max\ H(\mathbf{R}) \\ \text{s.t. } \mathbf{R} \in \mathbf{V} \\ T_j(s) \leq \tau_j(s) / (1+\delta) \quad \forall j,\ s \end{array} \tag{4.12}$$

The additional constraints related to rebuffering closely relate the buffer evolution at HAS clients to resource allocation at the base station. Intelligent resource allocation at the BS can help reduce rebuffering in video clients.

Enforcing the rebuffering constraints in Eq. (4.12) in a practical manner requires feedback from HAS clients. Each adaptive streaming user can feed back its media playback buffer level periodically to the BS scheduler in addition to the normal CQI feedback. The buffer-level feedback can be done directly over the RAN or more practically, indirectly through the video server.

Scheduling algorithms for multi-user wireless networks need to make decisions during every scheduling time slot (resource) t in such a way as to lead to a long-term optimal solution. The scheduling time slot for modern wireless networks is typically at much finer granularity than a (video) frame slot. A variant of the gradient scheduling algorithm called the Rebuffering-Aware Gradient Algorithm (RAGA) in [16] can be used to solve the optimization problem in Eq. (4.12) by using a token-based mechanism to enforce the rebuffering constraints. The RAGA scheduling decision in scheduling time slot t when the channel state is $m(t)$ can be summarized as follows:

$$\text{RAGA}: \ j = \arg\max_{j \in N} \left[e^{a_j(t) W_j(t)} \nabla H \left(R_j(t) \right) \cdot \mu_j^{m(t)} \right] \tag{4.13}$$

where $R_j(t)$ is the current moving-average service rate estimate for user j. It is updated every scheduling time slot as in the PF scheduling algorithm, i.e.

$$R_j(t+1) = (1-\beta) R_j(t) + \beta \mu_j(t) \tag{4.14}$$

where $\beta > 0$ is a small parameter that determines the time scale of averaging and $\mu_j(t)$ is the service rate of user j in time slot t. $\mu_j(t) = \mu_j^{m(t)}$ if user j was scheduled in time slot t and $\mu_j(t) = 0$ otherwise. $W_j(t)$ in Eq. (4.13) is a video-aware user token parameter and $a_j(t)$ is a video-aware user time-scale parameter, both of which are updated based on periodic media buffer-level feedback. These parameters hold the key to enforcing rebuffering constraints at the BS. Such a feedback mechanism has been defined in the DASH standard [1, 2] and is independent of specific client player implementation. For simplicity, we assume that such client media buffer-level feedback is available only at the granularity of a frame slot. Therefore, the user-token parameter and user-time-scale parameter are constant within a frame slot, i.e.

$$\begin{aligned} W_j(t) &= W_j^i \ \text{ for } \ i\tau \le t < (i+1)\tau \\ a_j(t) &= a_j^i \ \text{ for } \ i\tau \le t < (i+1)\tau \end{aligned} \tag{4.15}$$

Let B_j^i represent the buffer status feedback in frame slot i in units of media time duration. The difference between buffer levels from frame slot $(i-1)$ to frame slot i is given by

$$B_j^{i,\text{diff}} = B_j^i - B_j^{i-1}$$

A positive value for $B_j^{i,\text{diff}}$ indicates an effective increase in media buffer size in the previous reporting duration and a negative value indicates a decrease in media buffer size. Note that this difference depends on frame playback and download processes at the HAS client. To avoid rebuffering, we would like the rate of change at the client media buffer level to be greater than a certain positive threshold, i.e.

$$\left(B_j^{i,\text{diff}}/\tau\right) > \delta\,,\ \delta > 0 \tag{4.16}$$

The media-buffer-aware user-token parameter is updated every frame slot as follows:

$$W_j^i = max\left(W_j^{i-1} + \left(\delta\tau - B_j^{i,\text{diff}}\right), 0\right) \tag{4.17}$$

The intuitive interpretation of Eq. (4.17) is that if the rate of media buffer change for a certain user is below the threshold, the token parameter is incremented by an amount $(\delta\tau - B_j^{\text{diff}})$ that reflects the relative penalty for having a buffer change rate below the threshold. This increases its relative scheduling priority compared with other users whose media buffer change rate is higher. Similarly, when the rate of buffer change is above the threshold, the user-token parameter is decreased to offset any previous increase in scheduling priority. W_j^i is not reduced below zero, reflecting the fact that all users with a consistent buffer rate change greater than the threshold have scheduling priorities as per the standard proportional fair scheduler.

The video-aware parameter a_j^i determines the time scale over which rebuffering constraints are enforced for adaptive streaming users. A larger value of a_j^i implies greater urgency in enforcing the rebuffering constraints for user j. In a HAS scenario, the values of a_j^i can be set to reflect this relative urgency for different users. Therefore, we set a_j^i based on the media buffer level of user j in frame slot i as follows:

$$a_j^i = 1 + \phi * max\left(\frac{B_{\text{thresh}}^{\text{Steady}} - B_j^i}{B_{\text{thresh}}^{\text{Steady}}}, 0\right) \tag{4.18}$$

where ϕ is a scaling constant, B_j^i is the current buffer level in seconds for user j, and $B_{\text{thresh}}^{\text{Steady}}$ is the threshold for the steady-state operation of the HAS video client. If the buffer level B_j^i for user j is above the threshold, then $a_j^i = 1$ and if it is below the threshold, then a_j^i scales to give relatively higher priorities to users with lower buffer levels. This scaling of priorities based on absolute user buffer levels improves the convergence of the algorithm. The user-time-scale parameter a_j^i is set to 0 for non-adaptive streaming users, turning the metric in Eq. (4.13) into a standard PF metric. Note that the parameter $W_j(t)$ is updated based on the rate of media-buffer-level change, while the parameter $a_j(t)$ is updated based on the buffer levels themselves. Such an approach provides a continuous adaptation of user scheduling priorities based on media-buffer-level feedback (unlike an emergency-response-type response) and reduces the rebuffering percentage of users without significantly impacting video quality.

Figure 4.8 compares the rebuffering percentage and the Perceived Video Quality (PVQ) of resource allocation algorithm RAGA with standard Proportional Fair (PF), Proportional Fair with Barrier for Frames (PFBF), and GMR (Gradient with Minimum Rate) algorithms in a 100-user scenario. For GMR, we set the minimum rate for each video user to the rate of the

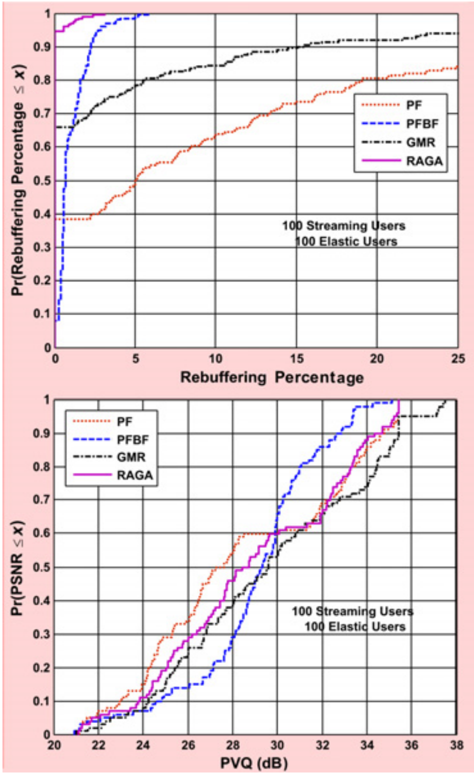


Figure 4.8 Rebuffering and average quality comparison for RAGA with different scheduling approaches

lowest representation level of the user's video. PVQ is computed as the difference between the mean and standard deviation of PSNR. Only played-out video frames are considered in the computation of PVQ. Observe that RAGA has the lowest rebuffering percentage among all the schemes across all the users. It has reduced the number of users experiencing rebuffering and also the amount of rebuffering experienced by the users. The PVQ using RAGA is better than PF scheduling for all users. GMR is better than PF in terms of rebuffering, but it still lags behind RAGA in rebuffering performance due to a lack of dynamic cooperation with the video clients. Although GMR appears to have marginally better PVQ than RAGA, this is at a huge cost in terms of increased rebuffering percentages. PFBF performs better than GMR in terms of peak rebuffering percentage but lags behind both PF and GMR in terms of the number of users experiencing rebuffering. Also, PFBF has better PVQ than all schemes for some users and worse than all schemes for others. The disadvantage with PFBF is that it reacts to low buffer levels in an emergency fashion and inadvertently penalizes good users to satisfy users with low buffer levels. RAGA continually adjusts the scheduling priorities of the users based on the rate of change of media buffer levels, thus improving the QoE of streaming users in terms of reduced rebuffering and balanced PVQ.

4.6 DASH over e-MBMS

As the multicast standard for Long-Term Evolution (LTE), enhanced Multimedia Broadcast Multicast Service (e-MBMS) was introduced by 3GPP to facilitate delivery of popular content to multiple users over a cellular network in a scalable fashion. Delivery of popular YouTube clips, live sports events, news updates, advertisements, file sharing, etc. are relevant use cases for eMBMS. eMBMS utilizes the network bandwidth more efficiently than unicast delivery by using the inherent broadcast nature of wireless channels. For unicast transmissions, retransmissions based on Automatic Repeat Request (ARQ) and/or Hybrid ARQ (HARQ) are used to ensure reliability. However, for a broadcast transmission, implementing ARQ can lead to network congestion with multiple users requesting different packets. Moreover, different users might lose different packets and retransmission could mean sending a large chunk of the original content again, leading to inefficient use of bandwidth as well as increased latency for some users. Application Layer Forward Error Correction (AL-FEC) is an error-correction mechanism in which redundant data is sent to facilitate recovery of lost packets. For this purpose, Raptor codes [17,18] were adopted in 3GPP TS 26.346 [19] as the AL-FEC scheme for MBMS delivery. Recently, improvements in the Raptor codes have been developed and an enhanced code called RaptorQ has been specified in RFC 6330 [20] and proposed to 3GPP. Streaming delivery (based on the H.264/AVC video codec and Real-time Transport Protocol (RTP)) over MBMS was studied in [21].

The forthcoming discussion presents the existing standardized framework in TS 26.346 [19] for live streaming of DASH-formatted content over eMBMS. The eMBMS-based live video streaming is over the FLUTE protocol [22] – file delivery over unidirectional transport – which allows for transmission of files via unidirectional eMBMS bearers. Each video session is delivered as a FLUTE transport object, as depicted in Figure 4.9. Transport objects are created as soon as packets come in. The IPv4/UDP/FLUTE header is a total of 44 bytes per IP packet. Protection against potential packet errors can be enabled through the use of AL-FEC. The AL-FEC framework decomposes each file into a number of source blocks of approximately equal size. Each source block is then broken into K source symbols of fixed symbol size T bytes. The Raptor/RaptorQ codes are used to form N encoding symbols from the original K source symbols, where $N > K$. Both Raptor and RaptorQ are systematic codes, which means that the original source symbols are transmitted unchanged as the first K encoding symbols. The encoding symbols are then used to form IP packets and sent. At the decoder, it is possible to recover the whole source block from any set of encoding symbols only slightly greater than K with a very high probability. Detailed comparisons between Raptor and RaptorQ are

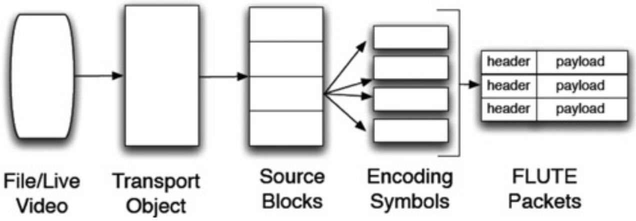


Figure 4.9 Transport-layer processing overview

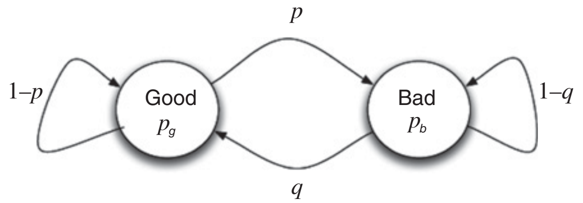


Figure 4.10 Markov model for simulating LTE RLC-PDU losses

presented in [23]. The choice of the AL-FEC parameters is made at the Broadcast Multicast Service Center (BMSC). For example, the BMSC has to select the number of source symbols K, the code rate K/N, and the source symbol size T. For a detailed discussion on the pros and cons of choosing these different parameters, the reader is referred to [24].

When simulating a live service, a long waiting time for encoding is not desirable. However, to ensure good Raptor/RaptorQ performance, a large value of K needs to be chosen. Thus, the minimum value of $K = K_{\min}$ is an important design parameter. A larger $K_{\min}$ causes a large startup delay, whereas a smaller $K_{\min}$ leads to poor performance. N encoding symbols are generated from K symbols using the AL-FEC (Raptor/RaptorQ) scheme. IP packets are then formed using these encoding symbols as payloads. The FLUTE packet is generated from the FLUTE header and payload containing the encoding symbols.

IP packets (RLC-SDUs (Service Data Units)) are mapped into fixed-length RLC-PDUs (Protocol Data Units). A 3GPP RAN1-endorsed two-state Markov model can be used to simulate LTE RLC-PDU losses, as shown in Figure 4.10. A state is good if it has less than 10% packet loss probability for the 1% and 5% BLER simulations, or less than 40% packet loss probability for the 10% and 20% BLER simulations.

The parameters in the figure are as follows: p is the transition probability from a good state to a bad state; q is the transition probability from a bad state to a good state; p_g is the BLER in a good state; p_b is the BLER in a bad state. It can be seen that the RAN model described above does not capture the coverage aspect of a cell, since it is the same for all users. For a more comprehensive end-to-end analysis, the following model can be used.

Instead of using a Markov model for all the users as above, a separate Markov model for each user in a cell can be used [24]. The received SINR data for each user is then used to generate a Multicast Broadcast Single-Frequency Network (MBSFN) sub-frame loss pattern. Such data can be collected for different MCS (Modulation and Coding Scheme) values. Using the sub-frame loss pattern for a given MCS, separate Markov models can be generated for each user in a cell. Note that this model is not fundamentally different from the RAN-endorsed model, but it accounts for the varying BLER distribution across users in a cellular environment. The BLER distribution depends on the specific deployment models and assumptions and could be different subject to different coverage statistics.

The performance bounds for eMBMS can be evaluated under different conditions. The bearer bit rate is assumed to be 1.0656 Mbits/s. Publicly available video traces can be used for video traffic modeling (http://trace.eas.asu.edu). Video traces are files mainly containing video-frame time stamps, frame types (e.g., I, P, or B), encoded frame sizes (in bits), and frame

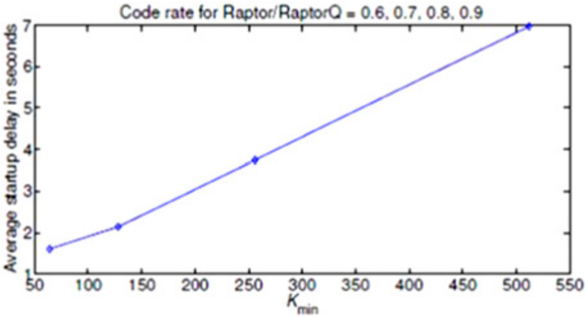


Figure 4.11 Startup delay as a function of $K_{\min}$

qualities (e.g., PSNR) in a Group of Pictures (GoP) structure. The length of an RLC-SDU is taken as 10 ms. The content length is set at 17,000 frames for each video trace. The video-frame frequency is considered to be 30 frames/s. The video frames are then used to generate source blocks and encoding symbols are generated using the AL-FEC framework (both Raptor/RaptorQ). The system-level simulations offer beneficial insights on the effect of system level and AL-FEC parameters on the overall QoE.

Different QoE metrics can be considered for multimedia delivery to mobile devices. In the case of file download or streaming of stored content, on user request, there is an initial startup delay after which streaming of video occurs and QoE can be measured by the initial startup delay and fraction of time that rebuffering occurs. The main contribution to startup delay for eMBMS live streaming is the AL-FEC encoding delay (i.e., when the service provider has to wait for a sufficient number of frames to be generated to ensure a large enough source block for efficient AL-FEC implementation). The source symbol size is chosen as $T = 16$ bytes. It is kept small in order to decrease the initial startup delay, so that a larger value of K can be chosen for the same source block.

The average startup delay (averaged over different code rates $K/N = 0, 6, 0.7, 0.8, 0.9$) is plotted in Figure 4.11 as a function of $K_{\min}$. As expected, the startup delay increases with increasing $K_{\min}$. The average PSNR of the received video stream is calculated using the offset trace file used for simulations. When a frame is lost, the client tries to conceal the lost frame by repeating the last successfully received frame. The rebuffering percentage is defined as the fraction of time that video playback is stalled in the mobile device. For live streaming, rebuffering occurs whenever two or more consecutive frames are lost. The client repeats the last successfully received frame and the video appears as stalled to the user. Video playback resumes as soon as one of the future frames is received successfully. The empirical Cumulative Density Function (CDF) of the PSNR and the rebuffering percentage for code rates 0.9 and 0.8 are shown in Figures 4.12 and 4.13, respectively. $K_{\min}$ is fixed to be 64. For detailed simulation parameters and algorithms, refer to [24]. It can be observed that improving the code rate improves the coverage from a QoE perspective, as it guarantees better PSNR and rebuffering for more users.

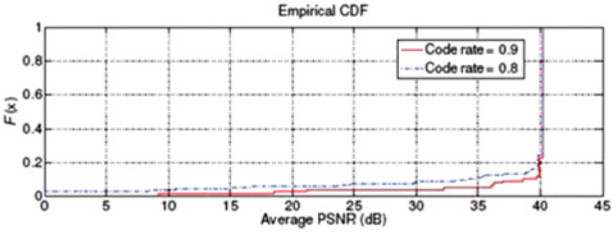


Figure 4.12 Performance comparisons for $K/N = 0.8, 0.9$: average PSNR

4.7 Server–Client Signaling Interface Enhancements for DASH

One of the most common problems associated with video streaming is the clients' unawareness of server and network conditions. Clients usually issue requests based on their bandwidth, unaware of the server's status which comprises factors such as

- the server's maximum upload rate and
- the number of clients streaming from the same server at the same time.

Thus, clients tend to request segments belonging to representations at the highest possible bit rates based on their perception, regardless of the server's condition. This kind of behavior often causes clients to compete for the available bandwidth and overload the server. As a result, clients could encounter playback stalls and pauses, which deteriorate QoE. Figure 4.14 shows a typical example of multiple clients streaming simultaneously. Initially, with only one streaming client, the available bandwidth for content streaming is high and the client gets the best possible quality based on his/her bandwidth. With more clients joining in the streaming process, clients starts to compete for the bandwidth and consequently the QoE drops. Greedy clients tend to eat up network bandwidth and stream at higher quality, leaving the rest of the clients to suffer much lower QoE.

Existing load-balancing algorithms blindly distribute bandwidth equally among streaming clients. However, an equal bandwidth-sharing strategy might not always be the best solution, since it may not provide the same QoE. For example, fast or complex-motion content, as in soccer or action movies, typically requires more bandwidth in order to achieve equal quality to low-motion content, such as a newscast.

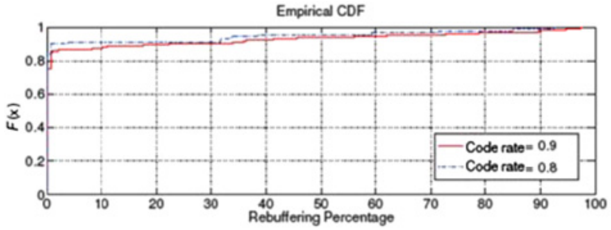


Figure 4.13 Performance comparisons for $K/N = 0.8, 0.9$: rebuffering percentage

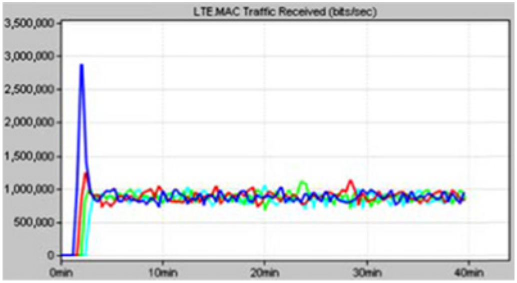


Figure 4.14 Download rate (throughput) for four clients streaming content from the same server and network

In our proposed solution, both the clients and the server share additional information through a feedback mechanism. Such information includes

- the average QoE measured at the client side and
- the server's upload rate.

The clients notify the server of their perceived QoE so far. This is in the form of statistics sent by the client regarding the client's average requested bit rate, average Mean Opinion Score (MOS), number of buffering events, etc. Other quality metrics can also be used. The server uses the client information in order to perform quality-based load balancing.

The server in return advises each client about the bandwidth limit it can request. In other words, the server can notify each client which DASH representations can be requested at any given time. This is achieved by sending the clients a special binary code, the Available Representation Code (ARC). ARC includes a bit for each representation, the Representation Access Bit (RAB), which can be either 0 or 1. The rightmost bit in the ARC corresponds to the representation with the highest bit rate, while the leftmost bit corresponds to the representation with the least bit rate.

As fluctuations to the server's upload rate occur, the server starts limiting the representations available to the clients. The server deactivates the representations available to clients in such a manner that at any point in time the maximum total bit rate requested by all clients does not exceed the server's upload rate. By defining such limits, the server remains at less risk of being overloaded and hence there are fewer delays in content transfer, leading to higher QoE at the streaming clients' side. The selection of which representation to be enabled or disabled is subject to server-based algorithms.

Different outcomes regarding the collective QoE of the streaming clients can be achieved depending on the algorithm selected for representation (de-)activation. In scenarios where the server gets overloaded with requests, limiting the representations available to clients can be useful in different ways. In the following sub-sections, two load-balancing approaches based on our server-assisted feedback will be described. These approaches are:

(a) Minimum QoE reduction. In this approach, the target of server load balancing is to minimize significant drops in quality for each user.
(b) Same average QoE. In this approach, the target is to have all clients perceive the same average QoE over time.

4.7.1 The Minimum Quality Reduction Approach

This algorithm's main focus is to minimize significant drops in quality on a per-user basis. In other words, higher priority is given to users who will experience a bigger gap in quality in case a representation is to be deactivated. This approach uses iterative steps to select the representation to be disabled or enabled when the server's upload rate changes or clients' requests exceed the maximum server upload rate. The procedure can be summarized as follows:

1. For each client, compute the quality reduction when the maximum available representation is disabled.
2. Disable the representation that causes minimum quality reduction.
3. Compute the sum of the bit rates of the maximum representation available at each client.
4. If the sum computed in step 3 is still higher than the available maximum upload rate, repeat steps 1 to 4; otherwise, go to step 5.
5. Stop.

As an example, Table 4.1 lists the typical quality changes experienced by two clients streaming different contents at different bit rates. Table 4.2 lists the corresponding outcome per iteration when the maximum upload rate changes from 2000 kbits/s to 1200 kbits/s.

In Table 4.1:

- $R_{\max}$ is the quality of the highest bit rate representation;
- PSNR is the corresponding average quality of each representation;
- ΔPSNR_i is the difference between the quality of the highest representation and representation i.

Table 4.1 Average PSNR per representation (bit rate) for two different contents

Representation		Client 1		Client 2	
I	R_i (kbits/s)	PSNR	$\Delta\text{PSNR}_i(R_{\max} - R_i)$	PSNR	$\Delta\text{PSNR}_i(R_{\max} - R_i)$
0	1000	46.13	—	37.97	—
1	800	43.70	2.43	37.49	0.48
2	600	40.57	5.56	36.80	1.17
3	400	36.58	9.55	34.49	3.48
4	300	32.14	13.99	32.55	5.42

Table 4.2 Tracing table for Algorithm 1 in case available bandwidth drops from 2000 kbits/s to 1200 kbits/s

Iteration number	Client 1			Client 2			Total bandwidth (kbps)
	$R_{i/\text{bit rate}}$	ΔPSNR_i	ARC	$R_{i/\text{bit rate}}$	ΔPSNR_i	ARC	
1	$R_{0/1000}$		11111	$R_{0/1000}$	0	11111	2000
2	$R_{1/800}$	2.43	11111	$R_{1/800}$	0.48	11110	1800
3	$R_{1/800}$	2.43	11111	$R_{2/600}$	1.17	11100	1600
4	$R_{1/800}$	2.43	11110	$R_{3/400}$	3.48	11100	1400
5	$R_{2/600}$	5.56	11110	$R_{3/400}$	3.48	11000	1200

The action performed at each iteration can be explained as follows:

1. At this stage, the total bandwidth requested by the clients (2000 kbits/s) exceeds that permissible by the server (1200 kbits/s) and thus a few representations need to be disabled.
2. The drop in quality experienced by disabling Client 2/R_0 is less than the drop in quality experienced by disabling Client 1/R_0 and thus Client 2/R_0 is deactivated by setting its RAB to 0.
3. The new total bandwidth (1800 kbits/s) is still higher than the maximum available, so Client 2/R_1 will be disabled as this still leads to a smaller drop in quality.
4. The new total bandwidth (1600 kbits/s) is still higher than the maximum available. At this point, the quality drop experienced by Client 1 upon deactivation of Client 1/R_1 is less than that experienced by Client 2 upon deactivating Client 2/R_2, thus Client 1/R_1 is disabled.
5. Given the new total bandwidth (1400 kbits/s), Client 2/R_2 is disabled and since the target bandwidth has now been reached, the algorithm stops.

Since the server is no longer overloaded, clients are at less risk of buffering stalls. In our approach, we used PSNR as a balancing criterion. PSNR values are pre-calculated for each DASH segment and stored in the DASH MPD. Other criteria, such as the MOS, can also be used with the same approach.

4.7.2 Same Average Quality Approach

The minimum quality reduction approach mainly exploits feedback sent from the server to clients but not the other way around. The approach discussed in this section exploits two-way feedback.

In this algorithm, the main focus is to set all clients to approximately the same average quality. The iterative procedure is described as follows:

1. Compute the sum of the bit rates of the maximum representation available at each client.
2. If the sum calculated in step 1 exceeds that supported by the server, go to step 3 (overflow phase). Else, if the sum is found to be less than the server's upload rate, go to step 4 (underflow phase).

3. Overflow phase:
   (a) Find the client with the highest average quality (the average quality value is computed by the client and sent to the server via a feedback mechanism). If the client has more than one representation activated go to step 3(b), else go to step 3(c).
   (b) i. Set client update flag to true.
      ii. Deactivate highest representation.
      iii. Replace the client's average quality with the quality corresponding to the average quality of the maximum representation permissible to that specific client.
      iv. If the sum of the bit rates of the maximum representation available at each client still exceeds the server's upload rate go to step 3(a), else stop.
   (c) Remove the client found in step 3(a) from the list of candidates. If the list of candidates is empty and the client update flag is false go to step 5, else if the list of candidates is empty but the client update flag is true then stop, else go to step 3(a).
4. Underflow phase:
   (a) Find the client with the lowest average quality. If the client has at least one deactivated representation and activating that representation will not cause an overflow, then
      i. Set client update flag to true.
      ii. Activate the representation.
      iii. Set the client's average quality to the average quality of the maximum representation permissible to him.
      iv. Go to step 4(c).

      Else, remove the client from the list of candidates and go to step 4(b).
   (b) If the list of candidates is empty and the client update flag is false then go to step 5, else if the list of candidates is empty but the client update flag is true then stop, else go to step 4(c).
   (c) Recompute the sum of the bit rates of the maximum representation available at each client. If an underflow still occurs go to step 4(a), else stop.
5. Stable phase:
   (a) Set MaxClient to the client with the highest average quality (MaxQ).
   (b) Set MinClient to the client with the least average quality (MinQ).
   (c) If the following conditions are satisfied go to step 5(d), else stop:
      – The difference between MaxQ and MinQ exceeds a specified threshold.
      – MaxClient has more than one activated representation.
      – MinClient has at least one deactivated representation.
      – The quality of the maximum representation available to MaxClient exceeds that available to MinClient.
   (d) Deactivate a representation from MaxClient. If the bandwidth saved as a result of the deactivation suffices to enable a representation for MinClient, a representation is activated. Go to step 4(c).

   Using the same values as in Table 4.1, an illustrative example is shown in Table 4.3 where the server's upload rate is also set to 1200 kbits/s.

The details of each iteration step can be explained as follows:

1. The PSNR values stated at the first iteration correspond to the average PSNR calculated by the clients and sent to the server. At this stage the total bandwidth that can be requested

Table 4.3 Tracing results for Algorithm 2 in case the bandwidth allowed drops from 2000 kbits/s to 1200 kbits/s

Iteration number	Client 1			Client 2			Total bandwidth (kbps)
	$R_{i/\text{bit rate}}$Max	PSNR$_i$	ARC	$R_{i/\text{bit rate}}$Max	PSNR$_i$	ARC	
1	$R_{0/1000}$	45	11111	$R_{0/1000}$	37	11111	2000
2	$R_{1/800}$	43.70	11110	$R_{0/1000}$	37.97	11111	1800
3	$R_{2/600}$	40.57	11100	$R_{0/1000}$	37.97	11111	1600
4	$R_{3/400}$	36.58	11000	$R_{0/1000}$	37.97	11111	1400
5	$R_{3/400}$	36.58	11000	$R_{1/800}$	37.49	11110	1200

by the clients exceeds that permissible by the server, and thus we need to disable a few representations.

2. Client 1 has higher average quality (PSNR = 45), thus a representation has been deactivated at his side and the average quality has been replaced by that of the highest representation permissible so far (PSNR = 43.70).

The algorithm continues until the sum of the highest bit rates permissible for each client does not exceed the server's upload rate. Using such an algorithm ensures that all clients stream at almost the same average quality.

Experimental results have verified that the use of server-assisted feedback approaches result in:

- Significant reduction in playback stalls at the client side as well as lower buffering time.
- Better QoE balancing between multiple clients.
- Better playback time. This is a side-effect of the reduction in buffering time, since clients usually continue to wait (or repeat the requests) until the required segment is retrieved.

On the contrary, there was little or no perceivable quality loss since clients – when aware of the server load condition – tended to request low-quality segments to avoid buffering events or long stalls.

4.8 Conclusion

We have given an overview of the latest DASH standardization activities at MPEG and 3GPP and reviewed a number of research vectors that we are pursuing with regard to optimizing DASH delivery over wireless networks. We believe that this is an area with a rich set of research opportunities and that further work could be conducted in the following domains.

(i) Development of evaluation methodologies and performance metrics to accurately assess user QoE for DASH services (e.g., those adopted as part of MPEG and 3GPP DASH specifications [1, 2]), and utilization of these metrics for service provisioning and optimizing network adaptation.

(ii) DASH-specific QoS delivery and service adaptation at the network level, which involves developing new Policy and Charging Control (PCC) guidelines, QoS mapping rules, and resource management techniques over radio access network and core IP network architectures.

(iii) QoE/QoS-based adaptation schemes for DASH at the client, network, and server (potentially assisted by QoE feedback reporting from clients), to jointly determine the best video, transport, network, and radio configurations toward realizing the highest possible service capacity and end-user QoE. The broad range of QoE-aware DASH optimization problems emerging from this kind of cross-layer cooperation framework includes investigation topics such as QoE-aware radio resource management and scheduling, QoE-aware service differentiation, admission control, QoS prioritization, and QoE-aware server/proxy and metadata adaptation.

(iv) DASH-specific transport optimizations over heterogeneous network environments, where content is delivered over multiple access networks such as WWAN unicast (e.g., 3GPP packet-switched streaming [3]), WWAN broadcast (e.g., 3GPP multimedia broadcast and multicast service [19]), and WLAN (e.g., WiFi) technologies.

References

[1] ISO/IEC 23009-1: ‘Information technology – dynamic adaptive streaming over HTTP (DASH) – Part 1: Media presentation description and segment formats.’

[2] 3GPP TS 26.247: ‘Transparent end-to-end packet switched streaming service (PSS); Progressive download and dynamic adaptive streaming over HTTP (3GP-DASH).’

[3] Sodagar, I., ‘The MPEG-DASH standard for multimedia streaming over the Internet.’ *IEEE Multimedia*, Oct/Dec, 2011, 62–67.

[4] Stockhammer, T., ‘Dynamic adaptive streaming over HTTP: Standards and design principles.’ Proceedings of ACM MMSys2011, San Jose, CA, February 2011.

[5] Oyman, O. and Singh, S., ‘Quality of experience for HTTP adaptive streaming services.’ *IEEE Communications on Magnetics*, **50**(4), 2012, 20–27.

[6] ISO/IEC 23009-2: ‘Information technology – dynamic adaptive streaming over HTTP (DASH) – Part 2: Conformance and reference software.’

[7] ISO/IEC 23009-3: ‘Information technology – dynamic adaptive streaming over HTTP (DASH) – Part 3: Implementation guidelines.’

[8] ISO/IEC 23009-4: ‘Information technology – dynamic adaptive streaming over HTTP (DASH) – Part 4: Segment encryption and authentication.’

[9] ITU-T Recommendation H.222.0|ISO/IEC 13818-1:2013: ‘Information technology – generic coding of moving pictures and associated audio information: Systems.’

[10] ISO/IEC 14496-12: ‘Information technology – coding of audio-visual objects – Part 12: ISO base media file.’

[11] 3GPP TS 26.234: ‘Transparent end-to-end packet switched streaming service (PSS); Protocols and codecs.’

[12] 3GPP TS 26.244: ‘Transparent end-to-end packet switched streaming service (PSS); 3GPP file format (3GP).’

[13] 3GPP TR 26.938: ‘Improved support for dynamic adaptive streaming over HTTP in 3GPP.’

[14] Akhshabi, S., Begen, A.C., and Dovrolis, C., ‘An experimental evaluation of rate-adaptation algorithms in adaptive streaming over HTTP.’ Proceedings of Second Annual ACM Conference on Multimedia Systems, San Jose, CA, 2011, pp. 157–168.

[15] Ramamurthi, V. and Oyman, O., ‘Link aware HTTP adaptive streaming for enhanced quality of experience. Proceedings of IEEE Globecom, Atlanta, GA, 2013.

[16] Ramamurthi, V. and Oyman, O., ‘Video-QoE aware radio resource allocation for HTTP adaptive streaming.’ Proceedings of IEEE ICC, Sydney, Australia, 2014 (to appear).

[17] Shokrollahi, A., ‘Raptor codes.’ Digital Fountain, Technical Report DR2003-06-001, June 2003.

[18] Luby, M., Shokrollahi, A., Watson, M., and Stockhammer, T., ‘Raptor forward error correction scheme for object delivery.’ RFC 5053 (proposed standard), IETF, October 2007.

[19] 3GPP TS 26.346: ‘Multimedia broadcast/multicast service (MBMS): Protocols and codecs.’ Third-Generation Partnership Project (3GPP), 2011. Available at: http://www.3gpp.org/ftp/Specs/archive/26 series/26.346/.

[20] Luby, M., Shokrollahi, A., Watson, M., Stockhammer, T., and Minder, L., ‘RaptorQ forward error correction scheme for object delivery.’ RFC 6330 (proposed standard), IETF, August 2011.

[21] Afzal, J., Stockhammer, T., Gasiba, T., and Xu, W., ‘Video streaming over MBMS: A system design approach.’ *Journal of Multimedia*, **1**(5), 2006, 23–35.

[22] Paila, T., Walsh, R., Luby, M., Roca, V., and Lehtonen, R., ‘File delivery over unidirectional transport.’ RFC 6726 (proposed standard), IETF, November 2012.

[23] Bouras, C., Kanakis, N., Kokkinos, V., and Papazois, A., ‘Evaluating RaptorQ FEC over 3GPP multicast services.’ 8th International Wireless Communications & Mobile Computing Conference (IWCMC 2012), August 27–31, 2012.

[24] Kumar, U., Oyman, O., and Papathanassiou, A., ‘QoE evaluation for video streaming over eMBMS.’ *Journal of Communications*, **8**(6), 2013, 352–358.

Acronyms

3GPP	Third-Generation Partnership Project
AL-FEC	Application Layer Forward Error Correction
ARQ	Automatic Repeat Request
AVC	Advanced Video Coding
BLER	Block Error Rate
BMSC	Broadcast Multicast Service Center
BS	Base Station
CDF	Cumulative Distribution Function
CQI	Channel Quality Information
DASH	Dynamic Adaptive Streaming over HTTP
DECE	Digital Entertainment Content Ecosystem
DLNA	Digital Living Network Alliance
DM	Device Management
DRM	Digital Rights Management
eMBMS	Enhanced MBMS
FLUTE	File Delivery over Unidirectional Transport
GMR	Gradient with Minimum Rate
HARQ	Hybrid ARQ
HAS	HTTP Adaptive Streaming
HbbTV	Hybrid Broadcast Broadband TV
HTTP	Hypertext Transfer Protocol
IETF	Internet Engineering Task Force
IP	Internet Protocol
IPTV	IP Television
ISOBMFF	ISO Base Media File Format
LTE	Long-Term Evolution
MBMS	Multimedia Broadcast and Multicast Service

MBSFN	Multicast Broadcast Single-Frequency Network
MCS	Modulation and Coding Scheme
MOS	Mean Opinion Score
MPD	Media Presentation Description
MPEG	Moving Picture Experts Group
NAT	Network Address Translation
OIPF	Open IPTV Forum
OMA	Open Mobile Alliance
PCC	Policy Charging and Control
PDP	Packet Data Protocol
PF	Proportional Fair
PFBF	Proportional Fair with Barrier for Frames
PSNR	Peak Signal-to-Noise Ratio
PSS	Packet-Switched Streaming Service
PVQ	Perceived Video Quality
PVR	Personal Video Recorder
QoE	Quality of Experience
QoS	Quality of Service
RAN	Radio Access Network
RLC	Radio Link Control
RTP	Real-time Transport Protocol
RTSP	Real-Time Streaming Protocol
SDU	Service Data Unit
SINR	Signal-to-Interference-and-Noise Ratio
SSIM	Structural Similarity
TCP	Transmission Control Protocol
TR	Technical Report
TS	Technical Specification
UE	User Equipment
URL	Uniform Resource Locator
WiFi	Wireless Fidelity
WLAN	Wireless Local Area Network
WWAN	Wireless Wide Area Network
W3C	World-Wide-Web Consortium
XLink	XML Linking Language
XML	Extensible Markup Language

5

No-Reference Approaches to Image and Video Quality Assessment

Anish Mittal[1], Anush K. Moorthy[2] and Alan C. Bovik[3]
[1]*Nokia Research Center, USA*
[2]*Qualcomm Inc., USA*
[3]*University of Texas at Austin, USA*

5.1 Introduction

Visual quality assessment as a field has gained tremendous importance in the past decade, as evinced by the flurry of research activity that has been conducted in leading universities and commercial corporations on topics that fall under its umbrella. The reason for this is the exponential growth of visual data that is being captured, stored, transmitted, and viewed across the world. Driving the previously unfathomable growth in communications, images and videos now form a major chunk of transmitted data. This is not surprising since, from the dawn of time, humans have been visual animals who have preferred images over the written word. One need only look at the amount of area devoted to visual signal processing in the human brain to surmise that vision and its perception forms a major chunk of neurological processing [1, 2]. Hence, researchers have attempted to decode human vision processing and have used models of the visual system for image-processing applications [3–7].

While an image can convey more than a thousand words, transmission of visual content occupies an equivalently large amount of bandwidth in modern communication systems, hence images and videos are compressed before storage or transmission. With increasing resolutions and user expectations, increasingly scarce bandwidth is being strained. While the user is concerned only with the final quality of the image or the video, the service provider attempts

Multimedia Quality of Experience (QoE): Current Status and Future Requirements, First Edition.
Edited by Chang Wen Chen, Periklis Chatzimisios, Tasos Dagiuklas and Luigi Atzori.

to provide said quality with the least possible expense of bandwidth. With increasing quality expectations, algorithm developers attempt to better image/video-processing algorithms so that visual quality at the output is enhanced. In all cases, however, the quantity being addressed is a highly subjective one, that of “visual quality” as perceived by a human observer. The goal of automated Quality Assessment (QA) is to attempt to produce an estimate of this human-perceived quality, so that such a quantitative measure may be used in place of subjective human perception. Our discussion here centers on quality assessment algorithms and their applications in modern image/video-processing systems.

Typically, QA algorithms are classified on the basis of the amount of information that is available to the algorithm. Full-Reference (FR) quality assessment algorithms require the distorted image/video whose quality needs to be accessed as well as the *clean,* pristine reference image/video for comparison [8–17], whereas Reduced-Reference (RR) approaches only use limited information regarding the reference image/video in lieu of the actual reference content itself, together with the distorted image [18–21]. Blind or No-Reference (NR) QA refers to automatic quality assessment of an image/video using an algorithm which only utilizes the distorted image/video whose quality is being assessed [22–45].

While tremendous progress has been made in understanding human vision, our current knowledge is far from complete. In the face of such incomplete knowledge, it is not surprising that researchers have focused on simpler FR algorithms [14,15,18]. FR algorithms themselves have many applications, especially in the field of image and video compression where the original image/video is available. Apart from applications, FR algorithms also provide researchers with a fundamental foundation on which NR algorithms can be built. The tools that proved useful in the FR framework have been modified suitably and later used for NR algorithms. For example, Natural-Scene-Statistics (NSS) models of images[16, 17] have been used successfully in recent NR algorithms [35–41,43–45]. Since FR research has been covered in this compendium and elsewhere [7,46,47], in this chapter we shall focus on NR algorithms.

One important aspect of quality assessment research is evaluating the performance of an algorithm. Since QA algorithms attempt to reproduce human opinion scores, it is obvious that a good algorithm is one that correlates well with human opinion of quality. In order to estimate human opinion of quality, large databases spanning a wide range of contents and visual impairments (such as those we detail later in this chapter) are created, and large-scale human studies are conducted. The human opinion scores thus produced represent the ground-truth and are used in evaluating algorithm performance. Apart from evaluating performance, these scores serve an additional, very important function in the case of NR algorithms.

NR QA approaches can be classified on the basis of whether the algorithm has access to subjective/human opinion prior to deployment. Algorithms could use machine learning techniques along with human judgments of quality during a “training” phase and then could attempt to reproduce human opinion during the “testing” phase. Such algorithms, which first learn human behavior from subjective quality data, are referred to as Opinion-Aware (OA) NR algorithms. Opinion-aware algorithms are the first step toward building a completely blind algorithm (i.e., one that is not only blind to the reference image, but also to the human opinion of quality). Such completely blind algorithms, which do not use subjective data on quality to perform blind quality assessment, are termed Opinion-Unaware (OU) algorithms. While both OA and OU algorithms have practical applications, OU NR algorithms hold more practical relevance. This is because it is impossible to anticipate all of the different distortions that may occur in a practical system. Further, no controlled database can possibly span all distortions and quality ranges well enough. Recent research in NR QA has spanned both OA and OU algorithms. While OA algorithms perform as well as FR algorithms, recent OU approaches have been catching up and perform exceedingly well without access to human opinion [39–45].

In this chapter, we shall detail recent no-reference approaches to image and video quality assessment. Specifically, we shall cover both opinion-aware and opinion-unaware models. Most of the approaches that we shall cover are based on understanding and modeling the underlying statistics of natural images and/or distortions using perceptual principles. These approaches measure deviations from statistical regularities and quantify such deviations, leading to estimates of quality. In this chapter, we shall analyze the motivation and the principles underlying such statistical descriptions of quality and describe the algorithms in detail. We shall provide exhaustive comparative analysis of these approaches and discuss the potential applications of no-reference algorithms. Specifically, we shall cover the case of distortion-unaware perceptual image repair and quality assessment of tone-mapped images. We then segue into a discussion of the challenges that lie ahead for the field to gain maturity and other practical application scenarios that we envision for these algorithms. We conclude the chapter with a discussion of some philosophical predictions of future directions that the field of automatic quality assessment of images and videos may take.

5.2 No-Reference Quality Assessment

Early research on no-reference quality assessment was focused toward predicting the quality of images afflicted with specific distortions [22–29]. Such approaches aim to model distortion-specific artifacts that can relate well to the loss in visual quality. For example, JPEG compressed images could be evaluated for their visual quality using the strength of edges at block boundaries [23, 24, 48, 49]. Such algorithms are restricted to the distortions they are designed for, limiting their scalability and usage in real scenarios, since the distortion type afflicting the image is almost never a known quantity. Having said that, these algorithms represented the first steps toward truly blind NR algorithms and hence form an important landmark in the field of quality assessment.

The next leap in NR algorithms was that of developing distortion-agnostic approaches to QA. These algorithms can predict the quality of the image without information on the distortion afflicting the image [30–45]. Some of these models are also capable of identifying the distortion afflicting the image [35, 39]; information that could potentially be used for varied applications [50].

Early general-purpose, distortion-unaware NR QA algorithms were developed using models that can learn to predict human judgments of image quality from databases of human-judged distorted images [30–39]. Such OA models, which use distorted images with co-registered human scores, have been shown to deliver high performance on images corrupted with different kinds of distortions and severities [51]. As we have mentioned before, while these algorithms are tremendously useful, they may be limited in their application since they are limited by the distortions that they are trained on, as well as bound by the quality ranges that the controlled training database has on offer. Since no controlled database can completely offer the variety of practical distortions, the algorithms trained on these databases remain partially crippled. This is not to say that the underlying model used to “train” these algorithms is at fault; in fact, as with the FR-to-NR transition, the models used for OA NR algorithms have been tweaked and utilized for opinion-unaware approaches.

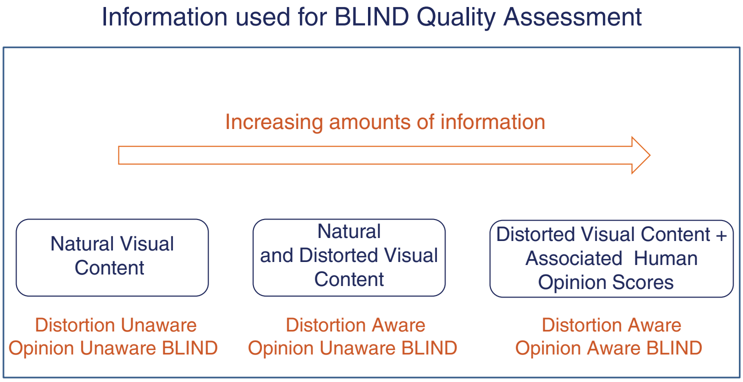

Information used for BLIND Quality Assessment
Increasing amounts of information
Natural Visual Content
Natural and Distorted Visual Content
Distorted Visual Content + Associated Human Opinion Scores
Distortion Unaware Opinion Unaware BLIND
Distortion Aware Opinion Unaware BLIND
Distortion Aware Opinion Aware BLIND

Figure 5.1 Blind quality assessment models requiring different amounts of prior information

OU approaches predict the visual quality of images without any access to human judgments during the training phase [39–45]. During the training phase, these algorithms may or may not have access to the distorted images. OU approaches which are trained on distorted images are limited by the distortion types and are referred to as Distortion-Aware (DA) algorithms. OU DA algorithms may be viewed as close cousins of the OA approaches, since both of them are limited by the training data available. OU Distortion-Unaware (DU) algorithms are those that do not utilize distorted images during the training phase. Since OU DU algorithms have no access to any distorted image or human opinion *a priori,* they represent a final frontier in NR QA research. OU DU algorithms find applications in uncontrolled environments such as highly unpredictable wireless networks or quality assessment of user-captured photographs and so on. It may seem that these algorithms have almost no information to make judgments on quality, but researchers find motivation in the fact that leading FR IQA models (such as the Structural SIMilarity index, SSIM [14]) are both opinion and distortion-unaware. Since one of the goals of NR QA research is to develop algorithms that could replace FR algorithms, OU DU NR algorithm development has caught the fancy of researchers in recent years [43–45]. A summary of NR algorithm classification is given in Figure 5.1.

5.2.1 Opinion-Aware Distortion-Aware Models

OA DA approaches make use of both distorted images and associated human judgments to develop QA models. Depending on the types of features they extract from the image/video, they can be categorized into codebook-based, ensemble-based, and Natural Scene Statistics (NSS)-based approaches.

5.2.1.1 Codebook-Based Approaches

The authors of [31, 32] use Gabor filter responses, based on local appearance descriptors to form a visual codebook post-quantization. The codebook feature space is then used to yield an estimate of quality. This is accomplished in one of two ways: (a) an example-based method or (b) a Support Vector Regression (SVR)-based method. The example-based method estimates the quality score of the test image using a weighted average of the quality scores of training images, where the authors assume that there exists a linear relationship between codeword histograms and quality scores. In contrast, the SVR-based method "learns" the mapping between the codeword histograms and the quality scores during the "training" phase. The approach is competitive in performance with other general-purpose NR IQA measures. However, its computational complexity limits its use in practical applications.

5.2.1.2 Ensemble-Based Approaches

Tang *et al.* [34] proposed an approach which learns an ensemble of regressors trained on three different groups of features – natural image statistics, distortion texture statistics, and blur/noise statistics. These regressors learn the mapping from feature space to quality and when deployed during the test phase, the algorithm reproduces the quality of the image using a combination of learned regressors. Another approach is based on a hybrid of curvelet, wavelet, and cosine transforms [52]. Although these approaches work on a variety of distortions, each set of features (in the first approach) and transforms (in the second approach) caters only for certain kinds of distortion processes, thereby limiting the deployment of these algorithms.

5.2.1.3 Natural Scene Statistics-Based Approaches

NSS-based approaches work on the rationale that all natural scenes obey statistical laws that are independent of the content of the scene being imaged. For instance, local quantities such as contrast are scale invariant in nature and follow heavy-tailed distributions [53]. In the case of distorted images, however, such scene statistics deviate from natural distributions, rendering them unnatural. These deviations, when quantified appropriately, can be utilized to evaluate the quality of the distorted image. This strategy of approaching the problem of NR QA from a natural scene perspective instead of a distortion-based perspective makes NSS-based NR approaches much less dependent on distortion-specific characteristics such as blocking. NSS models have proved to be very powerful tools for quality assessment in general, and have been used successfully for developing FR QA algorithms [16, 17, 25] and RR algorithms [54] in the past as well.

Recently, three successful blind/NR QA approaches to Image Quality Assessment (IQA) based on NSS were proposed [35–37], which exploit different NSS regularities in wavelet, DCT, and spatial domains, respectively. An NR IQA model developed in the wavelet domain, dubbed the Distortion Identification-based Image Verity and INtegrity Evaluation (DIIVINE) index, makes use of a series of statistical features derived from an NSS wavelet coefficient model. These features are subsequently used in the two-stage framework for QA, where the first stage identifies the distortion type and the second stage performs distortion-specific quality assessment [35]. While the approach performs at par with some of the successful FR IQA algorithms, the expensive computation of spatial correlation-based features makes it impractical for use in real-world applications. Their modification on the lines of proposed pairwise product-based features in the Blind/Referenceless Image Spatial QUality Evalu-ator (BRISQUE) model can alleviate the problem [37].

The DCT domain-based approach – BLInd Image Notator using DCT Statistics (BLIINDS-II index) – computes a small number of features from an NSS model of block DCT coefficients

[36]. Such NSS features, once calculated, are supplied to a regression function that predicts human judgments of visual quality. In comparison with DIIVINE, BLIINDS-II is a single-stage algorithm and instead of training multiple distortion-specific QA models, it only makes use of a single NSS model that is able to deliver highly competitive QA prediction power. Although BLIINDS-II uses a small number of features (4), the non-linear sorting of features makes the approach-computationally complex.

The third approach – BRISQUE – was developed with the express purpose of efficiency [37]. BRISQUE explores the possibility of a transform-free approach and operates directly on spatial pixel data. BRISQUE is based on the spatial NSS model of Ruderman [55] and uses pointwise statistics of locally normalized luminance signals and distribution of pairwise products of neighboring locally normalized luminance signals as features. Once these features are computed, a mapping from features to human judgment is learned using a regression module, yielding a measure of image quality.

5.2.2 Opinion-Unaware Distortion-Aware Models

This section summarizes approaches to NR QA that do not use human opinion scores to design QA models - OU DU NR QA algorithms. The advantage in such a scheme is that these approaches are not limited by the size of the databases with human judgments, thereby increasing their versatility. OU DA algorithms could either use large databases of reference and distorted images along with the corresponding FR algorithm scores as a proxy for human judgments, or in the ideal case, use only a set of pristine and distorted images together with the associated distorted categories.

5.2.2.1 Visual Words-Based Quality Assessment

Approaches based on visual words first decompose images using an energy-compacting filter bank and then divisively normalize the responses, yielding outputs that are well modeled using NSS models [41, 43]. Once such a representation is obtained, the image is divided into patches, and perceptually relevant NSS features are computed at each image patch. Features are computed from image patches obtained from both reference and distorted images to create distributions over visual words. Quality prediction is then accomplished by computing the Kullback–Leibler (KL) divergence between the visual word distribution of the distorted image and the signature visual word distribution of the space of exemplar images. One drawback of such an approach is that the creation of these visual word distributions from the features is lossy owing to the quantization involved in the process and could potentially affect predictions. Further, the approach is only as good as the diversity in the chosen training set of images and distortions and may not generalize to other distortions.

5.2.2.2 Topic Model-Based Learning for Quality Assessment

Algorithms based on topic models work on the principle that distorted images have certain latent characteristics that differ from those of “pristine” images [40]. These latent characteristics are explored through application of a "topic model" to visual words extracted from an assorted set of pristine and distorted images. Visual words, which are obtained using quality-aware NSS features, are used to determine the correct choice of latent characteristics, which are in turn capable of discriminating between pristine and distorted image content. The similarity between the probability of occurrence of the different topics in an unseen image and the distribution of latent topics averaged over a large number of pristine natural images is indicative of the image quality. The advantage of this approach is that it not only predicts the visual quality of the image, but also discovers groupings amongst artifacts of distortions in the corrupted images without any supervision. Unfortunately, in its current form, the approach does not perform as well as general-purpose OA models.

5.2.2.3 Full-Reference Image Quality Assessment-Based Learning for Quality Assessment

These approaches use quality scores produced by the application of an FR IQA algorithm on each distorted reference image pair in a training database as a proxy for human judgments of quality [42]. Distorted images and their reference versions are first partitioned into patches and a percentile pooling strategy is used to estimate the quality of each patch. The patches are then grouped into different groups based on their quality levels using special clustering techniques [56]. Quality-aware clustering is then applied to each group to learn the quality-aware centroids. Each patch of the distorted image is compared with the learned quality-aware centroids during the testing stage and the final quality score is assigned based on a simple weighted average. The score for each patch, once obtained, can be pooled to obtain the quality, of the image. This approach shows high correlation with respect to human judgments of image quality, and also high efficiency. As with all OA DU models, the approach is limited by the database of distortions that it is trained on and the quality estimates of the FR algorithm used as the proxy for human judgments.

5.2.3 Opinion-Unaware Distortion-Unaware Models

A holy grail of the blind/NR IQA problem is that of the design of perceptual models that can predict the quality of distorted images with as little prior knowledge of the images or their distortions as possible [41, 43, 44]. This section discusses the first steps toward completely blind approaches to NR QA (i.e., those approaches that do not make use of any prior distortion knowledge or subjective opinion of quality to predict the quality of images – OU DU models).

5.2.3.1 Visual words-based Quality Assessment

This approach is an extension of the OU DA visual word-based NR QA model of [41, 43]. While the crux of the approach remains the same as that summarized above, the DU extension differs in the way the visual words are formed during the training stage of the algorithm. In the DU case, instead of using both the distorted and reference images, the model uses only the natural undistorted reference images to form the visual codebook.

The feature-to-visual-word-distribution conversion is a lossy process due to the quantization involved and affects the accuracy of human judgment prediction. The approach described below overcomes these shortcomings and delivers performance on a par with the top-performing FR and NR IQA models that require training on human-judged databases of distorted images.

5.2.3.2 Multi-Variate Gaussian (MVG) Model-Based Quality Assessment

This NR OU-DU IQA model [44], dubbed Naturalness Image Quality Evaluator (NIQE), is based on constructing a collection of quality-aware features. This approach is based on the principle that all "natural" images captured are distorted in some form or another. For instance, during image capture, there is always a loss of resolution due to the low-pass nature of the lens; further, there exists defocus blur in different parts of the image depending on the associated depth of field. Since humans appear to more heavily weight their judgments of image quality from image regions which are in focus and hence appear sharper, salient quality measurements can be made from "sharp" patches in an image. From amongst a collection of natural patches, this approach uses data only from those patches that are richest in information (i.e., those that are less likely to have been subjected to a limiting distortion such as blur). The algorithm extracts quality-aware perceptual features from these patches to construct an MVG model. The quality of a given test image is then expressed as the distance between an MVG fit of the NSS features extracted from the test image whose quality is to be assessed, and an MVG model of the quality-aware features extracted from the corpus of natural images. Experimental results demonstrate that this approach performs as well as top-performing FR IQA models that require corresponding reference images and NR IQA models that require training on large databases of human opinions of distorted images (i.e., OA DU models).

5.2.4 No-Reference Video Quality Assessment

While there has been a lot of activity in the area of distortion-agnostic image quality assessment, the field of distortion-agnostic No-Reference Video Quality Assessment (NR VQA) has been relatively quiet. This is not to say that NR VQA algorithms do not exist, but most existing models are specific to the application that they were designed for and hence do not find widespread use. The popularity of distortion-specific measures for NR VQA can be attributed to two factors. First, VQA is a far more complex subject than IQA, and given that distortion-unaware models for NR IQA have just started to appear, such models for NR VQA are still forthcoming. Second, the lack of a universal measure for NR VQA and the need for blind quality assessment in many applications necessitates the development of such distortion-specific measures. Since this chapter has focused on distortion-specific NR IQA measures, and we believe that distortion-agnostic NR VQA algorithms are just around the corner, in this section we do not expressly summarize each proposed approach. Instead, we point the reader to relevant work in the area and to popular surveys which describe the algorithms in far more detail.

One of the most popular distortions that has been evaluated in the area of NR VQA is that of compression. Since video compression has been a popular subject for research and has tremendous practical applications, it comes as no surprise that many NR VQA metrics are geared toward compression. One artifact of compression that is commonly quantified is that of blocking [57–60], where edge strength at block boundaries is measured and correlated with perceptual quality. Techniques used include harmonic analysis based on Sobel edge detection [57], looking for correlations at 8 × 8 block boundaries in MPEG video [58], and using luminance masking along with edge-strength measures [59]. Another distortion that manifests due to compression is that of blur [61]. Jerkiness, which may be an artifact of the encoder, has also been evaluated for its quality using several techniques [62–66].

Researchers have also studied the quality effect of multiple coincident distortions on videos. The combination of blocking, ringing, and sharpness is evaluated in [67], where each distortion is evaluated using a technique that quantifies the distortion indicator. Packet-loss artifacts and their effect on visual quality, along with that of blocking due to compression, was studied in [68]. Another measure that evaluated the effect of blockiness, bluriness, and noisiness on visual quality was proposed in [69]. Apart from blocking and blurring, the authors of [70] evaluated the effects of motion artifacts on visual quality for 2.5G/3G systems. Bit-error impairments, noise, and blocking have been studied in [71], as has a motion-compensated approach [72].

Apart from spatial measures, researchers have also evaluated the effect of temporal quality degradations on visual perception. For example, motion information to evaluate the effect of frame-drop severity can be extracted from time-stamp information as in [65]. The authors segment the video and determine the motion activity for each of these segments, which is then used to quantify the significance of frame drops in the segments. Another measure of frame-drop quality is that in [73], where the discontinuity along time is computed using the mean-squared error between neighboring frames. The authors of [74] studied the effect of error concealment on visual quality. Channel distortion was modeled in [75], where information from macro blocks was used to quantify loss and its effect on visual quality. Jerkiness between frames was modeled using absolute frame differences between adjacent frames in [76].

The authors of [77] evaluate a multitude of factors to quantify overall qualities such as blur, blockiness, activity and so on. While this method still uses distortion-specific indicators of quality, it is one of the few NR VQA algorithms that does not limit itself to a particular application. A survey of such distortion-specific measures appears in [78].

The only truly distortion-agnostic measure for video quality was recently proposed in [79]. This algorithm is based on the algorithm for blind image quality assessment proposed in [36] and uses the principle of natural video statistics. As in the case of images, the assumption is that natural videos have certain underlying statistics that are destroyed in the presence of distortion. The deviation from naturalness is quantified to produce a quality measure. In [79], the authors use DCT coefficient differences between adjacent frames as the basis, and extract block-motion estimates to quantify motion characteristics. Features are then extracted by weighting DCT differences by the amount of motion and its effect on visual perception. The method is generic in nature and hence could be applied to a wide variety of distortions. The authors demonstrate that the algorithm quantifies quality for compression and packet loss using a popular video quality assessment database [80].

5.3 Image and Video Quality Databases

The performance of any IQA/VQA model is gauged by its correlation with human subjective judgments of quality, since the human is the ultimate receiver of the visual signal. 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. 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 (MOS/DMOS). Practical application of QA algorithms requires that these algorithms compute perceptual quality *efficiently*. In this section, we summarize the available image and video databases.

5.3.1 Image Quality Assessment Databases

- **LIVE Multiply Distorted Image Quality Database**. A subjective study [81] was conducted in two parts to obtain human judgments on images corrupted under two multiple distortion scenarios. The two scenarios considered are: (1) image storage, where images are first blurred and then compressed by a JPEG encoder; (2) camera image acquisition process, where images are first blurred due to a narrow depth of field or other defocus and then corrupted by white Gaussian noise to simulate sensor noise.
- **LIVE Image Quality Database**. The LIVE database [82] developed at the University of Texas at Austin, TX, contains 29 reference images and 779 distorted images at different image resolutions ranging from 634 × 438 to 768 × 512 pixels. Reference images are simulated with five different types of distortions to varying degrees – JPEG compression, JPEG2000 compression, additive Gaussian white noise, Gaussian blurring, and fast fading distortion, where the JPEG2000 compression bit stream is passed through a simulated Rayleigh fading channel. Human judgments were obtained from 29 subjects.
- **IRCCyN/IVC Image Quality Database (IVC)**. The IRCCyN/IVC database [83] developed at the Institut de Recherche en Communications et Cyberntique de Nantes (IRCCyN), France contains 10 reference images and 185 distorted images at an image resolution of 512 × 512 pixels. There are five distortion types in this database: JPEG compression, JPEG compression of only the luminance component, JPEG2000 compression, locally adaptive resolution coding, and Gaussian blurring. Each type of distortion was generated with five different amounts of distortion. Human judgments were collected from 15 subjects.
- **Tampere Image Quality Database**. The Tampere database [84] developed at the Tampere University of Technology, Finland contains 25 reference images and 1700 distorted versions of them at a resolution of 384 × 512 pixels. There are 17 different distortion types in the database including different types of noise, blur, denoising, JPEG and JPEG2000 compression, transmission of JPEG, JPEG2000 images with errors, local distortions, luminance, and contrast changes. Each type of distortion was generated with four different amounts. Human opinions were obtained from 838 subjects.
- **Categorical Subjective Image Quality (CSIQ) Database**. The CSIQ database [85] developed at Oklahoma State University, OK contains 30 reference images and 866 distorted images at a resolution of 512 × 512 pixels. Six distortion types are present in this database: JPEG compression, JPEG2000 compression, additive Gaussian white noise, additive Gaussian pink noise, Gaussian blurring, and global contrast decrements where each distortion type was generated with four or five varying degrees. The ratings were obtained from 35 subjects.

- **The Real Blur Image Database (RBID)**. The RBID [86] developed at the Uni-versidade Federal do Rio de Janeiro, Brazil, contains 585 blurred images captured from a real camera device with image sizes ranging from 1280 × 960 to 2272 × 1704 pixels. The images in this database are categorized into five different blur classes: unblurred, out of focus, simple motion, complex motion, and others. The ratings were collected from 20 subjects.

5.3.2 Video Quality Assessment Databases

- **LIVE Video Quality Database**. The LIVE VQA database [80] consists of 150 distorted videos created from 10 reference videos and span distortions such as MPEG-2 and H.264 compression, and simulated transmission of H.264 streams over IP and wireless channels. The videos in this database are at a resolution of 768 × 432 pixels and, as of this writing, the LIVE VQA database is a de-facto standard database to test the performance of any new VQA algorithm.
- **EPFL-PoliMI VQA Database**. Consisting of 156 video streams at CIF and 4CIF resolutions, this database incorporates packet-loss distortions due to transmission of H.264 streams over error-prone networks [87, 88]. With 40 subjects taking part in the study, this database can be used to measure the performance of VQA algorithms.
- **IRCCyN/IVC HD Video Database**. This database consists of a total of 192 distorted videos of 9–12 s duration at 1080i @ 50 fps [89]. The distorted videos were created using H.264 compression at different bit rates and were rated by 28 subjects using the Absolute Category Rating (ACR) scale [90]. The database also includes human opinion scores from a SAMVIQ test methodology [89].
- **MMSP Scalable Video Database**. Developed by researchers at EPFL, this scalable video database consists of compressed videos created by using two different scalable video codecs [91, 92]. Three HD videos were processed at three spatial and four temporal resolutions to create a total of 72 distorted videos which were rated using both a paired comparison and a single-stimulus one [90].
- **LIVE Mobile VQA Database**. Consisting of 200 distorted videos from 10 RAW HD (720p) reference videos, the LIVE Mobile VQA database is the first of its kind where temporal distortions such as time-varying video quality and frame-freezes of varying duration accompany the traditional compression and packet-loss distortion [93]. The videos were viewed and rated by human subjects on two mobile devices - a cellphone and a tablet - to produce the DMOS. With over 50 subjects, and 5300 summary subjective scores and time-sampled traces of quality, the LIVE Mobile VQA database is one of the largest and newest publicly available databases as of today.

A comparison of many of the databases mentioned here, as well as others, appears in [94]. The author of [94] also maintains a comprehensive list of image and video quality databases [95].

5.4 Performance Evaluation

In this section we summarize the performance of some of the image quality assessment algorithms discussed in this chapter and compare their performance to that of leading full-reference algorithms. We do not report NR VQA performance, since there exists only one truly blind, distortion-agnostic NR VQA algorithm [79].

We use the LIVE IQA database [82], which consists of 29 reference images and 779 distorted images spanning five different distortion categories - JPEG and JPEG2000 (JP2K) compression, additive white Gaussian noise (WN), Gaussian blur (blur), and a Rayleigh fast-fading channel distortion (FF) as the test bench. The database provides associated DMOS for each distorted image, representing its subjective quality. We list the performances of the three FR indices, Peak-Signal-to-Noise Ratio (PSNR), single-scale Structural SIMilarity index (SSIM) [14], and the Multi-Scale Structural SIMilarity index (MS-SSIM) [15]. While the PSNR is often criticized for its poor performance, it forms a baseline for QA algorithm performance and is still widely used to quantify visual quality, while the latter is a leading FR algorithm with high correlation with human performance – a goal that all NR algorithms attempt to achieve.

The NR algorithms evaluated are: CBIQ [31], LBIQ [34], BLIINDS-II [36], DIIVINE [35], and BRISQUE [96], all OA-DA algorithms; TMIQ [40], an OU-DA approach and NIQE [44], an OU-DU algorithm. The correlations for CBIQ [31] and LBIQ [34] were provided by the authors.

Since all of the IQA approaches that we compare require a training procedure to calibrate the regressor module, we divided the LIVE database randomly into chosen subsets for training and testing. While the OU approaches do not require such training (they are trained on an alternate database of distorted + natural images or natural images only, as summarized before), and

FR approaches do not need any training at all, to ensure a fair comparison across methods, the correlations of predicted scores with subjective opinion of visual quality are only reported on the test set. The dataset was divided into 80% training and 20% testing such that no overlap occurs between train and test content. This train–test procedure was repeated 1000 times to ensure that there was no bias due to the spatial content used for training. We report the median performance across all iterations.

We use Spearman's Rank-Ordered Correlation Coefficient (SROCC) and Pearson's (Linear) Correlation Coefficient (LCC) to test the model. SROCC and LCC represent the correlation between algorithm and subjective scores so that a value of 1 indicates perfect correlation. Since LCC is a linear correlation, all algorithm scores are passed through a logistic non-linearity [82] before computing LCC for mapping to DMOS space. This is a standard procedure that is used to detail the algorithm performance of QA algorithms. The SROCC and LCC values of the algorithms summarized in this chapter are tabulated in Tables 5.1 and 5.2, respectively.

The correlations demonstrate that OA-DA algorithms correlate well with human perception approaching the performance of leading full-reference algorithms such as MS-SSIM and besting popular measures such as PSNR. While the performance of TMIQ (the OU-DA measure) is not as good, the results are encouraging, especially for certain distortions such as JP2K and JPEG compression, where the performance is close with that of PSNR. The OU-DU measure of NIQE produces a performance comparable with that of the OA-DA measures and hence to that of successful FR IQA algorithms. While there remains room for improvement, the tables demonstrate that state-of-the-art NR algorithm performance is comparable with that of FR algorithms and NR approaches can be used successfully in place of FR approaches.

Table 5.1 Median SROCC across 1000 train–test combinations on the LIVE IQA database. *Italics* indicate (OA/OU)-DA no-reference algorithms and **bold face** indicates OU-DU model algorithms

	JP2K	JPEG	WN	Blur	FF	All
PSNR	0.8646	0.8831	0.9410	0.7515	0.8736	0.8636
SSIM	0.9389	0.9466	0.9635	0.9046	0.9393	0.9129
MS-SSIM	0.9627	0.9785	0.9773	0.9542	0.9386	0.9535
CBIQ	*0.8935*	*0.9418*	*0.9582*	*0.9324*	*0.8727*	*0.8954*
LBIQ	*0.9040*	*0.9291*	*0.9702*	*0.8983*	*0.8222*	*0.9063*
BLIINDS-II	*0.9323*	*0.9331*	*0.9463*	*0.8912*	*0.8519*	*0.9124*
DIIVINE	*0.9123*	*0.9208*	*0.9818*	*0.9373*	*0.8694*	*0.9250*
BRISQUE	*0.9139*	*0.9647*	*0.9786*	*0.9511*	*0.8768*	*0.9395*
TMIQ	*0.8412*	*0.8734*	*0.8445*	*0. 8712*	*0.7656*	*0.8010*
NIQE	**0.9172**	**0.9382**	**0.9662**	**0.9341**	**0.8594**	**0.9135**

5.5 Applications

This section summarizes some of the possible applications of NR QA that have been explored in the recent past. These applications are only meant to be representative and are in no way exhaustive. The section serves as a reference for the reader and is only a starting point in understanding visual quality applications.

5.5.1 Image Denoising

One topic that has received a large amount of interest from the image processing community is image denoising, where the goal is to "remove" the noise from the image, thereby making it "cleaner" [97]. Although a lot of progress has been made in the development of sophisticated denoising models, blind image denoising algorithms [96, 98] which denoise the image

Table 5.2 Median LCC across 1000 train–test combinations on the LIVE IQA database. *Italics indicate* (OA/OU)-DA no-reference algorithms and bold face indicates OU-DU model algorithms

	JP2K	JPEG	WN	Blur	FF	All
PSNR	0.8762	0.9029	0.9173	0.7801	0.8795	0.8592
SSIM	0.9405	0.9462	0.9824	0.9004	0.9514	0.9066
MS-SSIM	0.9746	0.9793	0.9883	0.9645	0.9488	0.9511
CBIQ	*0.8898*	*0.9454*	*0.9533*	*0.9338*	*0.8951*	*0.8955*
LBIQ	*0.9103*	*0.9345*	*0.9761*	*0.9104*	*0.8382*	*0.9087*
BLIINDS-II	*0.9386*	*0.9426*	*0.9635*	*0.8994*	*0.8790*	*0.9164*
DIIVINE	*0.9233*	*0.9347*	*0.9867*	*0. c9370*	*0.8916*	*0.9270*
BRISQUE	*0.9229*	*0.9734*	*0.9851*	*0.9506*	*0.9030*	*0.9424*
TMIQ	*0.8730*	*0. 8941*	*0.8816*	*0.8530*	*0.8234*	*0.7856*
NIQE	**0.9370**	**0.9564**	**0.9773**	**0.9525**	**0.9128**	**0.9147**

without knowledge of the noise severity remain relatively underexplored. Such blind denoising approaches generally combine blind parameter estimation with denoising algorithms to yield completely blind image denoising algorithms.

Initial approaches to blind denoising made use of empirical strategies such as L-curve methods [99–102], the discrepancy principle [103], cross validation [103–108], and risk-based estimates [109–113] of the reference image for parameter optimization. The use of perceptual optimization functions has been shown to yield better parameter estimates [114].

NSS-based blind image denoising approaches seek to reduce the amount of noise in a corrupted image without knowledge of the noise strength [98, 115]. In these approaches, the parameter being estimated is the noise variance, since most denoising algorithms assume that the noise is Gaussian in nature with zero mean and unknown variance. Although the approaches in [98, 115] discuss the estimation of the noise variance parameter only, they can be used to estimate other parameters, depending on the underlying noise model assumed in the image denoising algorithm used.

In [115], the authors exploit content-based statistical regularities of the image. While the approach works well, it is exhaustive and computationally intensive. This is because the image is denoised multiple times using different values of the noise variance and the quality of each denoised image is estimated using a no-reference content-evaluation algorithm and the best image picked from this chosen is set as the output denoised image.

The approach in [98] uses a different strategy. Here, the input parameter is estimated using statistical properties of natural scenes, where statistical features identical to those in [96] are extracted and then mapped onto an estimate of noise variance. One interesting observation that the authors make is that the denoised image produced when the algorithm is provided with the accurate noise variance estimate has lower perceptual quality (as measured by an algorithm) than one that is produced using a different (although incorrect) noise variance. The features extracted were hence designed so that the denoised image has the highest visual quality. During the training stage, given a large set of noisy images, each image was denoised with various values of the input noise variance parameter using [116], and its visual quality was evaluated using MS-SSIM [15]. The image having the highest perceptual quality as gauged by MS-SSIM [15] was selected, and the corresponding noise variance parameter was set as training input to the blind parameter estimation algorithm. This scheme is able to produce denoised images of higher visual quality, without knowledge of the actual noise variance even during the training stage. An extension of this approach would be to evaluate how this approach would perform while estimating parameters of other distortions such as blur or compression [50].

5.5.2 *Tone Mapping*

Most of the discussion in this chapter has focused on the quality of “regular” images/videos (i.e., those images and videos which have a limited dynamic range). In the recent past, High Dynamic Range (HDR) imaging techniques have gained tremendous popularity both in academia and in commercial products [117]. An HDR image is typically created from multiple “low” dynamic range images. For instance, in a camera, the exposure time is varied so that different parts of the incident spectrum of light are captured separately and then combined to form an HDR image. The saturation that is seen at the high or low end of the light spectrum in regular images is overcome by using the HDR technique. In order to combine multiple exposures, an algorithm called tone mapping is applied. The goal of a tone-mapping algorithm is to produce an HDR image that not only has a high perceptual quality, but also looks “natural.” One way to measure this is through a tone-mapping-specific IQA algorithm. Since HDR images have no traditionally defined “reference,” the problem is blind in nature. Recently, the authors of [118] proposed such a Tone-Mapping NR IQA algorithm called the tone-mapped image Quality Index (TMQI).

TMQI compares the low and high dynamic range images using a modification of the “structure” term of the SSIM [14] to account for the non-linear response of the human visual system. Such computation is undertaken at multiple scales, drawing inspiration from [15]. A naturalness term is also computed based on previously unearthed statistics of natural low-dynamic images [119,120]. A combination of the two measures leads to the TMQI. The authors demonstrate that the index correlates well with human perceptions of quality based on findings from previous subjective studies on HDR quality.

Apart from discussing the algorithm and evaluating its performance, the authors also demonstrate how the TMQI measure could be used to aid parameter tuning for tone-mapping algorithms. Another interesting use of the measure is its application in adaptive fusion of tone-mapped images, which takes advantage of multiple tone-mapping operators to produce a high-quality HDR image. The work in [118] is a pioneering effort in the field of tone-mapped quality assessment and demonstrates how useful NR QA algorithms could be in real-life applications.

5.5.3 Photo Selection

Photo capture is picking up at a very fast pace with the launch of new hand held devices and smart phones. Americans captured 80 billion digital photographs in 2011 [121], and this number is increasing annually. More than 250 million photographs are being posted daily on Facebook. Consumers are becoming overwhelmed by the amount of available digital visual content, and finding ways to review and control the quality of digital photographs is becoming almost impossible. With this objective in mind, a photo quality helper app [122] for android phones has also been designed recently as part of a senior design project by an undergraduate team at the University of Texas at Austin, which does an automatic judgment of photo quality and enables those photos which do not meet the user’s quality threshold to be discarded.

5.6 Challenges and Future Directions

While visual quality assessment has made giant leaps in the recent past, so much so that researchers have quickly moved on from intense research in FR QA to opinion and distortion-unaware NR QA, there remain multiple challenges that need to be addressed in order for the field to attain maturity. In this section, we discuss a handful of such challenges and speculate on possible research directions that may yield positive results.

- **Content Bias**. The discussion in this chapter has assumed that all content is equal. Algorithms for quality assessment are almost always content blind and only look at low-level feature descriptions such as texture to make quality judgments. Humans, however, relate to images at a much higher level. While distortions in texture and smooth regions definitely influence user perceptions of quality, the content of the image has an important role to play in the overall quality rating. Images that a subject likes, for example, a picture of the subject's baby, may be rated as higher quality even if the distortion on the image is unacceptable. Likes and dislikes may not always be personal. Sunsets, beaches, good-looking faces, etc. are universally liked. Current databases for visual quality assessment attempt to remove this bias by selecting content that does not necessarily invoke strong feelings of like or dislike. While this helps in understanding human perceptions of quality, real-life content seldom tends to be "boring." It is hence of interest to try to ferret out content preferences and use this as a seed in producing quality ratings. If the goal of visual quality assessment is to replace the human observer, human biases need to be modeled and the area of content bias needs to be explored.

- **Aesthetics**. Closely related to the content bias is the notion of aesthetics [123–128]. The aesthetics of an image or video capture how appealing a video is to a user. While there has been some research in the field of visual aesthetics [123, 124, 126–128] a lot more remains to be done. Current aesthetics algorithms more often than not can only classify images as pleasing and not-pleasing. Those which produce aesthetics ratings do not correlate with human perception as well as quality assessment algorithms. Apart from improving this performance, it is of interest to study the joint effect of aesthetics and distortions on an image. Is a pleasingly shot image as bad as a non-pleasingly shot one under the same distortion regime? This question ties in closely with content bias, for while aesthetics may be evaluated independently of content, an ideal algorithm that reproduces human behavior needs to understand and model human opinion on aesthetics, content, and quality, and all interaction between the three.
- **No-Reference Video Quality Assessment**. In comparison with advances in NR IQA, algorithms for NR VQA remain scarce. This is not without reason. The addition of motion information in the visual content complicates the problem of NR QA drastically. Indeed, even in the slightly easier problem of NR VQA, successful algorithms on one database do not necessarily translate into successful algorithms on another [80, 93]. The nature of the distortion in question, and its effect on motion information, needs to be studied in detail before any modeling effort for motion is undertaken. As can be imagined, this is not an easy task. One possible direction of effort that may yield results is that of task-specific quality assessment. For example, in the video teleconferencing use case, one is interested in producing high-quality facial images, where the face occupies a major portion of the screen. VQA algorithms could target the face and use it as the region-of-interest to gauge quality. Obviously, this will not translate to other use cases. However, given the lack of progress made in understanding motion models with respect to distortions, it may be worthwhile investigating application-specific models.
- **Human Behavior Modeling**. One aspect of video quality assessment that remains relatively unexplored is that of human behavioral modeling [129, 130]. When humans view a video, the overall satisfaction with the video is a function of the temporal dynamics that the human sees as the video plays out. The temporal variation in the quality as the user views the video, and its effect on overall quality perception, is an interesting field of study [93, 129]. For instance, one could ask the question: If a low-quality segment precedes a high-quality

segment, is the overall satisfaction higher than when the case is reversed? This is of course one of the many questions that can be asked and, in order to answer these questions, one needs databases that have been developed to study human responses to time-varying quality. It is only recently that such databases have started to appear [93]. Once human responses to temporal quality variation are decoded, models need to be designed to account for such human behavioral responses and incorporated into video quality assessment algorithms. The design of the databases that may answer these questions is in itself a pretty arduous task, and we do not expect the modeling to be any easier. However, once such models are available, the applications are tremendous. Apart from the obvious improvements to existing VQA algorithms, such models will allow for scheduling of video packets to be transmitted based on current and predicted future channel conditions. Scheduling visual information to maximize perceptual quality is a field that is still nascent, and remains of tremendous future interest.

5.7 Conclusion

In this chapter, we have summarized recent research in the field of no-reference or blind image and video quality assessment. The field of NR QA has recently witnessed a host of activity, and several high-performance algorithms have been developed to solve the problem of predicting human perceptions of quality without the need for a reference. The methods that we explored were broken down by the amount of information available to the algorithm during the "training" stage prior to deployment. As we summarized, even in the face of almost no information regarding the distortion type or human opinion on quality, there exist algorithms that predict visual quality with a fair degree of accuracy. This is an extremely encouraging fact, and we expect high-performing NR QA algorithms to approach the level of full-reference correlation on the near future. While great progress has been made on the image quality front, no-reference video quality assessment has received less attention. As we summarized, the complexity of motion and its interaction with distortion make NR VQA a difficult problem to solve.

While the field is certainly maturing, it is far from mature. We summarized a handful of challenges that the field of quality assessment needs to address. Primary amongst the challenges is that of deployment of QA algorithms in practical applications. While QA algorithms have been researched painstakingly, applying QA algorithms to traditional image-processing problems is still under explored. This is true in the case of FR QA algorithms [114], but even more so in the case of NR QA algorithms [50]. As NR QA algorithm performance peaks, we hope that the traditional measures of error, such as mean-squared error, will be replaced by far more meaningful perceptual quality measures. Of course, such a replacement requires demonstration of tangible success – a task that needs concentrated involvement of researchers in the field of quality assessment.

The future is still bright for visual quality measures, especially in areas that have not been explored much before – such as interactions between visual quality and visual tasks. It is but natural to posit separation of measurements of quality impairment (from capture, processing, compression, transmission, post-processing) from scene-dependent factors, so their effects on detection, recognition, or other tasks can be identified and mitigated. This is particularly true in high-distortion environments, such as the increasingly crowded wireless/mobile environment. A principled, ground-up approach is needed whereby the effects of blindly measured video quality degradations on visual tasks can be established. This is of particular importance in forthcoming wireless vision applications where severe distortions occur, and in security applications such as human tracking, which have taken on an increasingly important role in modern-day systems.

References

[1] Ullman, S. and Poggio, T., *Vision: A computational investigation into the human representation and process of visual information*. MIT Press, Cambridge, MA, 2010.

[2] Grady, D., 'The vision thing: Mainly in the brain.' *Discover*, **14**(6), 1993, 56–66.

[3] Bovik, A.C., *Handbook of Image and Video Process*. Academic Press, New York, 2010.

[4] Heeger, D.J. and Simoncelli, E.P., 'Model of visual motion sensing.' *Spatial Vision in Humans and Robots*, **19**, 1993, 367–392.

[5] Jayant, N., Johnston, J., and Safranek, R., 'Sig. compression based on models of human perception.' *Proceedings of the IEEE*, **81**(10), 1993, 1385–1422.

[6] Park, J., Seshadrinathan, K., Lee, S., and Bovik, A.C., 'VQ pooling: Video quality pooling adaptive to perceptual distortion severity.' *IEEE Transactions on Image Processing*, **22**(2), 2013, 610–620.

[7] Seshadrinathan, K. and Bovik, A.C., 'Automatic prediction of perceptual quality of multimedia sig.sa survey.' *Multimedia Tools and Applications*, **51**(1), 2011, 163–186.

[8] Chandler, D.M. and Hemami, S.S., 'VSNR: A wavelet-based visual Sig.-to-noise ratio for natural images.' *IEEE Transactions on Image Processing*, **16**(9), 2007, 2284–2298.

[9] Daly, S., 'The visible differences predictor: An algorithm for the assessment of image fidelity.' In Watson, A.B. (ed.), *Digital Images and Human Vision.* MIT Press, Cambridge, MA, 1993, pp. 179–206.

[10] Lubin, J., 'The use of psychophysical data and models in the analysis of display system performance.' In Watson, A.B. (ed.), *Digital Images and Human Vision.* MIT Press, Cambridge, MA, 1993, pp. 163–178.

[11] Teo, P.C. and Heeger, D.J., 'Perceptual image distortion.' IEEE International Conference on Image Processing, Vol. 2, 1994, pp. 982–986.

[12] Egiazarian, K., Astola, J., Ponomarenko, N., Lukin, V., Battisti, F., and Carli, M., 'New full-reference quality metrics based on HVS.' International Workshop on Video Processing and Quality Metrics, 2006.

[13] Wang, Z. and Bovik, A.C., 'A universal image quality index.' *IEEE Signal Processing Letters*, **9**(3), 2002, 81–84.

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

[15] Wang, Z., Simoncelli, E.P., and Bovik, A.C., 'Multiscale structural similarity for image quality assessment.' Asilomar Conference on Signals, Systems and Computers, Vol. 2, 2003, pp. 1398–1402.

[16] Sheikh, H.R., Bovik, A.C., and De Veciana, G., 'An information fidelity criterion for image quality assessment using natural scene statistics.' *IEEE Transactions on Image Processing*, **14**(12), 2005, 2117–2128.

[17] Sheikh, H.R. and Bovik, A.C., 'Image information and visual quality.' *IEEE Transactions on Image Processing*, **15**(2), 2006, 430–444.

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

[19] Li, Q. and Wang, Z., 'Reduced-reference image quality assessment using divisive normalization-based image representation.' *IEEE Journal of Selected Topics in Signal Processing*, **3**(2), 2009, 202–211.

[20] Chono, K., Lin, Y.C., Varodayan, D., Miyamoto, Y., and Girod, B., 'Reduced-reference image quality assessment using distributed source coding.' IEEE International Conference on Multimedia and Expo, 2008, pp. 609–612.

[21] Engelke, U., Kusuma, M., Zepernick, H.J., and Caldera, M., 'Reduced-reference metric design for objective perceptual quality assessment in wireless imaging.' *Signal Processing: Image Communication*, **24**(7), 2009, 525–547.

[22] Barland, R. and Saadane, A., 'Reference free quality metric using a region-based attention model for JPEG-2000 compressed images.' *Proceedings of SPIE*, 6059, 2006, 605 905-1–605 905-10.

[23] Chen, J., Zhang, Y., Liang, L., Ma, S., Wang, R., and Gao, W., 'A no-reference blocking artifacts metric using selective gradient and plainness measures.' *Pacific Rim Conference on Multimedia: Advances in Multimedia Information Processing*. Springer-Verlag, Berlin, 2008, pp. 894–897.

[24] Suthaharan, S., 'No-reference visually significant blocking artifact metric for natural scene images.' *Journal of Signal Processing*, **89**(8), 2009, 1647–1652.

[25] Sheikh, H.R., Bovik, A.C., and Cormack, L.K., 'No-reference quality assessment using natural scene statistics: JPEG2000.' *IEEE Transactions on Image Processing*, **14**(11), 2005, 1918–1927.

[26] Varadarajan, S. and Karam, L.J., 'An improved perception-based no-reference objective image sharpness metric using iterative edge refinement.' IEEE International Conference on Image Processing, 2008, pp. 401–404.

[27] Ferzli, R. and Karam, L.J., 'A no-reference objective image sharpness metric based on the notion of just noticeable blur (JNB).' *IEEE Transactions on Image Processing*, **18**(4), 2009, 717–728.

[28] Narvekar, N.D. and Karam, L.J., 'A no-reference perceptual image sharpness metric based on a cumulative probability of blur detection.' IEEE International Workshop on Quality of Multimedia Expo, 2009, pp. 87–91.

[29] Sadaka, N.G., Karam, L.J., Ferzli, R., and Abousleman, G.P., 'A no-reference perceptual image sharpness metric based on saliency-weighted foveal pooling.' IEEE International Conference on Image Processing, 2008, pp. 369–372.

[30] Li, X., 'Blind image quality assessment.' IEEE International Conference on Image Processing, Vol. 1, 2002, pp. 449–452.

[31] Ye, P. and Doermann, D., 'No-reference image quality assessment using visual codebook.' IEEE International Conference on Image Processing, 2011.

[32] Ye, P. and Doermann, D., 'No-reference image quality assessment using visual codebooks.' *IEEE Transactions on Image Processing*, **21**(7), 2012, 3129–3138.

[33] Gabarda, S. and Cristóbal, G., 'Blind image quality assessment through anisotropy.' *Journal of the Optical Society of America*, **24**(12), 2007, 42—51

[34] Tang, H., Joshi, N., and Kapoor, A., 'Learning a blind measure of perceptual image quality.' IEEE International Conference on Computer Vision Pattern Recognition, 2011.

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

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

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

[38] Mittal, A., Moorthy, A.K., and Bovik, A.C., 'Making image quality assessment robust.' Asilomar Conference on Signals, Systems and Computers, 2012, pp. 1718–1722.

[39] Mittal, A., Moorthy, A.K., and Bovik, A.C., 'Blind/referenceless image spatial quality evaluator.' Asilomar Conference on Signals, Systems and Computers, 2011, pp. 723–727.

[40] Mittal, A., Muralidhar, G.S., Ghosh, J., and Bovik, A.C., 'Blind image quality assessment without human training using latent quality factors.' *IEEE Signal Processing Letters*, **19**, 2011, 75–78.

[41] Mittal, A., Soundararajan, R., Muralidhar, G., Ghosh, J., and Bovik, A.C., 'Un-naturalness modeling of image distortions.' Vision Sciences Society, 2012.

[42] Xue, W., Zhang, L., and Mou, X., 'Learning without human scores for blind image quality assessment.' IEEE International Conference on Computer Vision Pattern Recognition, 2011.

[43] Mittal, A., Soundararajan, R., Muralidhar, G.S., Bovik, A.C., and Ghosh, J., 'Blind image quality assessment without training on human opinion scores.' In *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 2013, pp. 86510T–86510T.

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

[45] Mittal, A., Soundararajan, R., and Bovik, A.C., 'Prediction of image naturalness and quality.' *Journal of Vision*, **13**(9), 2013, 1056.

[46] Soundararajan, R. and Bovik, A.C., 'Survey of information theory in visual quality assessment.' *Signal, Image and Video Processing*, **7**(3), 2013, 391–401.

[47] Moorthy, A.K. and Bovik, A.C., 'Visual quality assessment algorithms: What does the future hold?' *Multimedia Tools and Applications*, **51**(2), 2011, 675–696.

[48] Meesters, L. and Martens, J.B., 'A single-ended blockiness measure for JPEG-coded images.' *Journal of Signal Processing*, **82**(3), 2002, 369–387.

[49] Wang, Z., Sheikh, H.R., and Bovik, A.C., 'No-reference perceptual quality assessment of JPEG compressed images.' International Conference on Image Processing, 1, 2002, 477–480.

[50] Moorthy, A.K., Mittal, A., and Bovik, A.C., 'Perceptually optimized blind repair of natural images.' *Signal Processing: Image Communication*, **28**, 2013, 1478–1493.

[51] Wang, Z. and Bovik, A.C., 'Reduced- and no-reference image quality assessment.' *IEEE Signal Processing Magazine*, **28**(6), 2011, 29–40.

[52] Shen, J., Li, Q., and Erlebacher, G., 'Hybrid no-reference natural image quality assessment of noisy, blurry, JPEG2000, and JPEG images.' *IEEE Transactions on Image Processing*, **20**(8), 2011, 2089–2098.

[53] Ruderman, D.L. and Bialek, W., 'Statistics of natural images: Scaling in the woods.' *Physical Review Letters*, **73**, 1994, 814–817.

[54] Soundararajan, R. and Bovik, A.C., 'RRED indices: Reduced reference entropic differencing for image quality assessment.' *IEEE Transactions on Image Processing*, **21**(2), 2011, 517–526.

[55] Ruderman, D.L., 'The statistics of natural images.' *Network Computation in Neural Systems*, **5**(4), 1994, 517–548.

[56] Chen, X. and Cai, D., 'Large scale spectral clustering with landmark-based representation.' Advances in Artificial Intelligence, 2011.

[57] Tan, K.T. and Ghanbari, M., 'Blockiness detection for mpeg2-coded video.' *IEEE Signal Processing Letters*, **7**(8), 2000, 213–215.

[58] Vlachos, T., 'Detection of blocking artifacts in compressed video.' *Electronics Letters*, **36**(13), 2000, 1106–1108.

[59] Suthaharan, S., 'Perceptual quality metric for digital video coding.' *Electronics Letters*, **39**(5), 2003, 431–433.

[60] Muijs, R. and Kirenko, I., 'A no-reference blocking artifact measure for adaptive video processing.' European Signal Processing Conference, 2005.

[61] Lu, J., 'Image analysis for video artifact estimation and measurement.' Photonics West-Electronic Imaging, 2001, pp. 166–174.

[62] Huynh-Thu, Q. and Ghanbari, M., 'Impact of jitter and jerkiness on perceived video quality.' International Workshop on Video Processing and Quality Metrics for Consumer Electronics, 2006.

[63] Huynh-Thu, Q. and Ghanbari, M., 'Temporal aspect of perceived quality in mobile video broadcasting.' *IEEE Transactions on Broadcasting*, **54**(3), 2008, 641–651.

[64] Mei, T., Hua, X.S., Zhu, C.Z., Zhou, H.Q., and Li, S., 'Home video visual quality assessment with spatiotemporal factors.' *IEEE Transactions on Circuits and Systems for Video Technology*, **17**(6), 2007, 699–706.

[65] Yang, K., Guest, C.C., El-Maleh, K., and Das, P.K., 'Perceptual temporal quality metric for compressed video.' *IEEE Transactions on Multimedia*, **9**(7), 2007, 1528–1535.

[66] Ou, Y.F., Ma, Z., Liu, T., and 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), 2011, 286–298.

[67] Caviedes, J.E. and Oberti, F., 'No-reference quality metric for degraded and enhanced video.' Visual Communication and Image Processing, 2003, pp. 621–632.

[68] Babu, R.V., Bopardikar, A.S., Perkis, A., and Hillestad, O.I., 'No-reference metrics for video streaming applications.' International Workshop on Packet Video, 2004.

[69] Farias, M.C.Q. and Mitra, S.K., 'No-reference video quality metric based on artifact measurements.' IEEE International Conference on Image Processing, Vol. 3, 2005, pp. III–141.

[70] Massidda, F., Giusto, D.D., and Perra, C., 'No reference video quality estimation based on human visual system for 2.5/3G devices.' Electronic Imaging, 2005, pp. 168–179.

[71] Dosselmann, R. and Yang, X.D., 'A prototype no-reference video quality system.' Fourth Canadian Conference on Computer and Robot Vision, 2007, pp. 411–417.

[72] Yang, F., Wan, S., Chang, Y., and Wu, H., 'A novel objective no-reference metric for digital video quality assessment.' *IEEE Signal Processing Letters*, **12**(10), 2005, 685–688.

[73] Pastrana-Vidal, R.R. and Gicquel, J.C., 'Automatic quality assessment of video fluidity impairments using a no-reference metric.' International Workshop on Video Processing and Quality Metrics for Consumer Electronics, 2006.

[74] Yamada, T., Miyamoto, Y., and Serizawa, M., 'No-reference video quality estimation based on error-concealment effectiveness.' Packet Video, 2007, pp. 288–293.

[75] Naccari, M., Tagliasacchi, M., Pereira, F., and Tubaro, S., 'No-reference modeling of the channel induced distortion at the decoder for H. 264/AVC video coding.' IEEE International Conference on Image Processing, 2008, pp. 2324–2327.

[76] Ong, E.P., Wu, S., Loke, M.H., *et al.*, 'Video quality monitoring of streamed videos.' IEEE International Conference on Acoustics, Speech and Signal Processing, 2009, pp. 1153–1156.

[77] Keimel, C., Oelbaum, T., and Diepold, K., 'No-reference video quality evaluation for high-definition video.' IEEE International Conference on Acoustics, Speech and Signal Processing, 2009, pp. 1145–1148.

[78] Hemami, S.S. and Reibman, A.R., 'No-reference image and video quality estimation: Applications and human-motivated design.' *Signal Processing: Image Communication*, **25**(7), 2010, 469–481.

[79] Saad, M.A. and Bovik, A.C., 'Blind quality assessment of videos using a model of natural scene statistics and motion coherency.' Asilomar Conference on Signals, Systems and Computers, 2012.

[80] Seshadrinathan, K., Soundararajan, R., Bovik, A., and Cormack, L., 'Study of subjective and objective quality assessment of video.' *IEEE Transactions on Image Processing*, **19**(6), 2010, 1427–1441.

[81] Jayaraman, D., Mittal, A., Moorthy, A.K., and Bovik, A.C., 'Objective image quality assessment of multiply distorted images.' Asilomar Conference on Signals, Systems and Computers, 2012, pp. 1693–1697.

[82] 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), 2006, 3440–3451.

[83] Le Callet, P. and Autrusseau, F., 'Subjective quality assessment IRCCYN/IVC database.' http://www.irccyn.ec-nantes.fr/ivcdb/.

[84] Ponomarenko, N., Lukin, V., Zelensky, A., Egiazarian, K., Carli, M., and Battisti, F., 'TID2008 – A database for evaluation of full-reference visual quality assessment metrics.' *Advances in Modern Radioelectronics*, **10**, 2009, 30–45.

[85] Larson, E.C. and Chandler, D.M., 'Most apparent distortion: Full-reference image quality assessment and the role of strategy.' *Journal of Electronic Imaging*, **19**(1), 2010, 011 006-1–011 006-21.

[86] Ciancio, A.G., da Costa, A.L.N.T., da Silva, E.A.B., Said, A., Samadani, R., and Obrador, P., 'No-reference blur assessment of digital pictures based on multifeature classifiers.' *IEEE Transactions on Image Processing*, **20**(1), 2011, 64–75.

[87] De Simone, F., Naccari, M., Tagliasacchi, M., Dufaux, F., Tubaro, S., and Ebrahimi, T., 'Subjective assessment of h.264/avc video sequences transmitted over a noisy channel.' Quality Multimedia Expo, 2009, pp. 204–209.

[88] De Simone, F., Tagliasacchi, M., Naccari, M., Tubaro, S., and Ebrahimi, T., 'A h.264/avc video database for the evaluation of quality metrics.' IEEE International Conference on Acoustics, Speech and Signal Processing, 2010, pp. 2430–2433.

[89] Péchard, S., Pépion, R., and Le Callet, P., 'Suitable methodology in subjective video quality assessment: A resolution dependent paradigm.' International Workshop on Image Media Quality and its Applications, 2008.

[90] Rec. ITU-R BT.500-11. http://www.dii.unisi.it/~menegaz/DoctoralSchool2004/papers/ITU-R_BT.500-11.pdf.

[91] Lee, J., De Simone, F., Ramzan, N., *et al.*, 'Subjective evaluation of scalable video coding for content distribution.' International Conference on Multimedia, 2010, pp. 65–72.

[92] Lee, J., De Simone, F., and Ebrahimi, T., 'Subjective quality evaluation via paired comparison: Application to scalable video coding.' *IEEE Transactions on Multimedia*, **13**(5), 2011, 882–893.

[93] Moorthy, A.K., Choi, L., Bovik, A.C., and De Veciana, G., 'Video quality assessment on mobile devices: Subjective, behavioral and objective studies.' *IEEE Selected Topics in Signal Processing*, **6**(6), 2012, 652–671.

[94] Winkler, S., 'Analysis of public image and video databases for quality assessment.' *IEEE Selected Topics in Signal Processing*, **6**(6), 2012, 616–625.

[95] Winkler, S., 'List of image and video quality databases.' http://stefan.winkler.net/resources.html.

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

[97] Buades, A., Coll, B., and Morel, J.M., 'A review of image denoising algorithms, with a new one.' *Multiscale Modeling & Simulation*, **4**(2), 2005, 490–530.

[98] Mittal, A., Moorthy, A.K., and Bovik, A.C., 'Automatic parameter prediction for image denoising algorithms using perceptual quality features.' IS&T/SPIE Electronic Imaging. International Society for Optics and Photonics, 2012, pp. 82910G–82910G.

[99] Hansen, P.C., 'Analysis of discrete ill-posed problems by means of the l-curve.' *SIAM Review*, **34**(4), 1992, 561–580.

[100] Hansen, P.C. and Oleary, D.P., 'The use of the l-curve in the regularization of discrete hx-posed problems.' *SIAM Journal on Scientific Computing*, **14**(6), 1993, 1487–1503.

[101] Regińska, T., 'A regularization parameter in discrete ill-posed problems.' *SIAM Journal on Scientific Computing*, **17**(3), 1996, 740–749.

[102] Oraintara, S., Karl, W., Castanon, D., and Nguyen, T., 'A method for choosing the regularization parameter in generalized Tikhonov regularized linear inverse problems.' International Conference on Image Processing, Vol. 1, 2000, pp. 93–96.

[103] Karl, W.C., 'Regularization in image restoration and reconstruction.' In Bovik, A.C. (ed.), *Handbook of Image and Video Processing*. Academic Press, New York, 2000, pp. 141–160.

[104] Craven, P. and Wahba, G., 'Smoothing noisy data with spline functions.' *Numerische Mathematik*, **31**(4), 1978, 377–403.

[105] Golub, G.H., Heath, M., and Wahba, G., 'Generalized cross-validation as a method for choosing a good ridge parameter.' *Technometrics*, **21**(2), 1979, 215–223.

[106] Nychka, D., 'Bayesian confidence intervals for smoothing splines.' *Journal of the American Statistical Association*, **83**, 1988, 1134–1143.

[107] Thompson, A.M., Brown, J.C., Kay, J.W., and Titterington, D.M., 'A study of methods of choosing the smoothing parameter in image restoration by regularization.' *Transactions on Pattern Analysis and Machine Intelligence*, **13**(4), 1991, 326–339.

[108] Galatsanos, N.P. and Katsaggelos, A.K., 'Methods for choosing the regularization parameter and estimating the noise variance in image restoration and their relation.' *Transactions on Image Processing*, **1**(3), 1992, 322–336.

[109] Ramani, S., Blu, T., and Unser, M., 'Monte-Carlo sure: A black-box optimization of regularization parameters for general denoising algorithms.' *Transactions on Image Processing*, **17**(9), 2008, 1540–1554.

[110] Blu, T. and Luisier, F., 'The sure-let approach to image denoising.' *Transactions on Image Processing*, **16**(11), 2007, 2778–2786.

[111] Luisier, F., Blu, T., and Unser, M., 'A new sure approach to image denoising: Interscale orthonormal wavelet thresholding.' *Transactions on Image Processing*, **16**(3), 2007, 593–606.

[112] Zhang, X. and Desai, M., 'Adaptive denoising based on sure risk.' *Signal Processing Letters*, **5**(10), 1998, 265–267.

[113] Donoho, D. and Johnstone, I., 'Adapting to unknown smoothness via wavelet shrinkage.' *Journal of the American Statistical Association*, **90**, 1995, 1200–1224.

[114] Channappayya, S.S., Bovik, A.C., and Heath, R.W., 'A linear estimator optimized for the structural similarity index and its application to image denoising.' International Conference on Image Processing, 2006, pp. 2637–2640.

[115] Zhu, X. and Milanfar, P., 'Automatic parameter selection for denoising algorithms using a no-reference measure of image content.' *IEEE Transactions on Image Processing*, **19**(12), 2010, 3116–3132.

[116] Dabov, K., Foi, A., Katkovnik, V., and Egiazarian, K., 'Image denoising by sparse 3-d transform-domain collaborative filtering.' *IEEE Transactions on Image Processing*, **16**(8), 2007, 2080–2095.

[117] Reinhard, E., Heidrich, W., Debevec, P., Pattanaik, S., Ward, G., and Myszkowski, K., *High Dynamic Range Imaging: Acquisition, display, and image-based lighting*. Morgan Kaufmann, Burlington, MA, 2010.

[118] Yeganeh, H. and Wang, Z., 'Objective quality assessment of tone mapped images.' *IEEE Transactions on Image Processing*, **22**(2), 2013, 657–667.

[119] UCID – Uncompressed Colour Image Database. http://www.staff.lboro.ac.uk/cogs/datasets/UCID/ucid.html, 2004.

[120] Computer vision test images. http://www.2.cs.cmu.edu/afs/cs/project/cil/www/v-images.html, 2005.

[121] 'Image obsessed.' *National Geographic*, 221, 2012, p. 35.

[122] Johnson, D., Chen, B., Chen, N., Pan, J., and Huynh, J., ‘Photo quality helper.’ https://play.google.com/store/apps/details?id=net.dorianj.rockout &feature=search result#?t=W251bGwsMSwxLDEsIm5ldC5kb3JpYW5qLnJvY2tvdXQiXQ.

[123] Li, C. and Chen, T., ‘Aesthetic visual quality assessment of paintings.’ *IEEE Journal of Selected Topics in Signal Processing*, **3**(2), 2009, 236–252.

[124] Datta, R., Joshi, D., Li, J., and Wang, J., ‘Studying aesthetics in photographic images using a computational approach.’ *Lecture Notes in Computer Science*, **3953**, 2006, 288.

[125] Datta, R., Li, J., and Wang, J., ‘Algorithmic inferencing of aesthetics and emotion in natural images: An exposition.’ IEEE International Conference on Image Processing, 2008, pp. 105–108.

[126] Ke, Y., Tang, X., and Jing, F., ‘The design of high-level features for photo quality assessment.’ IEEE Conference on Computer Vision Pattern Recognition, Vol. 1, 2006.

[127] Luo, Y. and Tang, X., ‘Photo and video quality evaluation: Focusing on the subject.’ European Conference on Computer Vision, 2008, pp. 386–399.

[128] Moorthy, A.K., Obrador, P., Oliver, N., and Bovik, A.C., ‘Towards computational models of visual aesthetic appeal of consumer videos.’ European Conference on Computer Vision, 2010.

[129] Seshadrinathan, K. and Bovik, A.C., ‘Temporal hysteresis model of time varying subjective video quality.’ IEEE International Conference on Acoustics, Speech and Signal Processing, 2011, pp. 1153–1156.

[130] Park, J., Seshadrinathan, K., Lee, S., and Bovik, A., ‘Spatio-temporal quality pooling accounting for transient severe impairments and egomotion.’ IEEE International Conference on Image Processing, 2010.

Acronyms

BLIINDS	BLInd Image Notator using DCT Statistics
BRISQUE	Blind/Referenceless Image Spatial Quality Evaluator
CSIQ	Categorical Subjective Image Quality
DA	Distortion Aware
DIIVINE	Distortion Identification-based Image Verity and INtegrity Evaluation
DMOS	Differential Mean Opinion Score
DU	Distortion Unaware
FR	Full Reference
HDR	High Dynamic Range
IQA	Image Quality Assessment
LCC	Linear Correlation Coefficient
MOS	Mean Opinion Score
MVG	Multi-Variate Gaussian
NIQE	Naturalness Image Quality Evaluator
NR	No Reference
NSS	Natural Scene Statistics
OA	Opinion Aware
OU	Opinion Unaware
PSNR	Peak-Signal-to-Noise Ratio
QA	Quality Assessment
RR	Reduced Reference
SROCC	Spearman Rank-Ordered Correlation Coefficient
SSIM	Structural SIMilarity Index
SVR	Support Vector Regression
TMQI	Tone-Mapped Quality Index
VQA	Video Quality Assessment

6

QoE Subjective and Objective Evaluation Methodologies

Hong Ren Wu
RMIT, Australia

The discussions in previous chapters have highlighted an increasing emphasis on Quality of Experience (QoE) [1] compared with Quality of Service (QoS) [2] in audio-visual communication, broadcasting and entertainment applications, which signals a transition from technology-driven services to user-centric (or perceived) quality-assured services [3]. QoE as defined by ITU SG 12[1] is application or service specific and influenced by user expectations and context [1], and therefore necessitates assessments of perceived service quality and/or utility (or usefulness) of the service [4]. Subjective assessment and evaluation of the service are imperative to establish the ground truth for objective assessment and measures which aid in the design/optimization of devices, products, systems, and services in either the online or offline mode [1]. Assessment or prediction of QoE in multimedia services will have to take account of at least three major factors, including audio signal quality perception, visual signal quality perception, and interaction or integration of the perceived audio and visual signal quality [5–7], considering coding, transmission, and application/service conditions [3, 6, 8, 9]. This chapter focuses on the issues underpinning the theoretical frameworks/models and methodologies for QoE subjective and objective evaluation of visual signal communication services.

Subjective picture quality assessment methods for television and multimedia applications have been standardized [10, 11]. Issues relevant to human visual perception and quality scoring or rating are discussed in this chapter, while readers are referred to the standards documents and/or other monographs regarding specific details of the aforementioned standards [10–12].

[1] International Telecommunication Union, Telecommunication Standardization Sector, Study Group 12.

Multimedia Quality of Experience (QoE): Current Status and Future Requirements, First Edition.
Edited by Chang Wen Chen, Periklis Chatzimisios, Tasos Dagiuklas and Luigi Atzori.

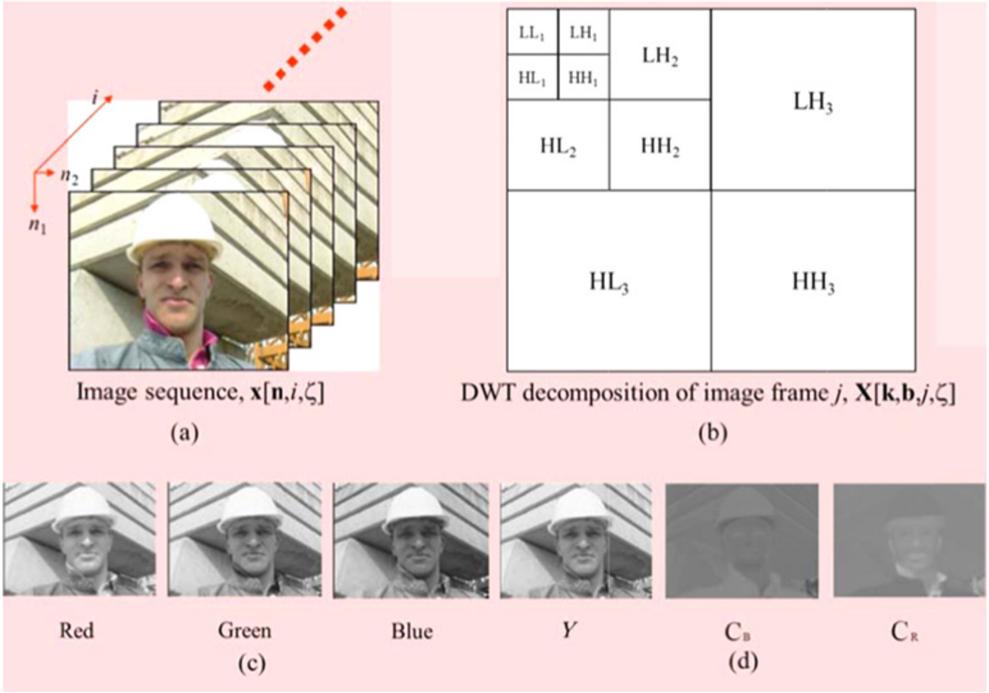


Figure 6.1 A color image sequence example: (a) image sequence; (b) three-level DWT decomposition sub-band designations; (c) three-color channel signals in *RGB* space; (d) three-color channel signals in $YC_{\text{B}}C_{\text{R}}$ space

The Human Visual System (HVS) can be modeled in either pixel domain or transform/sub-band decomposition domain, and the same can be said about picture quality metric formulations. Given a color image sequence or video as shown in Figure 6.1(a), $\mathbf{x}[\mathbf{n}, i, \zeta]$, N_1 pixels high and N_2 pixels wide, where $\mathbf{n} = [n_1, n_2]$ for $0 \le n_1 \le N_1 - 1$ and $0 \le n_2 \le N_2 - 1$ with I frames for $0 \le i \le I - 1$ in tricolor space $\Xi = \left\{Y, C_{\text{B}}, C_{\text{R}}\right\}^2$ (or $\Xi = \{R, G, B\}^3$) for $\zeta \in \{1, 2, 3\}$ corresponding to, for example, Y, C_{B}, C_{R} (or R, G, B) channels (cf. Figure 6.1(d) or (c)) [13, 14], respectively, its transform or decomposition is represented by $\mathbf{X} = [\mathbf{k}, \mathbf{b}, j, \zeta]$, as shown, for example, in Figure 6.1(b) for a Discrete Wavelet Transform (DWT) decomposition, where $\mathbf{k} = [k_1, k_2]$ defines the position (row and column indices) of a coefficient in a block of a frequency band $\mathbf{b}$ of slice j in the decomposition domain. For an s-level DWT decomposition, $\mathbf{b} = [s, \theta]$, where $\theta \in \left\{\theta_0|\text{LL band}, \theta_1|\text{LH band}, \theta_2|\text{HL band}, \theta_3|\text{HH band}\right\}$ and $s = 3$ per frame, as shown in Figure 6.1(b). It is noted that, as shown in Figure 6.1(c) and (d), three component channels in the $YC_{\text{B}}C_{\text{R}}$ color space are better decorrelated than those in the *RGB* color space, facilitating picture compression. Other color spaces, such as opponent

[2] Y, C_{B}, C_{R} represent digital luminance and color-difference signals, respectively.

[3] R, G, B represent digital red, green, blue signals, respectively.

color spaces, which are considered perceptually uniform[4] [15], have often been used in color picture quality assessment [16–18].

6.1 Human Visual Perception and QoE Assessment

Quantitative assessment of visual signal[5] quality as perceived by the HVS inevitably involves human observers participating in subjective tests which elicit quality ratings using one scale or another [4, 10–12]. Quantification of subjective test results is usually based on the Mean Opinion Score (MOS),[6] which indicates the average rating value qualified by a measure of deviation (e.g., standard deviation, variance, minimum and maximum values, or a confidence interval), acknowledging the subjectivity and statistical nature of the assessment [10–12]. This section discusses a number of issues associated with human visual perception, which affect subjective rating outcomes and, thereafter, their reliability and relevance when used as the ground truth for objective or computational metric designs. Models or approaches to computational QoE metric designs to date [19–27] are analyzed to appreciate their theoretical grounding and strength in terms of prediction accuracy, consistency, and computational complexity, as well as their limitations, with respect to QoE assessment and prediction for picture codec design, network planning and performance optimization, and service performance assessment.

Low-level human vision has been characterized by spatial, temporal and color vision, and visual attention (or foveation). There are well-formulated HVS models which have been successfully used in Image (or Video) Quality Assessment (IQA or VQA) and evaluation [19–21] and perception-based picture coder designs [3]. It is noted that the majority of these models have their parameters derived from a threshold vision test to estimate or predict the Just-Noticeable Difference (JND) [28], with a few exceptions [29,30]. When these threshold vision models are extended to supra-threshold experiments where most of the visual communications, broadcast, and entertainment applications to date apply [3], the selection of the model parameters relies heavily on a regression process to achieve the best fit to subjective test data [18]. Relevancy, accuracy, and reliability of subjective test data therefore directly affect the performance of objective QoE metrics [4, 7, 18, 25]. Four issues have emerged over the years of acquiring ground-truth subjective test data in terms of perceived picture quality and QoE, and are worth noting.

First and foremost, the picture quality perceived and thereafter the ratings given by human observers are affected by what they have seen or experienced prior to a specific subjective test. Contextual effects[7] (considered as short-term or limited test sample

[4] In perceptually uniform color space, equal distances in the, e.g., CIELab, color space represent equal perceived differences in appearance [17].

[5] Visual signals refer to digital pictures of natural scenes, including still images, video, image sequences, or motion pictures [3].

[6] Differential (or difference) mean opinion score, which is defined as “source – processed,” is also used.

[7] Contextual effects are fluctuations in the subjective rating of video or image sequences based on the impairment present in the preceding video sequences – e.g., a sequence with moderate impairment that follows a set of sequences with weak impairment may be judged lower in quality than if it followed sequences with strong impairment.

pool[8] effects) have been reported using standardized subjective test methods [31]. Affordability notwithstanding, users' expectations are understandably influenced by their benchmark experience or point of reference in what constitutes the "best" picture quality they have seen or experienced. This benchmark experience (long-term or global sample pool effects) will derive subsequent ratings on a given quality scale. For an observer who had never viewed or experienced, for example, an uncompressed YC_BC_R 4:4:4 component color video of Standard Definition (SD) [13] or full High Definition (HD) [14] on a broadcast video monitor designed for critical picture evaluation,[9] it would be highly uncertain what response one could hope to elicit from the observer when asked if any chroma distortion was present in the 4:2:2 or 4:2:0 component video in the subjective test. While chroma sub-sampling has been widely used in video products to take advantage of HVS' insensitivity to chroma component signals as an initial step to image data compression, the chroma distortions so caused are not always negligible as commonly believed, especially using quality (e.g., broadcast or professional-grade) video monitors. Figure 6.2 uses contrast-enhanced[10] difference images between the uncompressed *Barbara* test image in component YC_BC_R 4:4:4 format [13, 14] and those chroma sub-sampled images in component 4:2:2, 4:2:0, and 4:11 formats, respectively, to illustrate chromatic distortions. The same thing may be said about responses from observers used in an image or video quality subjective evaluation test, to whom various picture coding artifacts or distortions [32–34] are unknown. Issues associated with the consistency and reliability of subjective test results as reported or reviewed (cf., e.g., [7, 35]) aside, subjective picture quality assessment data collected from observers with limited benchmark experience is deemed not to have a reliable/desirable anchor (or reference) point, making the analysis results difficult to interpret or inconclusive [7], if not questionable, and does not inspire confidence in the application to objective quality metric design and optimization. In other words, using observers with minimum knowledge, experience, or expectations in subjective picture quality evaluation tests generates data with varying or no reference point, often then lowering expectations of what is considered as "Excellent" picture quality, and possibly running a real risk of racing to the bottom in the practice of quality assessment, control, and assurance for visual signal communications, broadcast, and entertainment applications.

Second, it has long been acknowledged that human perception and judgment in a psychophysical measurement task usually perform better in comparison tasks than casting an absolute rating.[11] Nevertheless, an Absolute Category Rating (ACR) or absolute rating scale has been used widely in subjective picture quality evaluations [7, 10–12]. To address the issue regarding fluctuations in subjective test data using absolute rating schemes due to the aforementioned contextual effects and the varying experience and expectations of observers, a Multiple Reference Impairment Scale (MRIS) subjective test method for digital video was

[8] The limited sample pool refers here to the set of test images or video sequences used in a particular subjective test run in comparison with what a person may have seen and experienced in his/her lifetime.

[9] An example is Sony's BVM-L230 23" Trimaster LCD Master (Reference) Monitor.

[10] Contrast enhancement was used to facilitate visualization of chromatic distortions due to chroma sub-sampling for this example in printed form or on a low-grade video or PC monitor.

[11] As the Chinese proverb goes, "不怕不识货, 就怕货比货." ("Shop around and compare, and then you will tell what is the best value for money.")

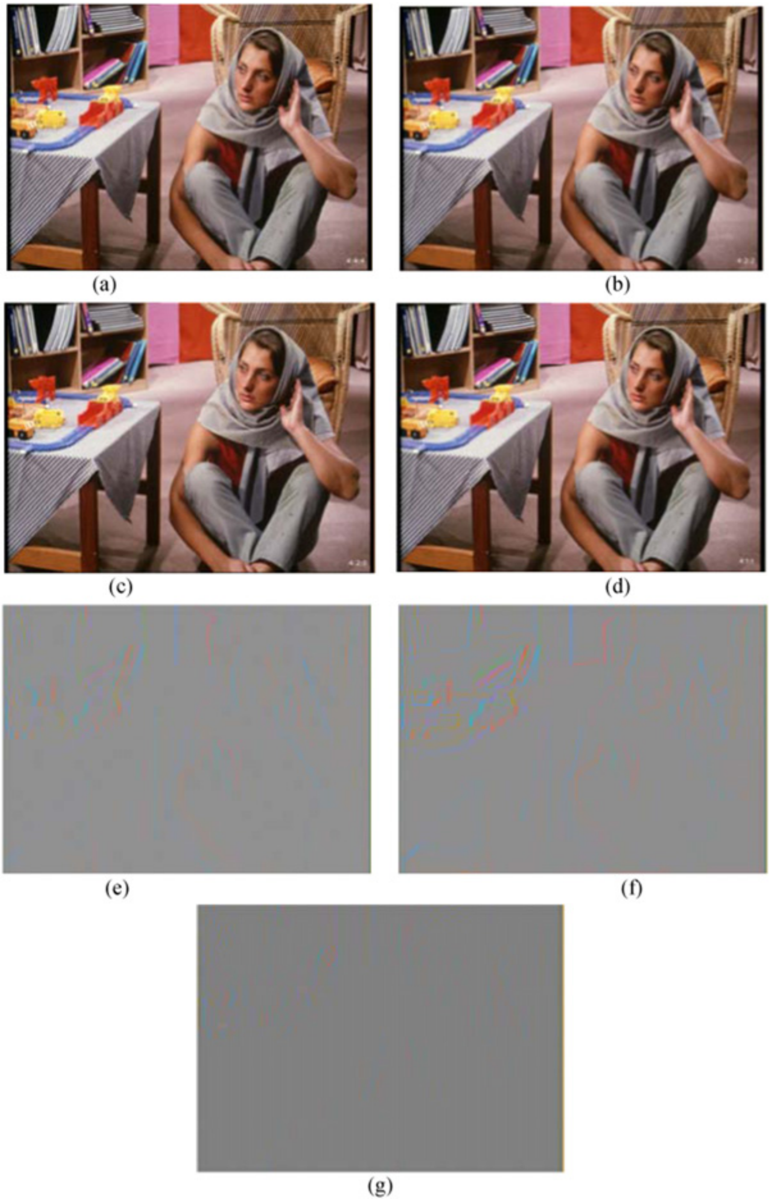


Figure 6.2 Examples of chroma distortions in *Barbara* test image due to chroma sub-sampling: (a) the uncompressed image in component YC_BC_R 4:4:4 format; (b) the image in component 4:2:2 format; (c) the image in component 4:2:0 format; (d) the image in component 4:1:1 format; (e) contrast-enhanced difference image between (a) and (b); (f) contrast-enhanced difference image between (a) and (c); (g) contrast-enhanced difference image between (a) and (d). Contrast enhancement was applied to difference images with a bias of 128 shown in (e)–(g) for better visualization in PDF format or using photo-quality printing

reported in [36], where a five-reference impairment scale (R5 to R1) was used with five reference pictures including the original, $\mathbf{x}_\mathrm{o}$, as uncorrupted picture reference, $\mathbf{x}_{\mathrm{R}5}$ (R5), reference distorted pictures defined, respectively, as $\mathbf{x}_{\mathrm{R}4}$, $\mathbf{x}_{\mathrm{R}3}$, $\mathbf{x}_{\mathrm{R}2}$, and $\mathbf{x}_{\mathrm{R}1}$ in terms of their perceptibility of impairment corresponding to perceptible but not annoying (R4), slightly annoying (R3), annoying (R2), and very annoying (R1). The observers compared the processed picture $\mathbf{x}_\mathrm{p}$ with the original $\mathbf{x}_{\mathrm{R}5}$ to determine if the impairment was perceptible, or with $\mathbf{x}_{\mathrm{R}i}$ for $i \in \{1, 2, 3, 4\}$ when there was a perceptible distortion to rate $\mathbf{x}_\mathrm{p}$ as better, similar, or worse than $\mathbf{x}_{\mathrm{R}i}$. This approach led to a comparative rating scale based on forced choice methods, which significantly reduced the deviation in the subjective test data and alleviated the contextual effects. Ideally, following this conventional distortion-detection strategy, each of the reference distortion scales as represented by $\mathbf{x}_{\mathrm{R}i}$ will be better off corresponding to JND levels [3, 36] or Visual Distortion Units (VDUs) [29].

Third, an issue not all together disassociated with the second is the HVS's response under two distinctive picture quality assessment conditions: where artifacts and distortions are at visibility sub-threshold or around the threshold (usually found in high-quality pictures) and at supra-threshold (commonly associated with medium and low-quality pictures). HVS models based on threshold vision test and detection of the JND have been available and widely adopted in objective picture quality/distortion measures/metrics [19–27]. While picture distortion measures based on these models have been successfully employed in perceptually lossless picture coding [3, 37, 38], applications of JND models to picture processing, coding, and transmission or storage at supra-threshold levels have revealed that the supposition of linear scaling JND models is not fully supported by various experimental studies [15, 29, 39] and, therefore, further investigations are required to identify, delineate, and better model HVS characteristics/responses under supra-threshold conditions [29,30, 35, 39, 40]. It was argued in [30] that for assessment of a high-quality picture with distortions near the visibility threshold, the HVS tends to look past the picture and perform a distortion detection task, whilst for evaluation of a low-quality picture with obviously visible distortions of highly supra-threshold nature, the HVS tends to look past (or to be more forgiving toward) the distortions and look for the content of the picture. This hypothesis is consistent with the HVS behavior as revealed by contextual effects. To cover a wide range of picture quality as perceived or experienced by human observers, HVS modeling and quantitative QoE measurement may need to take account of two, instead of one, assessment strategies which the HVS seems to adopt: distortion detection, which is commonly used in threshold vision tests and gradation of degradations of image appearance, which is practiced in supra-threshold vision tests [30].

Fourth, it is becoming increasingly clear that the assessment of QoE requires more than the evaluation of picture quality alone, with a need to differentiate the measurement of the perceived resemblance of a picture at hand to the original from that of the usefulness of the picture to an intended task. It was reported in [4] that QoP (perceived Quality of Pictures) in a five-scale ACR and UoP (perceived Utility of Pictures) anchored by Recognition Threshold (RT, assigned "0")[12] and Recognition Equivalence Class (REC, assigned "100")[13] could

[12] RT is defined in [4] as a perceived utility score threshold below which an image is deemed useless.

[13] In [4], REC defines a class of images whose perceived utility score is statistically equivalent to that of a perceptually lossless image with respective to and including the reference.

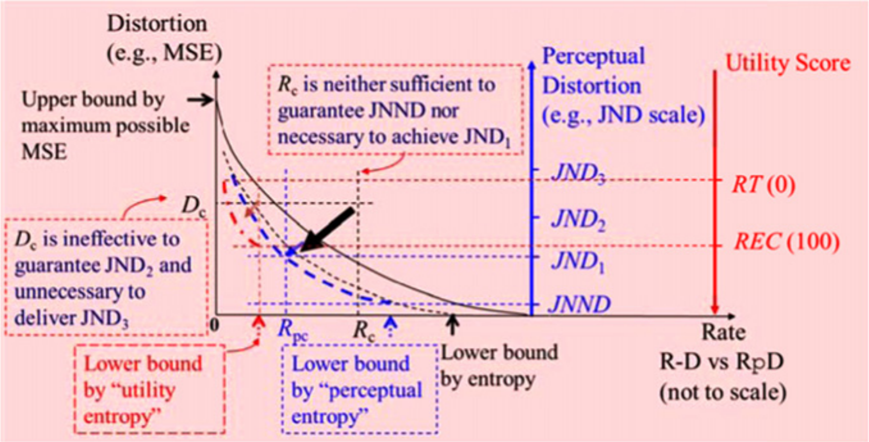

Distortion
(e.g., MSE)
Upper bound by maximum possible MSE
R_c is neither sufficient to guarantee JNND nor necessary to achieve JND_1
Perceptual Distortion (e.g., JND scale)
Utility Score
JND_3
RT (0)
D_c
JND_2
REC (100)
JND_1
D_c is ineffective to guarantee JND_2 and unnecessary to deliver JND_3
JNND
0
R_{pc}
R_c
Rate
Lower bound by "utility entropy"
Lower bound by "perceptual entropy"
Lower bound by entropy
R-D vs RpD
(not to scale)

Figure 6.3 R-D optimization considering a perceptual distortion measure [3] or a utility score [4] for QoE-regulated services compared with the MSE. *Source:* Adapted from Wu *et al.*, 2014 [91]

be approximated by a nonlinear function, and that QoP did not predict UoP well and vice versa. Detailed performance comparisons have been reported in [4] of natural scene statistical model [27] and image feature-based QoP and UoP metrics. Compared with subjective test methodology and procedures for QoP assessments which have been standardized over the years [10–12], subjective tests for UoP assessments for various specific applications may face further challenges ranging from the most critical to minimal efforts, and require participation of targeted human observers who have the necessary domain knowledge of intended applications, (e.g., radiologists and radiographers in medical diagnostic imaging) [37].

QoP and QoE assessments are not just for their own sakes, and they are linked closely to visual signal compression and transmission where Rate–Distortion (R-D) theory is applied for product, system, and service quality optimization [41–43]. From an R-D optimization perspective [44–46], it is widely understood that the use of raw mathematical distortion measures, such as the Mean Squared Error (MSE), do not guarantee visual superiority since the HVS does not compute the MSE [3, 47]. In RpD (Rate-perceptual-Distortion) optimization [48], where perceptual distortion or utility measures matter, the setting of the rate constraint, R_c, in Figure 6.3 is redundant from a perceptual distortion controled coding viewpoint. The perceptual bit-rate constraint, R_{pc}, makes more sense and delivers a picture quality comparable with JND_1. In comparison, R_c is neither sufficient to guarantee a distortion level at JNND (Just-Not-Noticeable Difference) nor necessary to achieve, for example, JND_1 in Figure 6.3. By the same token, D_c is ineffective at 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 bit rate required for information lossless picture coding [49,50], the perceptual entropy [48] sets the minimum bit rate required for perceptually lossless picture coding [37,38]. Similarly, in UoP-regulated picture coding in terms of a utility measure, *utility entropy* can be defined as the minimum bit rate required to reconstruct a picture and achieve complete feature recognition equivalent to perceptually lossless pictures, including the original as illustrated in Figure 6.3.

6.2 Models and Approaches to QoE Assessment

Objective QoP measures or metric designs for the purpose of QoE assessment can be classified based on the model and approach which they use and follow. The perceptual distortion or perceived difference between the reference and the processed visual signal can be formulated by applying the HVS process either to the two signals individually before visually significant signal differences are computed, or to the differences of the two signals in varying forms to weigh up their perceptual contributions to the overall perceptual score [19, 21].

6.2.1 Feature-Driven Models with Principal Decomposition Analysis

A 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 the HVS (e.g., Contrast Sensitivity Function (CSF), luminance adaption, and spatiotemporal masking effects), and optimizing coefficients to maximize the correlation of picture quality/distortion estimate with the MOS from subjective test data.

An objective Picture Quality Scale (PQS) was introduced by Miyahara in [51] and further refined in [52]. The design philosophy of PQS is summarized in [53], which leads to a metric construct consisting of the generation of visually adjusted and/or weighted distortion and distortion feature maps (i.e., images), the computation and normalization of distortion indicators (i.e., measures), decorrelated principal perceptual indicators by 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). Among the features considered by the PQS are a luminance coding error, considering the contrast sensitivity and brightness sensitivity described by Weber–Fechner's law [15], a 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. The PDA is used to decorrelate any overlap between these distortion indicators based on 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 measure [18].

A similar approach was followed by an early representative video quality assessment metric by ITS[14] [54], $\hat{s}$ (s-hat),[15] leading to the standardized Video Quality Metric (VQM) in the ANSI and the ITU-T objective perceptual video quality measurement standards [16, 55]. Seven

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

[15] $\hat{s}$ 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 [54].

parameters (including six impairment indicators/measures[16] and one picture quality improvement indicator/measure[17]) are used in linear combination to form the general VQM model with parameters optimized using the iterative nested least-squares algorithm to fit against a set of subjective training data. The general VQM model was reported in [55] to have performed statistically better than, or at least equivalent to, other models recommended in [16] 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 [52, 55, 87], which are deemed to be visually significant in the perception of picture quality, and a pooling strategy for formulation of an overall distortion measure with parameters optimized by a regression process to fit a set of subjective test data.

6.2.2 Natural Scene Statistics Model-Based Perceptual Metrics

The Natural Scene Statistics (NSS) model-based approach to QoP measurement is based on the hypothesis that modeling natural scenes and modeling HVS are dual problems, and QoP can be captured by NSS [27, 56]. Of particular interest are the Structural Similarity Index (SSIM) [57] and its variants, the Visual Information Fidelity (VIF) measure [58] and the texture similarity measure [59], the former two of which have been highly referenced and used in QoP performance benchmarking in recent years, as well as frequently applied to perceptual picture coding design using RpD optimization [3].

6.2.2.1 Structural Similarity [57]

Formulation of the SSIM is based on the assumption that structural 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 affecting perceived picture quality more than non-structural distortions such as a change in brightness and contrast, spatial shift or rotation, or a Gamma correction or change [47]. The SSIM replaces pixel-by-pixel comparisons with comparisons of regional statistics [57]. The SSIM for monochrome pictures measures the similarity between a reference/original image, $\mathbf{x}_{ref}[\mathbf{n}]$,[18] and a processed image, $\mathbf{x}_p[\mathbf{n}]$, N_1 pixels high and N_2 pixels wide, in *luminance* as approximated by picture mean intensities, $\mu_{\mathbf{x}_{ref}}$ and $\mu_{\mathbf{x}_p}$, *contrast* as estimated by picture standard deviations, $\sigma_{\mathbf{x}_{ref}}$and $\sigma_{\mathbf{x}_p}$, and *structure* as measured by the cross-correlation coefficients between $\mathbf{x}_{ref}[\mathbf{n}]$ and $\mathbf{x}_p[\mathbf{n}]$, $\sigma_{\mathbf{x}_{ref}\mathbf{x}_p}$. It is defined as follows [57]:

$$SSIM(\mathbf{x}_{ref}, \mathbf{x}_p) = \left(\mathcal{L}[\mathbf{x}_{ref}, \mathbf{x}_p]\right)^{\alpha} \left(\mathcal{C}[\mathbf{x}_{ref}, \mathbf{x}_p]\right)^{\beta} \left(\mathcal{S}[\mathbf{x}_{ref}, \mathbf{x}_p]\right)^{\gamma}, \tag{6.1}$$

[16] The six impairment measures consider blur-ringing, block distortion, jerky or unnatural motion, luminance and chrominance channel noise, and error blocks due to transmission error or packet loss [55].

[17] The picture quality improvement measure considers edge sharpening and enhancement [55].

[18] It is assumed that in $\mathbf{x}[\mathbf{n}, i, \zeta]$ of Figure 6.1, $i = 1$ for a single frame and $\zeta = 1$ considering the luminance component, Y, in $YC_{\mathrm{B}}C_{\mathrm{R}}$ space [57].

where the luminance, contrast, and structure similarity measures are, respectively,

$$\mathcal{L}[\mathbf{x}_{ref}, \mathbf{x}_p] = \frac{2\mu_{\mathbf{x}_{ref}}\mu_{\mathbf{x}_p} + a_l}{\mu^2_{\mathbf{x}_{ref}}\mu^2_{\mathbf{x}_p} + a_l}, \tag{6.2}$$

$$\mathcal{C}[\mathbf{x}_{ref}, \mathbf{x}_p] = \frac{2\sigma_{\mathbf{x}_{ref}}\sigma_{\mathbf{x}_p} + a_c}{\sigma^2_{\mathbf{x}_{ref}}\sigma^2_{\mathbf{x}_p} + a_c}, \tag{6.3}$$

and

$$\mathcal{S}[\mathbf{x}_{ref}, \mathbf{x}_p] = \frac{\sigma_{\mathbf{x}_{ref}\mathbf{x}_p} + a_s}{\sigma_{\mathbf{x}_{ref}}\sigma_{\mathbf{x}_p} + a_s}. \tag{6.4}$$

a_l, a_c, and a_s are constants to avoid instability, with values selected proportional to the dynamic range of pixel values; $\alpha > 0$, $\beta > 0$, and $\gamma > 0$ are parameters to define the relative importance of the three components, and for the vector index set, $\mathcal{N} = \{\mathbf{n} = [n_1, n_2] | 0 \le n_1 \le N_1 - 1; 0 \le n_2 \le N_2 - 1\}$, encompassing all pixel locations of $\mathbf{x}_{ref}[\mathbf{n}]$ and $\mathbf{x}_p[\mathbf{n}]$, with $card(\cdot)$ denoting the cardinality of a set,

$$\mu_{\mathbf{x}_{ref}} = \frac{1}{card(\mathcal{N})}\sum_{\mathbf{n}\in\mathcal{N}} \mathbf{x}_{ref}[\mathbf{n}] \text{ and } \mu_{\mathbf{x}_p} = \frac{1}{card(\mathcal{N})}\sum_{\mathbf{n}\in\mathcal{N}} \mathbf{x}_p[\mathbf{n}], \tag{6.5}$$

$$\begin{aligned}\sigma_{\mathbf{x}_{ref}} &= \left(\frac{1}{card(\mathcal{N})}\sum_{\mathbf{n}\in\mathcal{N}} \left(\mathbf{x}_{ref}[\mathbf{n}] - \mu_{\mathbf{x}_{ref}}\right)^2\right)^{\frac{1}{2}} \\ \text{and } \sigma_{\mathbf{x}_p} &= \left(\frac{1}{card(\mathcal{N})}\sum_{\mathbf{n}\in\mathcal{N}} \left(\mathbf{x}_p[\mathbf{n}] - \mu_{\mathbf{x}_p}\right)^2\right)^{\frac{1}{2}},\end{aligned} \tag{6.6}$$

and

$$\sigma_{\mathbf{x}_{ref}\mathbf{x}_p} = \left(\frac{1}{card(\mathcal{N})}\sum_{\mathbf{n}\in\mathcal{N}} \left(\mathbf{x}_{ref}[\mathbf{n}] - \mu_{\mathbf{x}_{ref}}\right)\left(\mathbf{x}_p[\mathbf{n}] - \mu_{\mathbf{x}_p}\right)\right)^{\frac{1}{2}}. \tag{6.7}$$

To address the issues with the non-stationary nature of spatial (and temporal) picture and distortion signals, as well as the visual attention of the HVS, the SSIM is applied locally (e.g., to a defined window), leading to windowed SSIM, $SSIM_W(\mathbf{x}_{ref}, \mathbf{x}_p)$. Sliding this window across the entire picture pixel-by-pixel will result in, for example, a total number of M $\mathrm{SSIM_W}$ values

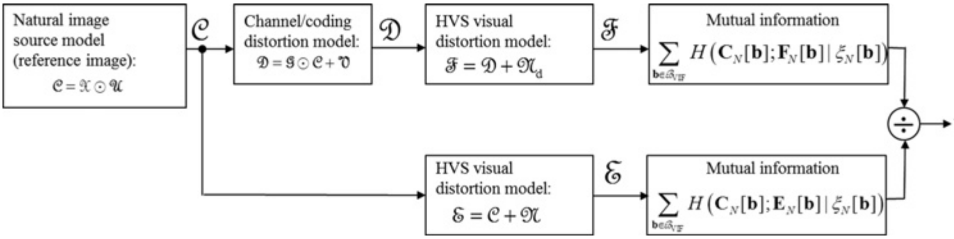

Natural image source model (reference image):
$\mathfrak{C} = \mathfrak{X} \odot \mathfrak{U}$
$\mathfrak{C}$
Channel/coding distortion model:
$\mathfrak{D} = \mathfrak{G} \odot \mathfrak{C} + \mathfrak{V}$
$\mathfrak{D}$
HVS visual distortion model:
$\mathfrak{F} = \mathfrak{D} + \mathfrak{N}_d$
$\mathfrak{F}$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

Figure 6.4 An information-theoretic framework used by VIF measurement (after [58]), where $\mathcal{C} = \{\mathbf{C}_k | k \in \mathcal{K}\} = \mathcal{X} \odot \mathcal{U} = \{\xi_k \cdot \mathbf{U}_k | k \in \mathcal{K}\}$ is the GSM, an 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 sub-band, $\mathcal{U} = \{\mathbf{U}_k | k \in \mathcal{K};\ \mathbf{U}_{k_1} \text{is independent of } \mathbf{U}_{k_2} \forall k_1 \neq k_2 \text{ and } k_1,k_2 \in \mathcal{K}\}$ a Gaussian vector RF with zero mean and covariance $\mathbf{C}_{\mathcal{U}}$, $\mathcal{X} = \{\xi_k | k \in \mathcal{K}\}$ an RF of positive scalars, symbol "⊙" defines the element-by-element product of two RFs [58], and $\mathcal{K}$ is the set of location indices in the wavelet decomposition domain; $\mathcal{D} = \{\mathbf{D}_k | k \in \mathcal{K}\} = \mathcal{G} \odot \mathcal{C} + \mathcal{V} = \{g_k \cdot \mathbf{C}_k + \mathbf{V}_k | k \in \mathcal{K}\}$, the RF representing the distorted image in the same sub-band, $\mathcal{G} = \{g_k | k \in \mathcal{K}\}$ a deterministic scalar field, $\mathcal{V} = \{\mathbf{V}_k | k \in \mathcal{K}\}$ a stationary additive zero-mean Gaussian noise RF with variance $\mathbf{C}_{\mathcal{V}} = \sigma^2_{\mathcal{V}} \mathbf{I}$ which is white and independent of $\mathcal{X}$ with identity matrix $\mathbf{I}$ and $\mathcal{U}$; $\mathcal{E} = \{\mathbf{E}_k | k \in \mathcal{K}\} = \mathcal{C} + \mathcal{N}$ and $\mathcal{F} = \{\mathbf{F}_k | k \in \mathcal{K}\} = \mathcal{D} + \mathcal{N}_d$ modeling HVS visual distortions to the reference $\mathcal{C}$ and channel/coding distortion $\mathcal{D}$, respectively, with RFs $\mathcal{N}$ and $\mathcal{N}_d$ being zero-mean uncorrelated multivariate Gaussian of M dimensions with covariance $\mathbf{C}_{\mathcal{N}} = \mathbf{C}_{\mathcal{N}_d} = \sigma^2_{\mathcal{N}} \mathbf{I}$ and $\sigma^2_{\mathcal{N}}$ the variance of the visual noise; $\mathbf{b}$ is the sub-band index and $\mathcal{B}_{\text{VIF}}$ the selected sub-band critical for VIF computation

for each pixel location in an entire picture. The overall SSIM is then computed as the average of the relevant SSIMs, as follows:

$$SSIM_{mean}(\mathbf{x}_{ref}, \mathbf{x}_p) = \frac{1}{M} \sum_{i=1,\ldots,M} SSIM_{W_i}(\mathbf{x}_{ref}, \mathbf{x}_p). \tag{6.8}$$

An 11 × 11 circular-symmetric Gaussian weighting function with standard deviation of 1.5 samples normalized to a unity sum was used in [57] for computation of the mean, standard deviation, and cross deviation in (6.5)–(6.7), respectively, to avoid blocking artifacts in the SSIM map. The SSIM has been extended to color images [60] and video [61, 90].

6.2.2.2 Visual Information Fidelity [58]

VIF formulation takes an information-theoretic approach to QoP assessment, where mutual information is used as the measure formulating source (natural scene picture statistics) model, distortion model, and HVS "visual distortion"[19] model. As shown in Figure 6.4, a Gaussian

[19] In [58], it is referred to as "HVS distortion visual noise."

Scale Mixture (GSM) model, $\mathcal{C}$, in the wavelet decomposition domain[20] is used to represent the reference picture. A Random Field (RF), $\mathcal{D}$, models the attenuation such as blur and contrast changes, and additive noise of the channel and/or coding which represent equal perceptual annoyance from 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, $\mathcal{N}$, corresponding to reference (or $\mathcal{N}_d$ for processed) in the wavelet domain. For a selected sub-band $\mathbf{b}$, where $\mathbf{b} = [s, \theta]$ with level s and orientation θ, in the wavelet transform domain, the VIF measure is defined as

$$VIF = \frac{\sum\limits_{\mathbf{b}\in\mathcal{B}_{\text{VIF}}} H\left(\mathbf{C}_N[\mathbf{b}]; \mathbf{F}_N[\mathbf{b}] | \xi_N[\mathbf{b}]\right)}{\sum\limits_{\mathbf{b}\in\mathcal{B}_{\text{VIF}}} H\left(\mathbf{C}_N[\mathbf{b}]; \mathbf{E}_N[\mathbf{b}] | \xi_N[\mathbf{b}]\right)}, \tag{6.9}$$

where the mutual information between the reference image and the perceived image in the same sub-band $\mathbf{b}$ is defined as $H\left(\mathbf{C}_N[\mathbf{b}]; \mathbf{E}_N[\mathbf{b}] | \xi_N[\mathbf{b}]\right)$, with $\xi_N[\mathbf{b}]$ being a realization of N elements in $\mathcal{X}$ for a given reference image, and that between the processed image and the image perceived by HVS is $H\left(\mathbf{C}_N[\mathbf{b}]; \mathbf{F}_N[\mathbf{b}] | \xi_N[\mathbf{b}]\right)$, with $\mathbf{C}_N[\mathbf{b}] = \left[\mathbf{C}_1, \mathbf{C}_2, ..., \mathbf{C}_N\right] \in \mathcal{C}, \xi_N[\mathbf{b}] = \left[\xi_1, \xi_2, ..., \xi_N\right] \in \mathcal{X}, \mathbf{D}_N[\mathbf{b}] = \left[\mathbf{D}_1, \mathbf{D}_2, ..., \mathbf{D}_N\right] \in \mathcal{D}, \mathbf{E}_N[\mathbf{b}] = \left[\mathbf{E}_1, \mathbf{E}_2, ..., \mathbf{E}_N\right] \in \mathcal{E}$, $\mathbf{F}_N[\mathbf{b}] = \left[\mathbf{F}_1, \mathbf{F}_2, ..., \mathbf{F}_N\right] \in \mathcal{F}$, and $\mathcal{B}_{\text{VIF}}$ the selected sub-band critical for VIF computation.

When there is no distortion, the VIF equals unity. When the 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.

6.2.2.3 Textural Similarity

The Structure Texture Similarity Metric (STSIM) measures perceived texture similarity between a reference picture and a processed counterpart to address an issue with SSIM, which tends to give low similarity values to textures which are perceptually similar. The framework used by the STSIM consists of sub-band decomposition (e.g., using steerable filter banks), computation of a set of statistics including the mean, variance, horizontal and vertical auto-correlations, and cross-band correlation, statistical comparisons and pooling scores across statistics, sub-bands and window positions. More detailed discussions and reviews of various textural similarity metrics can be found in [59].

6.2.3 HVS Model-Based Perceptual Metrics

The HVS model-based approach devises picture quality metrics to simulate the human visual perception using a model to characterize low-level vision for picture quality estimation, in terms of spatial vision, temporal vision, color vision, and foveation. Three types of HVS model

[20] The GSM model in the wavelet domain is an RF expressed as a product of two independent RFs and approximates key statistical features of natural pictures [62].

have emerged, including JND models, multichannel CGC models, and supra-threshold models, which have been applied successfully to picture quality assessment and perceptual picture coding design using RpD optimization [3]. The multichannel structure of the HVS decomposes the visual signal into several spatial, temporal, and orientation bands where masking parameters will be determined based on human visual experiments [16, 18].

6.2.3.1 JND Models

The HVS cannot perceive all changes in an image/video, nor does it respond to varying changes in a uniform manner [15, 63]. In picture coding, JND threshold detection-based HVS models are reported extensively [19–26, 28] and used in QoP assessment, perceptual quantization for picture coding, and perceptual distortion measures in RpD performance optimization for visual signal processing and transmission services [3].

The 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. The JND model can be represented in either the spatiotemporal domain or the transform/decomposition domain, or both. Examples of JND models are found with CSF, CM, and LA modeling in the DCT domain [64–67], and CSF and CM modeling using sub-band decomposition [68–71]; or in the pixel domain [72], where the key issue is to differentiate edge from textured regions [73].

A general luminance JND modeling in the sub-band decomposition domain is given by [3, 26]

$$JND_{SD}[\mathbf{k},\mathbf{b},j] = \mathcal{V}_{TBase}[\mathbf{k},\mathbf{b},j]\prod_{\wp} M_{\wp}[\mathbf{k},\mathbf{b},j] \tag{6.10}$$

where $\mathcal{V}_{TBase}[\mathbf{k},\mathbf{b},j]$ is the base visibility threshold at the location $\mathbf{k}$ in sub-band $\mathbf{b}$ of frame j determined by spatiotemporal CSF, and $\mathcal{M}_{\wp}[\mathbf{k},\mathbf{b},j]$, $\wp \in \{intra, inter, temp, lum, \dots\}$, represents different elevation factors due to intra-band (*intra*) masking, inter-band (*inter*) masking, temporal (*temp*) masking, luminance (*lum*) adaptation, and so on. The frame index j is redundant for single-frame images.

It is well known that HVS sensitivity reaches its maximum at the fovea over two degrees of the visual angle and decreases toward the peripheral retina, which spans 10–15° of visual angle [74]. While JND accounts for the local response, Visual Attention (VA) models the global response. In the sub-band decomposition domain, Foveated JND (FJND) can be modeled as follows [3]:

$$FJND_{SD}[\mathbf{k},\mathbf{b}] = JND_{SD}[\mathbf{k},\mathbf{b}] \cdot \mathcal{M}_{va}[V[\mathbf{k}]] \tag{6.11}$$

where $JND_{SD}[\mathbf{k},\mathbf{b}]$ is defined in (6.10), $\mathcal{M}_{va}[V[\mathbf{k}]]$ denotes the modulatory function determined by $V[\mathbf{k}]$ and usually taking a smaller value with larger $V[\mathbf{k}]$, which denotes the VA estimation corresponding to spatial frequency location **k** in block **b**. JND is a special case of FJND when VA is not considered and $\mathcal{M}_{va}[V[\mathbf{k}]]$ reduces to unity (i.e., $\mathcal{M}_{va}[V[\mathbf{k}]] = 1$).

There are two approaches to JND modeling for color pictures (i.e., modeling of Just-Noticeable Color Difference (JNCD)). Each color component channel can be modeled independently in a similar way to that in which the luminance JND model is formulated. Alternatively, JNCD can be modeled by a base visibility threshold of distortion for all colors, $\mathrm{JNCD}_{00}(\mathbf{n})$,[21] modulated by the masking effect of the non-uniform neighborhood (measured by the variance) represented by $\mathcal{M}_{nbh}[\mathbf{n}]$ and a scale function $\mathcal{M}_{slf}[\mathbf{n}]|_{\varsigma=1}$, modeling the masking effect induced primarily by local changes of luminance (measured by the gradient of the luminance component), assuming that the CIELAB color space $\Xi = \{L, a, b\}$ is used, and $\zeta = 1$ corresponds to the L component, as follows:

$$JNCD[\mathbf{n}] = JNCD_{00}[\mathbf{n}]g\mathcal{M}_{nbh}[\mathbf{n}]g\mathcal{M}_{slf}[\mathbf{n}]\Big|_{\varsigma=1}, \tag{6.12}$$

where **n** is the pixel coordinate vector in a pixel domain formulation.

Based on the JND model, a Peak Signal-to-Perceptual-Noise Ratio (PSPNR) was devised in [76] as follows:

$$PSPNR\,(i,\zeta) = \frac{255 \times 255}{\frac{1}{card(\mathcal{N})}\sum\limits_{\mathbf{n}\in\mathcal{N}}\left(\left|\mathbf{x}_{ref}[\mathbf{n},i,\zeta] - \mathbf{x}_{rec}[\mathbf{n},i,\zeta]\right| - JND_{ST}[\mathbf{n},i,\zeta]\right)^2 \delta[\mathbf{n},i,\zeta]}, \tag{6.13}$$

where $\mathcal{N} = \{\mathbf{n} = [n_1, n_2]|0 \le n_1 \le N_1 - 1; 0 \le n_2 \le N_2 - 1\}$, $\mathbf{x}_{ref}$ and $\mathbf{x}_{rec}$ are the reference and the reconstructed pictures, respectively,

$$\delta[\mathbf{n},i,\zeta] = \begin{cases} 1 \text{ if } \left|\mathbf{x}_{ref}[\mathbf{n},i,\zeta] - \mathbf{x}_{rec}[\mathbf{n},i,\zeta]\right| \ge JND_{ST}\,[\mathbf{n},i,\zeta], \\ 0 \qquad\qquad\qquad \text{otherwise} \end{cases} \tag{6.14}$$

$$JND_{ST}[\mathbf{n},i,\zeta] = JND_S[\mathbf{n},i,\zeta] \cdot \mathcal{M}_T\,[ILD\,[\mathbf{n},i,\zeta]], \tag{6.15}$$

$$\begin{aligned} JND_S[\mathbf{n},i,\zeta] &= JND_p[\mathbf{n}] \\ &= JND_{pL}\,[\mathbf{n}] + JND_{pT}[\mathbf{n}] - \kappa g\min\left\{JND_{pL}[\mathbf{n}], JND_{pT}[\mathbf{n}]\right\}. \end{aligned} \tag{6.16}$$

[21] The base color difference visibility threshold is determined using the color difference as measured by a more uniform color metric, calculated in polar coordinates of the CIELAB space with luminance, chroma, and hue components given by [17, 75]

$$\Delta E_{00} = \sqrt{\left(\frac{\Delta L'}{\alpha_L \cdot S_L}\right)^2 + \left(\frac{\Delta C_{ab}'}{\alpha_C \cdot S_C}\right)^2 + \left(\frac{\Delta H_{ab}'}{\alpha_H \cdot S_H}\right)^2 + R_T\left(\frac{\Delta C_{ab}'}{\alpha_C \cdot S_C}\right) \cdot \left(\frac{\Delta H_{ab}'}{\alpha_H \cdot S_H}\right)},$$

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.

In (6.16), the luminance adaptation factor $JND_{pL}[\mathbf{n}]$ at pixel location $\mathbf{n}$ can be decided according to the luminance in the pixel neighborhood; the texture masking factor $JND_{pT}[\mathbf{n}]$ can be determined via the weighted average of gradients around $\mathbf{n}$ [72] and refined with more detailed signal classification [73]; κ accounts for the overlapping effect between JND_{pL} and JND_{pT}, and $0< \kappa \leq 1$. For video, factor $JND_{p}[\mathbf{n}]$ can be multiplied further by an elevation factor $\mathcal{M}_T\,[ILD\,[\mathbf{n}, i, \zeta]]$ as in (6.15) to account for the temporal masking effect, which is depicted by a convex function of inter-frame change formulated in (6.17) [76]:

$$ILD\,[\mathbf{n}, i] = \frac{\mathbf{x}[\mathbf{n}, i] - \mathbf{x}[\mathbf{n}, i-1] + \mathbf{x}_{BG}[\mathbf{n}, i] - \mathbf{x}_{BG}[\mathbf{n}, i-1]}{2}, \tag{6.17}$$

where $\mathbf{x}[\mathbf{n}, i]$ denotes the pixel value of the ith frame and $\mathbf{x}_{BG}[\mathbf{n}, i]$ the average background luminance of the ith frame.

When $JND_{ST}[\mathbf{n}, i, \zeta]]|_{\forall\zeta} \equiv 0$ in (6.13), the PSPNR reduces to the PSNR.

A Visual Signal-to-Noise Ratio (VSNR) is devised in [77] using a wavelet-based visual model of masking and summation, which claims low computational complexity and low memory requirements.

6.2.3.2 Multichannel Vision Model

Contrast Gain Control (CGC) [78] has been used successfully in varying implementations for JND detection, QoP assessment, and perceptual picture coding in either standalone or embedded forms [18–21, 37, 43, 79–82], with Picture Quality Rating (PQR) extended from the original Sarnoff's Visual Discrimination Model–JNDmetrix$^{\text{TM}}$ [83, 84], extensively documented in ITU-T J.144 recommendation and frequently used as a benchmark [16].

An example of the CGC model in the visual decomposition domain used in [82] is described briefly for embedding a perceptual distortion measure in RpD optimization of a standard compliant coder. As shown in Figure 6.5, it consists of a frequency transform (with 9/7 filter) [85], CSF weighting, intra-band and inter-orientation masking, detection and pooling.

Given the Mallat DWT decomposition [86] of an image, $\mathbf{x}\,[\mathbf{n}, \zeta]$, for $\mathbf{n} \in \mathcal{N} = \{\mathbf{n} = [n_1, n_2] | 0 \leq n_1 \leq N_1 - 1; 0 \leq n_2 \leq N_2 - 1\}$, $\mathbf{X}_{DWT}\,[\mathbf{k}, \mathbf{b}, z]$, where $\mathbf{b} = [s, \theta]$ defines the decomposition level or scale $s \in \{1, 2, ..., 5\}$ representing five levels and orientation $\theta \in \Theta = \{\theta_0|\text{LL band}, \theta_1|\text{LH band}, \theta_2|\text{HL band}, \theta_3|\text{HH band}\}$ representing three orientations, and $\mathbf{k} = [k_1, k_2]$ with k_1 and k_2 as the row and column spatial frequency indices within the band specified by $\mathbf{b}$, the CGC model for a designated color channel has a masking response function of the form [82]

$$\mathbf{X}_z\,[\mathbf{k}, \mathbf{b}] = \rho_z \frac{\mathbf{E}_z\,[\mathbf{k}, \mathbf{b}]}{\mathbf{I}_z\,[\mathbf{k}, \mathbf{b}] + \sigma_z^q}, \tag{6.18}$$

where ζ is assumed to be 1 (representing the luminance Y component) and omitted to simplify the mathematical expressions, $\mathbf{E}_z\,[\mathbf{k}, \mathbf{b}]$ and $\mathbf{I}_z\,[\mathbf{k}, \mathbf{b}]$ are the excitation and inhibition functions,

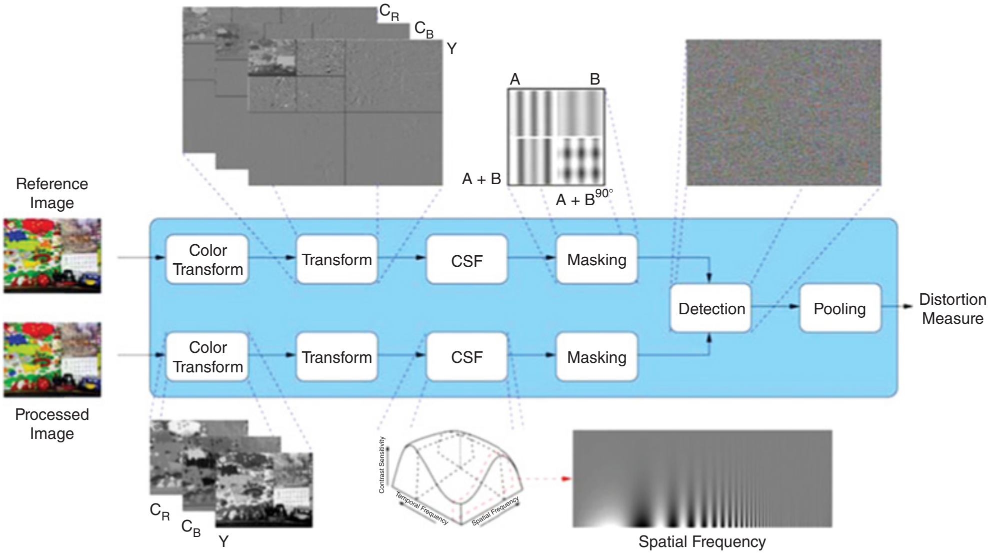


Figure 6.5 Multichannel contrast gain control model. *Source:* Tan *et al.*, 2014 [82]. Reproduced with permission of Dr. Tan

ρ_z and σ_z are the scaling and saturation coefficients, and $z \in \{\Theta, \Upsilon\}$, with Θ and Υ specifying inter-orientation and intra-frequency masking domains, respectively.

The excitation and inhibition functions of the two domains (i.e., $z \in \{\Theta, \Upsilon\}$) are given as follows:

$$\mathbf{E}_{\Theta}\left[\mathbf{k},\mathbf{b}\right] = \mathbf{X}_{CSF}^{p_{\Theta}}\left[\mathbf{k},\mathbf{b}\right], \tag{6.19}$$

$$\mathbf{E}_{\Upsilon}\left[\mathbf{k},\mathbf{b}\right] = \mathbf{X}_{CSF}^{p_{\Upsilon}}\left[\mathbf{k},\mathbf{b}\right], \tag{6.20}$$

$$\mathbf{I}_{\Theta}\left[\mathbf{k},\mathbf{b}\right] = \sum\nolimits_{\theta\in\Theta} \mathbf{X}_{CSF}^{q}\left[\mathbf{k},\mathbf{b}\right], \tag{6.21}$$

and

$$\mathbf{I}_{\Upsilon}\left[\mathbf{k},\mathbf{b}\right] = \frac{8}{card\left(\mathbf{M}_s(\mathbf{k})\right)} \sum\nolimits_{\mathbf{l}\in\mathbf{M}_s(\mathbf{k})} \mathbf{X}_{CSF}^{q}\left[\mathbf{l},\mathbf{b}\right] + \lambda^2, \tag{6.22}$$

where the exponents p_z and q represent, respectively, the excitatory and inhibitory nonlinearities and are governed by the condition $p_z > q > 0$ according to [78], $\mathbf{M}_s(\mathbf{k})$ is a neighborhood area surrounding $\mathbf{X}_{CSF}\left[\mathbf{k},\mathbf{b}\right]$, whose population is dependent on the frequency level, $s = \{1, 2, 3, 4, 5\}$ (from lowest to highest; cf. Figure 6.1(b)), such that $card\left(\mathbf{M}_s(\mathbf{k})\right) = (2s+1)^2$, and $\mathbf{X}_{CSF}\,[\mathbf{k},\mathbf{b}]$ contains the weighted transform coefficients, accounting for the CSF and defined as

$$\mathbf{X}_{CSF}\,[\mathbf{k},\mathbf{b}] = W_\delta \mathbf{X}_{DWT}\,[\mathbf{k},\mathbf{b}]\,. \tag{6.23}$$

In (6.21), $\sum_{\theta\in\Theta} \mathbf{X}^q_{CSF}\,[\mathbf{k},\mathbf{b}]$ represents the sum of transformed coefficients spanning all oriented bands. The variation in neighborhood windowing associated with $\sum_{\mathbf{l}\in\mathbf{M}_s(\mathbf{k})} \mathbf{X}^q_{CSF}\,[\mathbf{l},\mathbf{b}]$ in (6.22) addresses the uneven spatial coverage between different resolution levels in a multiresolution transform. The spatial variance, λ^2, in (6.22) has been added to the inhibition process to account for texture masking [74]. $\lambda^2 = \frac{1}{L_{1_\lambda} L_{2_\lambda}} \sum_{\mathbf{l}\in\mathbf{L}_\lambda} \left(\mathbf{X}_{CSF}\,[\mathbf{l},\mathbf{b}] - \mu\right)^2$, where $\mathbf{L}_\lambda$ denotes the code block $L_{1_\lambda} \times L_{2_\lambda}$ and μ the mean of $\mathbf{L}_\lambda$. In (6.23), W_δ for $\delta \in \{\mathrm{LL}, 1, 2, ..., 5\}$ represents six CSF weights, one for each resolution level plus an additional weight for the isotropic (LL) band and $W_\delta \propto 1/\mathcal{V}_{TBase}[\mathbf{k},\mathbf{b}]$ [76]. $\mathcal{V}_{TBase}[\mathbf{k},\mathbf{b}]$ is the base visibility threshold at the location $\mathbf{k}$ in sub-band $\mathbf{b}$ determined by spatiotemporal CSF.

In [82], a simple squared-error (or l_2-norm-squared) function is used to detect the visual difference between the visual masking responses of the reference, $\mathbf{X}_{\mathrm{Ref}z}[\mathbf{b},\mathbf{k}]$, and processed CSF-weighted DWT coefficients, $\mathbf{X}_{\mathrm{Pro}z}[\mathbf{b},\mathbf{k}]$, respectively, to form a perceptual distortion measure PDM as

$$PDM\,[\mathbf{b},\mathbf{k}] = \sum\nolimits_{z\in\{\Theta,\Upsilon\}} g_z \left|\mathbf{X}_{\mathrm{Ref}z}[\mathbf{b},\mathbf{k}] - \mathbf{X}_{\mathrm{Proc}z}[\mathbf{b},\mathbf{k}]\right|^2. \tag{6.24}$$

Here, g_z is the gain factor associated with inter-orientation (g_Θ) and intra-frequency (g_Υ) masking. In (6.24), $PDM\,[\mathbf{k},\mathbf{b}]\,|_{\mathbf{b}=[1,\theta_0]}$ is computed separately, since the LL band contains a substantial portion of the image energy in the transform domain, exhibiting a higher level of sensitivity to changes than that of all oriented bands at all resolution levels:

$$PDM\,[\mathbf{b},\mathbf{k}]\,|_{\mathbf{b}=[1,\theta_0]} = k_\Upsilon \frac{(\mathbf{X}_{\mathrm{Proc}z}[\mathbf{b},\mathbf{k}]|_{\mathbf{b}=[1,\theta_0]})^{p_\Upsilon}}{(\mathbf{X}_{\mathrm{Ref}z}[\mathbf{b},\mathbf{k}]|_{\mathbf{b}=[1,\theta_0]})^q + \sigma^q_\Upsilon}. \tag{6.25}$$

Here, k_Υ is a scaling constant, $\mathbf{X}_{\mathrm{Proc}z}[\mathbf{b},\mathbf{k}]|_{\mathbf{b}=[1,\theta_0]}$ and $\mathbf{X}_{\mathrm{Ref}z}[\mathbf{b},\mathbf{k}]|_{\mathbf{b}=[1,\theta_0]}$ are, respectively, the processed and reference visually weighted DWT coefficients for the LL band of the lowest resolution level, and σ^q_Υ is as defined in (6.18).

6.2.3.3 Supra-threshold Vision Models

A wide range of picture processing and compression applications require cost-effective solutions and belong to, more often than not, the so-called supra-threshold domain, where processing distortions or compression artifacts are visible. A supra-threshold wavelet coefficient quantization experiment 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 with respect to JND$_2$ and JND$_3$ are parallel to that of JND$_1$ [29].

The Most Apparent Distortion (MAD) measures supra-threshold distortion using a detection model and appearance model in the form of [30]

$$MAD = \left(D_{\text{detection}}\right)^{\alpha} \left(D_{\text{appearance}}\right)^{1-\alpha}, \tag{6.26}$$

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

$$\alpha = \frac{1}{1 + \beta_1 \left(D_{\text{detection}}\right)^{\beta_2}}, \tag{6.27}$$

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

6.2.4 Lightweight Bit-Stream-Based Models [9]

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 time. While no-reference picture quality metrics provide feasible solutions [25, 88], investigations have been prompted into lightweight QoE methods and associated standardization activities. There are at least three identifiable models, including the parametric model, packet-layer model, and bit-stream-layer model. With very limited information acquired or extracted from the transmission payload, stringent transmission delay constraints and limited computational resources, these models share a common technique – that is, optimization of perceptual quality or distortion predictors via, for example, regression or algorithms of similar trade using ground truth subjective test data (e.g., MOS or DMOS) and optimization criteria such as Pearson linear correlation, Spearman rank-order correlation, outlier ratio, and Root Mean Square Error (RMSE) [18, 25].

6.2.4.1 Parametric Model

Relying on Key Performance Indicators (KPIs) collected by network equipment via statistical analysis, a crude prediction of perceived picture quality or distortion is formulated using (bit) Rate (R) and Packet Loss Rate (PLR) along with side information (e.g., codec type and video resolution), which may be used to assist the adaptation of model parameters to differently coded pictures. Since the bit rate does not correlate well with the MOS data for pictures of varying content, packet loss occurring at different locations in a bit stream may have significantly different impact on perceived picture quality [3]; the quality estimation accuracy based on this model is limited, while the computation required is usually trivial.

6.2.4.2 Packet-Layer Model

With more information available via packet header analysis, distortions at picture frame level can be estimated better with information on coding parameters such as frame type and bit rate per frame, frame rate and position of lost packets, as well as PLR. The temporal complexity of video content can be estimated using ratios between bit rates of different type of frame. This enables temporal pooling for better quality or distortion prediction.

6.2.4.3 Bit-Stream-Layer Model

By accessing the media payload as well as packet-layer information, this model allows picture quality estimation either with or without pixel information [21, 88].

6.3 Offline and Online Evaluation

Evaluation of QoP for visual communication, broadcasting, and entertainment services may be conducted at different points between the source and the receiver [6, 9, 25, 88] for the purpose of product, system, or service provider quality control, monitoring and regulation/optimization, performance benchmarking; QoP monitoring and regulation/optimization along transmission path(s) within the network (e.g., at nodes) [89]; and QoP advisory and feedback at the receiver. The suitability of QoP measures based on various models and approaches to software or hardware online or offline performance evaluation depends on the measurement point/location in the encoding, transmission, and decoding chain, the availability of the reference video sequence(s), obtainable hardware and/or software computing resources, and the computational complexity of the QoP metrics. Table 6.1 shows the feasibility of QoP measures based on various models and approaches for online or offline assessments.

6.4 Remarks

To conclude this chapter, a number of observations can be made with respect to the current state of play in QoE for visual signal compression and transmission.

First, HVS model-based quality metrics have higher computational complexity than feature-driven-based, NSS-based, or lightweight quality measures, which makes software online solutions to QoE assessment all but impractical based on current computing technologies, if not entirely impossible, for most quality monitoring applications. Hardware online solutions have been demonstrated for full-reference quality assessments, which have a higher degree of system complexity and incur considerably more cost compared with alternative approaches.

Second, existing IQA and VQA metrics [16] have demonstrated their ability and success in grading the quality of pictures, corresponding to the traditional ACR subjective test data [11, 16]. However, it remains a challenge whether or not these metrics can be equally effective and able to produce accurate and robust values which correspond to JNND, JND_1, JND_2,

Table 6.1 Feasibility of QoP measures for online or offline assessment
On-: online evaluation: Off-: offline evaluation: Y: suitable; N: unsuitable.

		Encoding								**Network nodes**				**Decoding**				
		Coder R-DO				**Coder Evaluation**				**Evaluation**				**Evaluation**				
		Hardware		**Software**		**Hardware**		**Software**		**Hardware**		**Software**		**Hardware**		**Software**		
Type of metrics		On-	Off-	On-	Off-	On-	Off-	On-	Off-	On-	Off-	On-	Off-	On-	Off-	On-	Off-	**Computation Complexity**
HVS model	JND model based	Y	Y	N	Y	Y	Y	Y	Y									Moderate to high
	Multichannel model based	Y	Y	N	Y	Y	Y	N	Y									High
	Suprathreshold vision model based																	Moderate to high
Feature	PQS	Y	Y	N	Y	Y	Y	N	Y									Moderate to high
	s-hat	Y	Y	N	Y	Y	Y	N	Y									Moderate
	VQM	Y	Y	N	Y	Y	Y	N	Y									Moderate
NSS	SSIM	Y	Y		Y	Y	Y	Y	Y									Moderate
Model	VIF	Y	Y		Y	Y	Y		Y									High
	STSIM	Y	Y		Y	Y	Y		Y									Moderate
Light weight	Parametric model									Y		Y		Y		Y		Low
	Packet layer model									Y		Y		Y		Y		Low
	Bit-stream layer model									Y	Y	Y	Y	Y	Y	Y	Y	Low to moderate

etc., respectively, for quality-driven perceptually lossless and/or perceptual quality-regulated coding and transmission applications.

Third, there has been an obvious lack of reports on HVS modeling and perceptual distortion measures which capture 3-D video coding artifacts and distortions for 3-D visual signal coding and transmission applications.

Fourth, there have been very limited investigations into QoE assessment which integrates audio and visual components beyond the preliminary based on human perception and integrated human audiovisual system modeling [6, 7].

Significant theoretical and practical contributions to QoE research and development are required to complete the ongoing transition in audiovisual communications, broadcasting, and entertainment systems, and applications from the best-effort rate-driven technology-centric service to a quality-driven user-centric quality-ensured experience [3].

Acknowledgments

H.R. Wu is indebted to all his past and present collaborators and co-authors of joint publications relevant to the subject matter for their invaluable contributions to the material this chapter is sourced from and based on. Special thanks go to Professor W. Lin of Nanyang Technological University, Singapore, Dr. A.R. Reibman of AT&T Research, USA, Professor F. Pereira of Instituto Superior Tecnico-Insituto de Telecomunicacoes, Portugal, Professor S.W. Hemami of Northeastern University, USA, Professor F. Yang of Xidian University, China, Professor S. Wan of Northwestern Polytechnical University, China, Professor L.J. Karam of Arizona State University, USA, Professor K.R. Rao of University of Texas at Arlington, USA, Dr. D.M. Tan of HD2 Technologies Pty Ltd, Australia, Dr. D. Wu of HD2 Technologies Pty Ltd, Australia, Dr. T. Ferguson of Flexera Software, Australia, and Dr. C.J. van den Branden Lambrecht.

References

[1] International Telecommunication Union, Telecommunication Standardization Sector (ITU-T), ‘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, July 2008.

[2] International Telecommunication Union, Telecommunication Standardization Sector (ITU-T), ‘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, December 2011.

[3] Wu, H.R., Reibman, A., Lin, W., Pereira, F., and Hemami, S., ‘Perception-based visual signal compression and transmission’ (invited paper). Special Issue on Perception-Based Media Processing. *Proceedings of the IEEE*, **101**(9), 2013, 2025–2043.

[4] Rouse, D.M., Hemami, S.S., Pépion, R., and Le Callet, P., ‘Estimating the usefulness of distorted natural images using an image contour degradation measure.’ *Journal of the Optical Society of America A*, **28**(2), 2011, 157–188.

[5] International Telecommunication Union, Telecommunication Standardization Sector (ITU-T), ‘Opinion model for video-telephony applications.’ Rec. G.1070, April 2007.

[6] Coverdale, P., Möller, S., Raake, A., and Takahashi, A., ‘Multimedia quality assessment standards in ITU-T SG12.’ *IEEE Signal Processing Magazine*, **28**(6), 2011, 91–97.

[7] Pinson, M.H., Ingram, W., and Webster, A., ‘Audiovisual quality components.’ *IEEE Signal Processing Magazine*, **28**(6), 2011, 60–67.

[8] International Telecommunication Union, Telecommunication Standardization Sector (ITU-T), 'P.NBAMS ToR.' SG12 Doc. TD-379, September 2010.

[9] Yang, F. and Wan, S., 'Bitstream-based quality assessment for networked video: A review., *IEEE Communications Magazine*, November 2012, pp. 203–209.

[10] International Telecommunication Union, Radiocommunication Sector (ITU-R), 'Methodology for the subjective assessment of the quality of television pictures.' Rec. BT.500-13, January 2012.

[11] International Telecommunication Union, Telecommunication Standardization Sector (ITU-T), 'Subjective video quality assessment methods for multimedia applications.' Rec. P.910, April 2008.

[12] Corriveau, P., 'Video quality testing.' In Wu, H.R. and Rao, K.R. (eds), *Digital Video Image Quality and Perceptual Coding*. CRC Press, Boca Raton, FL, 2006, pp. 125–153.

[13] International Telecommunication Union, Radiocommunication Sector (ITU-R), 'Studio encoding parameters of digital television for standard 4:3 and wide-screen 16:9 aspect ratios.' Rec. BT.601-7, March 2011.

[14] International Telecommunication Union, Radiocommunication Sector (ITU-R), 'Parameter values for the HDTV standards for production and international programme exchange.' Rec. BT.709-5, April 2002.

[15] Wandell, B.A., *Foundations of Vision*. Sinauer, Sunderland, MA, 1995.

[16] International Telecommunication Union, Telecommunication Standardization Sector (ITU-T), 'Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference.' Rec. J.144, March 2004.

[17] Zhang, X. and Wandell, B.A., 'Color image fidelity metrics evaluated using image distortion maps.' *Signal Processing*, **70**, 1998, 201–214.

[18] Yu, Z., Wu, H., Winkler, S., and Chen, T., 'Objective assessment of blocking artifacts for digital video with a vision model' (invited paper). *Proceedings of the IEEE*, **90**(1), 2002, 154–169.

[19] Watson, A.B. (ed.), *Digital Images and Human Vision*. MIT Press, Cambridge, MA, 1993.

[20] van den Branden Lambrecht, C. (ed.), Special Issue on Image and Video Quality Metrics. *Signal Processing*, **70**(3), 1998.

[21] Wu, H.R. and Rao, K.R. (eds), *Digital Video Image Quality and Perceptual Coding*. CRC Press, Boca Raton, FL, 2006.

[22] Muntean, G.-M., Ghinea, G., Frossard, P., Etoh, M., Speranza, F., and Wu, H.R. (eds), Special Issue on Quality Issues on Multimedia Broadcasting. *IEEE Transactions on Broadcasting*, **54**(3), 2008.

[23] Karam, L.J., Ebrahimi, T., Hemami, S., *et al.* (eds), Special Issue on Visual Media Quality Assessment. *IEEE Journal on Selected Topics in Signal Processing*, **3**(2), 2009.

[24] Lin, W., Ebrahimi, T., Loizou, P.C., Moller, S., and Reibman, A.R. (eds), Special Issue on New Subjective and Objective Methodologies for Audio and Visual Signal Processing. *IEEE Journal on Selected Topics in Signal Processing*, **6**(6), 2012.

[25] Hemami, S.S. and Reibman, A.R., 'No-reference image and video quality estimation: Applications and human-motivated design.' *Signal Processing: Image Communication*, **25**, 2010, 469–481.

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

[27] Bovik, A.C., 'Automatic prediction of perceptual image and video quality' (invited paper). Special Issue on Perception-Based Media Processing. *Proceedings of the IEEE*, **101**(9), 2013, 2008–2024.

[28] Lin, W., 'Computational models for just-noticeable difference.' In Wu, H.R. and Rao, K.R. (eds), *Digital Video Image Quality and Perceptual Coding*. CRC Press, Boca Raton, FL, 2006, pp. 281–303.

[29] Ramos, M.G. and Hemami, S.S., 'Suprathreshold wavelet coefficient quantization in complex stimuli: Psychophysical evaluation and analysis.' *Journal of the Optical Society of America A*, **18**(10), 2001, 2385–2397.

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

[31] Corriveau, P., Gojmerac, C., Hughes, B., and Stelmach, L., 'All subjective scales are not created equal: The effect of context on different scales.' *Signal Processing*, **77**(1), 1999, 1–9.

[32] Plompen, R., Motion video coding for visual telephony. PTT Research Neher Laboratories, 1989.

[33] ANSI T1.801.02-1995, Digital Transport of Video Teleconferencing/Video Telephony Signals – Performance Terms, Definitions, and Examples. American National Standard for Telecommunications, ANSI, 1995.

[34] Yuen, M. and Wu, H.R., 'A survey of hybrid MC/DPCM/DCT video coding distortions.' *Signal Processing*, **70**, 1998, 247–278.

[35] Chandler, D.M., 'Seven challenges in image quality assessment: Past, present, and future research.' *Signal Processing*, 2013, **2013**, 1–53.

[36] Wu, H.R., Yu, Z., and Qiu, B., 'Multiple reference impairment scale subjective assessment method for digital video.' Proceedings of the 14th International Conference on Digital Signal Processing, Santorini, Greece, July 2002, Vol. 1, pp. 185–189.

[37] Wu, D., Tan, D.M., Baird, M., DeCampo, J., White, C., and Wu, H.R., 'Perceptually lossless medical image coding.' *IEEE Transactions on Medical Imaging*, **25**(3), 2006, 335–344.

[38] Oh, H., Bilgin, A., and Marcellin, M.W., 'Visually lossless encoding for JPEG2000.' *IEEE Transactions on Image Processing*, **22**(1), 2013, 189–201.

[39] Peli, E., 'Contrast in complex images.' *Journal of the Optical Society of America A*, **7**(10), 1990, 2032–2040.

[40] Chandler, D.M. and Hemami, S.S., 'Effects of natural images on the detectability of simple and compound wavelet subband quantization distortions.' *Journal of the Optical Society of America A*, **20**(7), 2003, 1164–1180.

[41] Budrikis, Z.L., 'Visual fidelity criterion and modeling.' *Proceedings of the IEEE*, **60**(7), 1972, 771–779.

[42] Mannos, J.L. and Sakrison, D.J., 'The effects of a visual fidelity criterion on the encoding of images.' *IEEE Transactions on Information Theory*, **20**(4), 1974, 525–536.

[43] Pica, A., Isnardi, M., and Lubin, J., 'HVS based perceptual video encoders.' In Wu, H.R. and Rao, K.R. (eds), *Digital Video Image Quality and Perceptual Coding*. CRC Press, Boca Raton, FL, 2006, pp. 337–360.

[44] Shannon, C.E., 'Coding theorems for a discrete source with a fidelity criteria.' IRE National Convention Record, Vol. 7, 1959, pp. 142–163.

[45] Berger, T. and Gibson, J.D., 'Lossy source coding.' *IEEE Transactions on Information Theory*, **44**(10), 1998, 2693–2723.

[46] Jayant, N.S. and Noll, P., *Digital Coding of Waveforms: Principles and Applications to Speech and Video*. Prentice-Hall, Englewood Cliffs, NJ, 1984.

[47] Wang, Z. and Bovik, A.C., 'Mean squared error: Love it or leave it.' *IEEE Signal Processing Magazine*, **26**(1), 2009, 98–117.

[48] Jayant, N.S., Johnston, J., and Safranek, R., 'Signal compression based on models of human perception.' *Proceedings of the IEEE*, **81**(10), 1993, 1385–1422.

[49] Shannon, C.E., 'A mathematical theory of communication.' *Bell Systems Technology Journal*, **27**, 1948, 379–423.

[50] Clarke, R.J., *Transform Coding of Images*. Academic, New York, 1985.

[51] Miyahara, M., 'Quality assessments for visual service.' *IEEE Communications Magazine*, **26**, 1988, 51–60.

[52] Miyahara, M., Kotani, K., and Algazi, V.R., 'Objective picture quality scale (PQS) for image coding.' *IEEE Transactions on Communications*, **46**(9), 1998, 1215–1226.

[53] Miyahara, M. and Kawada, R., 'Philosophy of picture quality scale.' In Wu, H.R. and Rao, K.R. (eds), *Digital Video Image Quality and Perceptual Coding*. CRC Press, Boca Raton, FL, 2006, pp. 181–223.

[54] Webster, A.A., Jones, C.T., Pinson, M.H., Voran, S.D., and Wolf, S., 'An objective video quality assessment system based on human perception.' *SPIE Proceedings: Human Vision, Visual Processing, and Digital Display IV*, 1913, **1993**, 15–26.

[55] Pinson, M.H. and Wolf, S., 'A new standardized method for objectively measuring video quality.' *IEEE Transactions on Broadcasting*, **50**(3), 2004, 312–322.

[56] Párraga, C.A., Troscianko, T., and Tolhurst, D.J., 'The effects of amplitude-spectrum statistics on foveal and peripheral discrimination of changes in natural images, and a multi-resolution model.' *Vision Research*, **45**(25&26), 2005, 3145–3168.

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

[58] Sheikh, H. and Bovik, A.C., 'Image information and visual quality.' *IEEE Transactions on Image Processing*, **15**(2), 2006, 430–444.

[59] Pappas, T.N., Neuhoff, D.L., de Ridder, H., and Zujovic, J., 'Image analysis: Focus on texture similarity' (invited paper). Special Issue on Perception-Based Media Processing. *Proceedings of the IEEE*, **101**(9), 2013, 2044–2057.

[60] Hassan, M. and Bhagvati, C., 'Structural similarity measure for color images.' *International Journal of Computer Applications*, **43**(14), 2012, 7–12.

[61] Wang, Z. and Li, Q., 'Video quality assessment using a statistical model of human visual speed perception.' *Journal of the Optical Society of America A*, **24**(12), 2007, B61–B69.

[62] Wainwright, M.J., Simoncelli, E.P., and Wilsky, A.S., 'Random cascades on wavelet trees and their use in analyzing and modeling natural images.' Applied and Computational Harmonics Analysis, 11, 2001, 89–123.

[63] Sekuler, R. and Blake, R. *Perception*, 3rd edn. McGraw-Hill, New York, 1994.

[64] Ahumada, A.J., 'Luminance-model-based DCT quantization for color image compression.' *SPIE Proceedings: Human Vision, Visual Processing and Digital Display III*, 1666, **1992**, 365–374.

[65] Peterson, H.A., Ahumada, A.J., and Watson, A.B., 'Improved detection model for DCT coefficient quantization.' *SPIE Proceedings: Human Vision, Visual Processing and Digital Display*, 1913, **1993**, 191–201.

[66] Watson, A.B., 'DCTune: A technique for visual optimization of DCT quantization matrices for individual images.' Society for Information Display Digest of Technical Papers XXIV, 1993, pp. 946–949.

[67] Safranek, R.J., 'A JPEG compliant encoder utilizing perceptually based quantization.' *SPIE Proceedings: Human Vision, Visual Processing and Digital Display V*, **2179**, 1994, 117–126.

[68] Safranek, R.J. and Johnston, J.D., 'A perceptually tuned subband image coder with image dependent quantization and post-quantization.' Proceedings of IEEE ICASSP, 1989, pp. 1945–1948.

[69] Chou, C.-H. and Li, Y.-C., 'A perceptually tuned subband image coder based on the measure of just-noticeable distortion profile.' *IEEE Transactions on Circuits and Systems for Video Technology*, **5**, 1995, 467–476.

[70] Höntsch, I. and Karam, L.J., 'Locally adaptive perceptual image coding.' *IEEE Transactions on Image Processing*, **9**(9), 2000, 1285–1483.

[71] Liu, Z., Karam, L.J., and Watson, A.B., 'JPEG2000 encoding with perceptual distortion control.' *IEEE Transactions on Image Processing*, **15**(7), 2006, 1763–1778.

[72] Yang, X.K., Lin, W.S., Lu, Z.K., Ong, E.P., and Yao, S.S., 'Just noticeable distortion model and its applications in video coding.' *Signal Processing: Image Communication*, **20**(7), 2005, 662–680.

[73] Liu, A., Lin, W., Paul, M., Deng, C., and Zhang, F., 'Just noticeable difference for image with decomposition model for separating edge and textured regions.' *IEEE Transactions on Circuits and Systems for Video Technology*, **20**(11), 2010, 1648–1652.

[74] Daly, S., 'Engineering observations from spatiovelocity and spatiotemporal visual models.' In van den Branden Lambrecht, C.J. (ed.), *Vision Models and Applications to Image and Video Processing*. Kluwer, Norwell, MA, 2001.

[75] Chou, C. and Liu, K., 'A perceptually tuned watermarking scheme for color images.' *IEEE Transactions on Image Processing*, **19**(11), 2010, 2966–2982.

[76] Chou, C.-H. and Chen, C.-W., 'A perceptually optimized 3-D subband image codec for video communication over wireless channels.' *IEEE Transactions on Circuits and Systems for Video Technology*, **6**(2), 1996, 143–156.

[77] Chandler, D.M. and Hemami, S.S., 'VSNR: A wavelet-based visual signal-to-noise ratio for natural images.' *IEEE Transactions on Image Processing*, **16**(9), 2007, 2284–2298.

[78] Watson, A.B. and Solomon, J.A., 'A model of visual contrast gain control and pattern masking.' *Journal of the Optical Society of America A*, **14**(9), 1997, 2379–2391.

[79] van den Branden Lambrecht, C.J., 'Perceptual models and architectures for video coding applications.' Ph.D. dissertation, Swiss Federal Institute of Technology, Zurich, 1996.

[80] Winkler, S., 'A perceptual distortion metric for digital color video.' *SPIE Proceedings: Human Vision and Electronic Imaging IV*, **3644**, 1999, 175–184.

[81] Tan, D.M., Wu, H.R., and Yu, Z., 'Perceptual coding of digital monochrome images.' *IEEE Signal Processing Letters*, **11**(2), 2004, 239–242.

[82] Tan, D.M., Tan, C.-S., and Wu, H.R., 'Perceptual colour image coder with JPEG2000.' *IEEE Transactions on Image Processing*, **19**(2), 2010, 374–383.

[83] Visual Information Systems Research Group, 'A methodology for imaging system design and evaluation.' Sarnoff Corporation, 1995.

[84] Visual Information Systems Research Group, 'Sarnoff JND vision model algorithm description and testing.' Sarnoff Corporation, 1997.

[85] Antonini, M., Barlaud, M., Mathieu, P., and Daubechies, I., 'Image coding using wavelet transform.' *IEEE Transactions on Image Processing*, **1**(2), 1992, 205–220.

[86] Mallat, S.G., 'A theory for multiresolution signal decomposition: The wavelet representation.' *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **11**(7), 1989, 674–693.

[87] International Telecommunication Union, Radiocommunication Sector (ITU-R), ‘Objective perceptual video quality measurement techniques for standard definition digital broadcast television in the presence of a full reference.’ Rec. BT.1683, June 2004.

[88] Yang, F., Wan, S., Xie, Q., and Wu, H.R., ‘No-reference quality assessment for networked video via primary analysis of bit stream.’ *IEEE Transactions on Circuits and Systems for Video Technology*, **20**(11), 2010, 1544–1554.

[89] Hioki, W., *Telecommunications*, 2nd edn. Prentice-Hall, Englewood Cliffs, NJ, 1995.

[90] Wang, Z., Lu, L., and Bovik, A.C., ‘Video quality assessment based on structural distortion measurement.’ *Signal Processing: Image Communication*, **19**(2), 2004, 121–132.

[91] Wu, H.R., Lin, W., and Ngan, K.N., ‘Rate-perceptual-distortion optimization (RpDO) based picture coding – issues and challenges.’ Proceedings of 19th International Conference on Digital Signal Processing, Hong Kong, August 2014, pp. 777–782.

Acronyms

DWT	Discrete Wavelet Transform
HVS	Human Visual System
IQA	Image Quality Assessment
MRIS	Multiple Reference Impairment Scale
NSS	Natural Scene Statistics
PDA	Principal Decomposition Analysis
PQS	objective Picture Quality Scale
QoE	Quality of Experience
QoP	perceived Quality of Picture
QoS	Quality of Service
REC	Recognition Equivalence Class
RT	Recognition Threshold
UoP	perceived Utility of Picture
VDU	Visual Distortion Unit
VQA	Video Quality Assessment

7

QoE Control, Monitoring, and Management Strategies

Maria G. Martini, Chaminda T.E.R. Hewage, Moustafa M. Nasrall and Ognen Ognenoski
Kingston University, UK

7.1 Introduction

New multimedia systems and services have higher quality requirements, not limited to connectivity: users expect services to be delivered according to their demands in terms of quality. In recent years, the concept of Quality of Service (QoS) has been extended to the new concept of Quality of Experience (QoE) [1], as the former only focuses on network performance (e.g., packet loss, delay, and jitter) without a direct link to perceived quality, whereas the latter reflects the overall experience of the consumer accessing and using the provided service. Experience is user- and context-dependent, that is, it involves considerations about subjective quality and users' expectations based on the cost of the service, their location, the type of service, and the convenience of using the service. However, subjective QoE evaluation is time-consuming, costly, and not suitable for use in closed-loop adaptations, hence there is a growing demand for objective QoE monitoring and control: objective, rather than subjective, QoE evaluation can be used to enable user-centric design of novel multimedia systems, including wireless systems based on recent standards (such as WiMAX and 3GPP LTE/LTE-A) through an optimal use of the available resources based on the aforementioned objective utility index.

The main aim of achieving a satisfactory QoE for the users of a system can be afforded at different layers of the protocol stack. Dynamic rate control strategies, also optimized across the users, can be considered at the application layer in order to allocate the available resources according to users' requirements and transmission conditions. Rate control was originally adopted with the goal of achieving a constant bit rate, then with the goal of adapting the

Multimedia Quality of Experience (QoE): Current Status and Future Requirements, First Edition.
Edited by Chang Wen Chen, Periklis Chatzimisios, Tasos Dagiuklas and Luigi Atzori.

source data to the available bandwidth [2]. Dynamic adaptation to variable channel and network conditions (i.e., by exploiting the time-varying information about lower layers) can also be performed.

Packet scheduling schemes across multiple users can be considered at the Medium Access Control (MAC) layer in order to adapt each stream to the available resources [3]. Content-aware scheduling can also be considered, as in [4]. At the physical layer, Adaptive Modulation and Coding (AMC) can be exploited to improve the system performance, by adapting the relevant parameters to both the channel and the source characteristics. In addition, throughput variations, resulting in lower QoE, can be smoothed out through a number of methods, including interference shaping [5], or compensated via appropriate buffering.

This chapter focuses on QoE monitoring, control, and management for different types of service. The remainder of the chapter is organized as follows. Section 7.2 focuses on QoE monitoring, describing subjective and objective methodologies, the need for real-time monitoring, and relevant technical solutions. Section 7.3 focuses on QoE management and control in different scenarios, including wireless scenarios and adaptive streaming over HTTP. The case of transmission to multiple users is addressed as an example requiring the joint management of different QoE requirements from different users. Finally, conclusions are drawn in Section 7.4.

7.2 QoE Monitoring

QoE monitoring is the key to assessing the overall enjoyment or annoyance of emerging multimedia applications and services. This could lead further to designing a system which maximizes user experience. For instance, the monitored QoE at different nodes could be used to optimize the system parameters and to maximize the user QoE in general. However, monitoring QoE is a challenge due to a range of factors associated with QoE, such as human factors (e.g., demographic and socioeconomic background), system factors (e.g., content- and network-related influences), and contextual factors (e.g., duration, time of day, and frequency of use). The overall experience can be monitored, analyzed, and measured by QoE-related parameters, which quantify the user's overall satisfaction with a service [1, 6].

Therefore, QoE monitoring goes beyond conventional QoS monitoring, which focuses on the performance of the underlying network. QoS measurements are typically obtained using objective methods, whereas monitoring and understanding QoE requires a multi-disciplinary and multi-technological approach. The methods adopted to measure QoE should account for all the factors associated with user perception or experience. At least, the major factors need to be identified in order to comprehensively evaluate the user experience for a given application. While the monitoring of each individual aspect of QoE remains a challenge, understanding the interaction among these aspects and its overall effect is a far greater challenge. With the advancement of technology, measuring certain aspects of QoE has become feasible (e.g., via psychophysical measurements and physiological measurements[7]). However, more work has to be done in order to understand the overall experience, which could be a result of multi-disciplinary research. The following discusses the state-of-the-art of QoE monitoring and technologies, potentially enabling QoE-driven applications and services.

A common challenge of any multimedia application or service provider is to ensure that the offered services meet at least the minimum quality expected by the users. The use of QoE measurements enables us to measure the overall performance of the system from the user's perspective [8]. The factors influencing QoE are specific to certain applications. For instance, perceptual video quality is the major QoE factor for video delivery services. In this case, accurate video quality measurements and monitoring at different system nodes will enable us to achieve maximum user QoE. A few solutions are commercially available to measure the QoE of different applications (e.g., Microsoft's Quality of Experience Monitoring Server).

QoE monitoring tools can be classified into two main categories, namely active monitoring and passive monitoring. In active monitoring, specific traffic is sent through the network for performance evaluation, whereas in passive monitoring devices are placed in measuring points to measure the traffic features as it passes. Both methodologies have their own advantages and disadvantages. For instance, passive monitoring does not incur additional overheads to the user traffic, whereas active monitoring inserts traffic specifically for QoE monitoring. The monitoring of QoE can take place at different stages of the end-to-end application delivery chain, for example at the head/server end or at the last mile (e.g., media gateway, home gateway, or Set-Top Box (STB)). Even though the measurements taken at the receiver provide the best knowledge of user experience, QoE monitoring at intermediate nodes will also provide a good indication of the effect of these technologies.

Traditionally, QoE measuring and monitoring is a human-based process: due to the subjective nature of QoE, the outcome should be evaluated by human observers. Therefore, subjective quality evaluation tests are the golden standard to monitor the QoE of novel media applications. Several international standards describe perceptual image and video quality evaluation procedures (e.g., ITU-R BT.500-13 [9]). However, specific standards targeting more general QoE monitoring have not been reported in the literature so far. Defining such a procedure is a challenge because of the several perceptual and physiological aspects attached to QoE. For instance, in 3D video, increased binocular disparity may provide the user an increased depth perception, which could also induce discomfort due to increased parallax. Therefore, emerging QoE monitoring procedures should be able to measure the overall effect of these affecting factors. Even though subjective QoE measuring provides the best judgment, it comes with several disadvantages. For instance, we need controlled test environments, several human subjects, time and effort to obtain subjective measurements. In addition, these standardized measurement methods cannot be deployed in real-time QoE monitoring scenarios.

Objective QoE monitoring procedures will overcome several disadvantages associated with subjective QoE measurements. However, they may lack the accuracy of the results obtained with subjective procedures. Unlike subjective tests, objective methods can employ simple algorithms to calculate different factors of QoE. For instance, Peak Signal-to-Noise Ratio (PSNR) can easily be calculated to understand the quality of images/video. However, these measures may or may not correlate well with subjective measurements. Therefore, hybrid QoE measurement tools will enable more reliable and accurate measurements, since they account for impairments which affect user perception. For stereoscopic video, the possibility of using objective quality measures of individual left and right views to predict 3D video subjective quality is discussed in [10, 11]. These methodologies evaluate both image and depth perception-related artifacts, which is key to understanding the perceived QoE by end users.

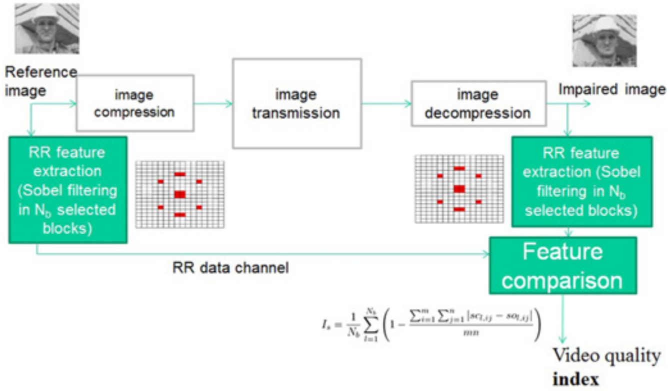

Reference image
image compression
image transmission
image decompression
Impaired image
RR feature extraction (Sobel filtering in N_b selected blocks)
RR feature extraction (Sobel filtering in N_b selected blocks)
Feature comparison
RR data channel
$I_s = \frac{1}{N_b}\sum_{l=1}^{N_b}\left(1 - \frac{\sum_{i=1}^{m}\sum_{j=1}^{n}|sc_{l,ij} - so_{l,ij}|}{mn}\right)$
Video quality
index

Figure 7.1 Reduced-reference quality metric example. *Source:* Martini *et al.*, 2012 [12] and Martini *et al.*, 2012 [13]. Reproduced with permission of Springer and Elsevier

Objective QoE monitoring tools cannot always be used in real-time monitoring tasks. For instance, monitoring the objective image or video quality in real time needs a reference image or video for comparison with the received image or video (to calculate, e.g., PSNR). It is not practical to send the original video sequence to the receiver side to measure the quality, due to the high bandwidth demand. To overcome this problem, Reduced-Reference (RR) and No-Reference (NR) quality measurements are used to measure the quality in real time. RR methods compute a quality metric based on limited information about the original source, whereas NR – or blind – metrics do not need any reference information for quality evaluation.

The RR quality evaluation methods in [12,13] extract features from the original image/video sequence. Such features are transmitted and compared with the result of the evaluation of these features at the receiver side (see Figure 7.1). The two metrics in [12, 13] adopt different strategies for quality evaluation based on feature comparison. Similarly, the RR quality evaluation method for 3D video proposed in [14] evaluates the perceived quality at the receiver side using the edge information extracted from the original 3D video sequence as side-information. This method accounts for both image artifacts as well as depth-map-related artifacts, which is important to describe the overall perception. Similar approaches are necessary for real-time QoE monitoring, in order to measure the quality at the receiver side with limited information from the sender side [15, 16]. Designing such a system will always be a challenge, because of the range of perceptual attributes attached to QoE. However, a careful identification of QoE factors (influential factors) and understanding the effect of end-to-end technologies on these aspects will enable researchers and developers to design optimum QoE measurement and monitoring tools.

Quality assessment of the received multimedia information is crucial for two main purposes:

- system performance evaluation;
- “on the fly” system monitoring and adaptation.

In the first case, the goal is to assess the final quality, reflecting the subjective quality experienced by the users (through subjective tests or objective metrics well matching subjective results). In the second case, while matching subjective results is also of importance, the main requirements are the possibility of calculating the video quality metric in real time and without reference to the original transmitted signal. QoE-driven system adaptation is addressed in Section 7.3.

Different transmission technologies result in different types of quality impairment. For instance, in transmission via RTP/UDP, packet losses are the major source of impairments. In this case, the RTP Control Protocol (RTCP) and its extended version RTCP-XR [17] enable monitoring QoS (e.g., via reports on packet loss rates) and also QoE (e.g., with voice quality metrics for Voice over IP (VoIP)).

In transmission via TCP, due to retransmissions of lost packets, delay is the main reason for QoE reduction. The remainder of this section focuses on QoE monitoring for TCP-based video streaming, focusing in particular on QoE monitoring for Dynamic Adaptive Streaming over HTTP (DASH).

7.2.1 QoE Monitoring for Adaptive Streaming over HTTP/TCP

The recent standards for HTTP streaming (e.g., DASH [18–20], developed by the Motion Pictures Expert Group (MPEG)) support a streaming client–server model with user-oriented control, where initially metadata is exchanged allowing the user to learn about content type and availability. The user fetches content from the network according to its preferences and network conditions, adaptively switching between the available multimedia content utilizing the HTTP protocol. The advantages of this technique compared with previous streaming techniques (i.e., RTP/RTSP streaming, HTTP progressive streaming) make MPEG-DASH a promising framework for adaptive transmission of multimedia data to end users.

There are two main sets of relevant parameters regarding the QoE paradigm for HTTP adaptive streaming. The first set refers to parameters that influence the overall QoE, whereas the second set refers to observable parameters that can be used directly for the derivation of QoE metrics. The important parameters from the first set are summarized on a per-level basis as follows.

- Video level: bit rate, frame rate, resolution, type of encoder.
- Transport level: sequence and timing of HTTP request, parallel TCP connections, segment durations, frequency of MPD updates.
- Radio and network level: parameters associated with bandwidth allocation, multi-user scheduling, radio access network, modulation, coding, OFDMA, time and frequency, resource/burst allocations.

The second set of parameters refers to parameters that are observed and taken into direct consideration for the derivation of QoE metrics (e.g., video quality, initial delay, frequency of rebuffering events).

With reference to the aforementioned second set of parameters, three levels of QoS for HTTP video streaming are addressed in [21]: network QoS, application QoS, and user QoS. The authors refer to user QoS as a set of observable parameters that reflect the overall QoE. The general idea of the authors is to investigate the effect of network QoS on user QoS (i.e., QoE), with application QoS as bridging level between them. Initially, they use analytical models and empirical evaluations to derive the correlation between application and network QoS, and then perform subjective experiments to evaluate the relationship between application QoS and user QoE. The network QoS is observed with active measurements and refers to the network path performance between the server and the client in terms of Round Trip Time (RTT), packet loss, and network bandwidth. Further, the following application-level QoS metrics are observed:

- T_{init}, initial buffering time;
- T_{rebuff}, mean duration of rebuffering event;
- f_{rebuff}, rebuffering frequency.

T_{init} refers to the period between the starting time of loading the video and the starting time of playing the video, T_{rebuff} measures the average duration of rebuffering event, and f_{rebuff} denotes the frequency of rebuffering event.

The user QoE is measured via the Mean Opinion Score (MOS) according to the ITU-T P.911 recommendation [22] via subjective measurements. This approach is adopted since objective metrics such as PSNR and MSE evaluate only distortion and do not take into account the Human Visual System (HVS), hence they are not suitable for video streaming.

The authors propose to estimate the QoE based on application-level QoS parameters as follows:

$$MOS = 4.32 - 0.067T_{\text{init}} - 0.742f_{\text{rebuff}} - 0.106T_{\text{rebuff}}.$$

The most significant input within this model is the frequency of rebuffering, outlined as the main factor affecting users' QoE.

The authors in [23] propose a no-reference QoE model based on Random Neural Networks (RNNs). This QoE estimation approach models two main problems regarding adaptive video streaming: playout interruptions and video quality variations due to lossy compression during the encoding of different video quality levels. The QoE estimation model considers the following parameters: the quantization parameter used in the video compression and playout interruptions that occur during the video playout. This model is a no-reference model, and hence simple compared with full- or partial-reference QoE models.

The work in [24] similarly considers video playout interruptions, but not the effect of changing the video bit rate as in the case of adaptive video streaming. The QoE model elaborated utilizes the PSQA method based on RNNs [23]. When the parameters affecting the QoE change, a new PSQA module is designed based on subjective tests. The idea is to have several distorted samples evaluated subjectively by a panel of human observers. Then the results of this evaluation are used to train an RNN in order to capture the relationship between the parameters and the QoE. During these tests the authors keep the resolution of the videos and the frame rate constant, and different video qualities are produced via different quantization parameters. The effects from the network (i.e., packet losses, delay, jitter) are included in the playout interruptions. Playout interruptions are represented as a function of three measurable parameters: the total number of playout interruptions N, the average value of interruption delays D_{avg}, and the maximum value of interruption delay $D_{\max}$. These parameters are measured over an interval containing a fixed duration of video data. The RNN-based QoE model estimation can be summarized in the following points:

- Users are more sensitive to video playout interruptions compared with QP increase for lower QP values.
- The user's QoE starts to decrease significantly for very high QP values.
- The user's QoE drops faster with increasing $D_{\max}$ and reaches a saturation point after 6 to 8 s.

Furthermore, it is outlined that, when the network bandwidth decreases, it is recommended to use a coarser representation (i.e., a lower bit rate) rather than risking even a single playout interruption. Hence, when considering the trade-off between QP and playout interruptions, the latter should be kept to a minimum at the cost of decreasing QP.

In the end, the RNN-based QoE is compared with a freeze distortion model [24] through Mean Square Error (MSE). The freeze distortion model does not take into consideration the values of the QP, hence it is significantly outperformed even for very high QP values.

An extensive study to understand the video quality impact on user engagement is presented in [25], where a large data set of different content types is used and parameters at the client side (join time, buffering ratio, average bit rate, rendering quality, and rate of buffering events) are measured in order to observe the effect of these parameters on the QoE. The analysis shows that the buffering ratio has the largest impact on the QoE regardless of the content type. Furthermore, application- and content-specific conclusions are derived; for example, the average bit rate is more significant for live content compared with VoD content.

The authors in [26] state that the buffer underrun is not enough to represent the viewers' QoE by presenting subjective tests for TCP streaming. A no-reference metric (pause intensity) is proposed, which is a product of the average pause duration and the pause frequency. The metric is derived utilizing an equation-based TCP model.

7.3 QoE Management and Control

7.3.1 Cross-Layer Design Strategies

Cross-Layer Design (CLD) solutions should be investigated in order to optimize the global system based on a QoE criterion. As an example, in [27] a CLD approach is considered with multi-user diversity, which explores source and channel heterogeneity for different users.

Typically, CLD is performed by jointly designing two layers in the protocol stack [28–31]. In [15], CLD takes the form of a network-aware joint source and channel coding approach, where source coding (at the application layer) and channel coding and modulation (at the physical layer) are jointly designed by taking the impact of the network into account. In [29], cross-layer optimization also involves two layers, the application layer and the MAC layer of a radio communications system. The proposed model for the MAC layer is suitable for a transmitter without instantaneous Channel State Information (CSI). A way of reducing the amount of exchanged control information is considered, by emulating the layer behavior in

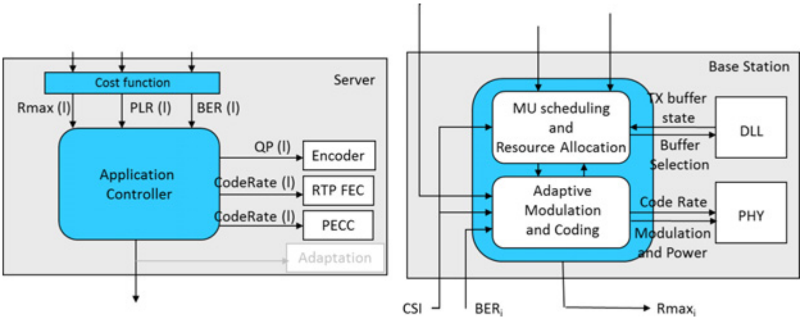


Figure 7.2 Application controller (left) and base station controller (right) proposed in the OPTIMIX European project

the optimizer based on a few model parameters to be exchanged. The parameters of the model are determined at the corresponding layer, and only these model parameters are transmitted as control information to the optimizer. The latter can tune the model to investigate several layer states without the need to exchange further control information with the layer. A significant reduction of control information to be transmitted from a layer to the optimizer is achieved, at the expense of the control information from the optimizer to the layers that might increase slightly.

The work in [28] includes in the analysis MAC-PHY and APP layers, presenting as an example a MAC/application-layer optimization strategy for video transmission over 802.11a wireless LANs based on classification.

The CONCERTO and the OPTIMIX European projects address(ed) CLD strategies, the cross-information to be exchanged, and the strategies to pass such information among the layers in mobile networks. In order to control the system parameters based on the observed data, two controller units were proposed in the OPTIMIX project: one at the application layer (APP) and one at the base station (BSC) to control lower-layer parameters [32] and in particular resource allocation among the different users based on the (aggregated) multiple feedback. A block diagram of the two controllers is shown in Figure 7.2.

The two controllers operate at different time scales, since more frequent updates are possible at the base station controller, and rely on different sets of observed parameters. For instance, the application-level controller outputs the parameters for video encoding and application-layer error protection based on the collected information on available bandwidth, packet loss ratio, and bit error rate. The base station controller performs resource allocation among users based on information on channel conditions and bit error rate, as well as quality information from the application layer. The goal of the proposed system is to provide a satisfactory quality of experience to video users, hence video quality is the major target and evaluation criterion, not neglecting efficient bandwidth use, real-time constraints, robustness, and backward compatibility.

Owing to the wide range of QoE-aware strategies (the reader can see the recent special issues [1, 33] for other examples), we focus in the following on two categories: QoE management for DASH and QoE management in wireless shared channels.

7.3.2 QoE Management for Dynamic Adaptive Streaming

In dynamic adaptive video streaming, adaptation can be performed [34–36] based on the QoE parameters discussed in Section 7.2.1. The different system parameters can be selected based on QoE requirements. In addition, video configurations can be adapted in real time with the target of optimizing the QoE [37].

For instance, the work in [38] investigates the required client's buffer size for the prescribed video quality by targeting memory-constrained applications. This approach considers video streaming over TCP and provides an analytical model for the buffer size as a function of the TCP transmission parameters, network characteristics (packet loss and round trip time), and probability of buffer underrun at the playback.

A network-oriented proposal for QoE management is presented in [39], where video adaptation is performed at a proxy at the edge between Internet and wireless core, in order to improve the DASH QoE in cellular broadband access networks. The proxy performs global optimization over multiple DASH flows by splitting the connection toward the client (hence, increasing the average throughput). Video quality-aware dynamic prioritization is adopted (low-rate streams have high priority to preserve minimal QoE), and fairness is introduced. Further, an adaptive controller is introduced in the network in order to minimize the cost function. This function depends on video distortion, bit-rate variations, and playback jitter at the clients, and is defined as a weighted sum.

The work in [40] presents a network-driven QoE approach for DASH, which jointly considers the content characteristics and wireless resource availability in Long-Term Evolution (LTE) networks in order to enhance the HTTP streaming user experience. This is achieved by rewriting the client's HTTP request at a proxy in the mobile network, since the network operator has better information regarding network conditions. The DASH client is proxy-unaware and can play the obtained segments from the network.

A DASH system with client-driven QoE adaptation is presented in [41]. The proposed adaptation logic in this work combines TCP throughput, content-related metrics, buffer level, and user requirements and expectations, which enhances the viewing experience at the user side. The analysis outlines that regarding DASH, the representation switch rate and media encoding parameters should be quantified together during the design of the QoE system. In the worst case, the proposed automated QoE model results in a 0.5-point divergence compared with the evaluation of human subjects.

An approach for using user-viewing activities in improving the QoE is presented in [42]. These activities are used to mitigate temporal structure impairments of the video. In this study, by using subjective tests, these activities are linked with the network measurements and user QoE. The results show that network information is insufficient to capture the user's dissatisfaction with the video quality. In addition, the impairments in the video can trigger user activities (pausing and reducing screen size); therefore, inclusion of pause events improves the prediction capability of the proposed model.

7.3.3 QoE Delivery to Multiple Users Over Wireless Systems

When transmitting multimedia signals to multiple users over wireless systems (see the example scenario in Figure 7.3), the trade-off between resource utilization and fairness among users

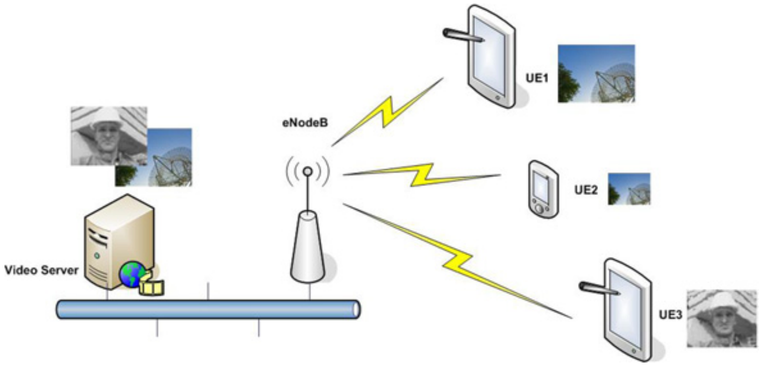


Figure 7.3 Wireless delivery to multiple users – example scenario

has to be addressed. On the one hand, the interest of network operators is to maximize the exploitation of the resources (e.g., assigning more resources to the user(s) experiencing better channel conditions). On the other hand, this strategy can result in unsatisfied users, since users experiencing worse channel conditions would not be served and would not meet their QoE requirements. For this reason, fairness among users has to be considered in scheduling and resource allocation. In recent years, several approaches have been proposed, with the goal of jointly maximizing the quality experienced by different users.

In [32] we addressed the aforementioned trade-off by focusing on fairness, targeting the maximization of the minimum weighted quality among different users. This approach results in a good level of fairness among users. However, without a proper admission control strategy, this could lead to weak exploitation of the resources and if a single user experiences very bad channel conditions, the attempt to serve this user with a reasonable quality may dramatically jeopardize the quality received by other users.

Content awareness is a key feature in providing QoE and, in recent years, a number of relevant approaches have emerged. The authors of [4, 43] investigated a content-aware resource allocation and packet scheduling for video transmission over wireless networks. They presented a cross-layer packet scheduling approach, transmitting pre-encoded video sequences over wireless networks to multiple users. This approach is used for Code Division Multiple Access (CDMA) systems, and it can be adapted for Orthogonal Frequency Division Multiple Access (OFDMA) systems such as IEEE 802.16 and LTE wireless networks. The data rates of the served users are dynamically adjusted depending on the channel quality and the gradient of a content-aware utility function, where the utility takes into account the distortion of the received video.

In multimedia applications, the content of a video packet is critical for determining the importance of the packets. Utility functions can be defined as either a function of the head-of-line packet delay, a function of each flow's queue length, or a function of each user's current average throughput. In terms of content, the utility gained due to transmitting the packet, the size of the packet in bits, and the decoding deadline for the packet (i.e., each frame's time stamp) can be considered. In addition, CSI can inform the scheduling decision. The method adopted in [4, 43] consists of ordering the packets of the encoded video according to their relative contribution to the final quality of the video, and then constructing a utility function for each packet in which its gradient reflects the contribution of each packet to the perceived video quality. Hence, the utility function is defined as a function of the decoded video quality (i.e., based on the number of packets already transmitted to a user for every frame). Further, robust data “packetization” at the encoder and realistic error concealment at the decoder are considered. The proposed utility function enables optimization in terms of the actual quality of the received video. The authors provide an optimal solution where video packets are decoded independently and a simple error concealment approach is used at the decoder. Moreover, with complex error concealment a proper solution is provided where a distortion utility is calculated. Performance evaluation is carried out and it is noticed that the proposed content-aware scheduler outperforms content-independent approaches in particular for video streaming applications. The parameters used in this scheduler are achievable rate, CSI from User Equipment (UE), a weighting parameter for fairness purposes across users (which is based on the distortion in a user’s decoded video based on the previous transmissions), and three features of each packet which are utility gained due to packet transmission, decoding deadline, and packet size.

A content-aware downlink packet scheduling scheme for multi-user scalable video delivery over wireless networks is proposed in [44]. The scheduler uses a gradient-based scheduling framework, as elaborated earlier in [4], along with the Scalable Video Coding (SVC) schemes. The reason for using SVC is to provide multiple high-quality video streams over different prevailing channel conditions for multiple users. The scheduler proposed in [44] outperforms the traditional content-independent scheduling approaches. Furthermore, the SVC encoder offers the potential to utilize packet prioritization strategies without degrading/compromising system performance. Packet prioritization can be signaled to the MAC layer (i.e., scheduler), in conjunction with the utility metrics of each packet. A distortion model is also proposed in order to efficiently and accurately predict the distortion of an SVC-encoded video stream. This model has been employed in this work in order to prioritize source packets in the queue based on their estimated impact on the overall video quality. The parameters used are achievable rate for every user, loss probability, user’s estimated channel state, and expected distortion.

A detailed review of content-aware resource allocation schemes for video transmission over wireless networks is given in [45], although this does not include some of the most recent approaches.

The authors of [46, 47] discuss a scheduling and resource allocation strategy for multi-user video streaming over Orthogonal Frequency Division Multiplexing (OFDM) downlink systems. The authors utilize SVC for encoding the video streams. This work utilizes only the temporal and quality scalability (not spatial scalability) for video coding via the adaptive resource allocation and scheduling. The authors propose a gradient-based scheduling and resource allocation strategy. This strategy prioritizes different users by considering adaptively adjusted priority weights, computed based on the video content, deadline requirements, and transmission history. A delay function is designed to cope with the effect of the deadline approaching, for which the possibility of delay violation is reduced.

The aim of the work presented in [43, 47] is to maximize the average PSNR of all SVC video users under a constrained power transmission, time-varying channel conditions, and variable-rate video content. The obtained results show that the proposed scheduler outperforms the content-blind and deadline-blind algorithms with a gain of as much as 6 dB in terms of average PSNR when the network is saturated. The parameters considered for this scheduler are average throughput to control fairness, dynamic weight based on the target delay (i.e., dynamic/desirable bit rate for the unfinished sub-flow of the video streams), video content (e.g., packet sizes at temporal and quality layers), achievable rate, CSI, length of unfinished sub-flow, playback deadline, priority weight computed based on the corresponding distortion decrease, bit rate required to deliver the current sub-flow of the video streams which consider the target delay and the bits remaining for the unfinished sub-flow, and different delay functions to decrease the delay violation which is mainly adjusted to achieve remarkable results.

In [48], the authors presented a scheduling algorithm that can be tuned to maximize the throughput of the most significant video packets, while minimizing the capacity penalty due to quality/capacity trade-off. It is shown that the level of content awareness required for optimum performance at the scheduler, and the achieved capacity, are highly sensitive to the delay constraint.

The authors of [49] propose a distortion-aware scheduling approach for video transmission over OFDM-based LTE networks. The main goal of this work is to reduce the end-to-end distortion in the application layer for every user in order to improve the video quality. Hence, parameters from the Physical (PHY), MAC, and Application (APP) layers are taken into consideration. At the APP layer, the video coding rate is extracted; at the MAC layer, Physical Resource Block (PRB) scheduling and channel feedback are exchanged; and at the PHY layer, modulation and coding parameters are used. Parameters used are frame distortion caused by lost slice, waiting time, transmitting time, latency bound, video distortion caused by Block Error Rate (BLER) – a function of modulation and coding scheme of PRB – Signal-to-Interference-Noise Ratio (SINR) of wireless channel, the dependency of the video content under the constraint of the transmitting delay, and different coding rate. Simulation results show that the proposed gradient-based cross-layer optimization can improve the video quality.

The authors of [50] propose a novel cross-layer scheme for video transmission over LTE-based wireless networks. The scheme takes into consideration the I-based and P-based packets from the APP layer, scheduling packets according to their importance from the MAC layer, and channel states from the PHY layer. The work in [50] aims to improve the perceived video quality for each user and improve the overall system performance in terms of spectral efficiency. It is assumed that I packets are more important than P packets for each user. The reason is that the loss of an I packet may lead to error propagation within the Group of Pictures (GoP). Hence, the packet scheduling algorithm at the MAC layer is adapted to prioritize I packets over P packets for each video sequence. Results show that the proposed cross-layer scheme performs better in terms of system throughput and perceived video quality. The parameters used are achievable rate, service rate requirement I and P packet queues, CSI, and I packets being more important than P packets since the loss due to error propagation may lead to error propagation within the GoP.

The authors of [51] propose a cross-layer algorithm of packet scheduling and resource allocation for multi-user video transmission over OFDMA-based wireless systems. The goal of this work is to maintain fairness across different users and minimize the received video distortion of every user by adopting a proper packet scheduling and radio resource allocation approach. Similar work is done in [52]. Furthermore, when video streams are requested by the end user, video packets of the corresponding video streams are ordered according to their contribution to the reconstructed video quality which is estimated before transmission. Then, video packets are buffered in order at the base station. Hence, video content information from the APP layer (i.e., indicating the importance of each packet on the reconstructed video quality), queue state information from the MAC layer (i.e., indicating the order of the video packets in the buffer), channel state information from the PHY layer, playback delay, and video packet size are parameters used in the proposed algorithm.

A quality-aware fair downlink packet scheduling algorithm for scalable video transmission over LTE systems was proposed in [53]: we addressed quality fairness by relying on the Nash bargaining solution. We proposed a downlink scheduling strategy for scalable video transmission to multiple users over OFDMA systems, such as the recent LTE/LTE-A wireless standard. A novel utility metric based on video quality was used in conjunction with our proposed quality-driven scheduler. The proposed metric takes into account the frame dependency in a video sequence. The streams are encoded according to the SVC standard, and hence organized in layers where the upper layers cannot be decoded unless the lower layers are correctly received. The proposed metric is called frame significance throughput. Results showed that our proposed strategy outperforms throughput-based strategies, and in particular it enables the operator of the mobile system to select the level of fairness for different users in a cell based on its business model. The system capacity in terms of satisfied users can be increased by 20% with the proposed quality-based utility, in comparison with advanced state-of-the-art throughput-based strategies.

A channel and content-aware 3D video downlink scheduling combined with a prioritized queuing mechanism for OFDMA systems is proposed in [54]. The idea behind the prioritized queuing mechanism is to prioritize the most important video layers/components with the goal of enhancing the perceived 3D video quality at the receiver. We focused on color plus depth 3D video and considered the different importance of diverse components with respect to the perceived quality. 3D video is encoded using an SVC video encoder. The priority values of the encoded 3D video components are signaled from the application layer to the MAC layer via cross-layer signaling. The users feed back their sub-channel gain to the base station, which is then used at the MAC layer for the resource allocation process. Hence, the proposed scheduler is designed to guarantee that the important layers are scheduled at every scheduling epoch over the sub-channels with higher gain. The Packet Loss Ratio (PLR) is increased for the prioritized color/depth layers at the MAC layer, at the expense of a small increase in the PLR for the less perceptual video layers. Video layers highly affected by packet losses are discarded, so as not to waste radio resources. The prioritization scheme results in a global quality improvement in the prioritized case. The parameters used are the Head of Line (HoL) packet delay, a weight that controls the throughput fairness among users, the fractional rate based on video-layer bit rate, SINR, Dependency/Temporal/Quality (DTQ) identifications of the SVC video stream, and the maximum tolerable delay based on the playout time (i.e., frame rate).

The authors of [55] propose a content-aware scheduler to allocate resources for downlink scalable video transmission over LTE-based systems. The goal of this scheduler is to control the video quality, as this can be done by allocating PRB to the users based on their available resources, link quality, and device capability. In addition, the number of available PRBs and the link quality control the scheduler decision in choosing the profile level and assigning the required number of PRBs for each user. The parameters used are SVC profile levels, Channel Quality Indicator (CQI), number available, along with a quality-driven scheduler. The quality of the video is obtained by considering two methods: one is based on a no-reference metric and the other is based on a full-reference metric (i.e., dependent on knowledge of the PSNR of the original video).

A packet scheduling approach for wireless video streaming is proposed in [56]. The wireless network is a 3G-based network. The proposed approach involves applying different deadline thresholds to video packets with different importance in order to obtain different packet loss ratios. The importance value of a video packet is determined by its relative position within its GoP and motion texture context.

The authors of [57, 58] discuss a cross-layer system design between the streaming server and the mobile Worldwide Interoperability for Microwave Access (WiMAX) base station for SVC streaming over mobile WiMAX. The aim of this work is to support the QoS of video streaming services effectively over the WiMAX network. It is worth noting that transmission packets can be classified into multiple levels of importance when using the SVC standard. The authors of [59] investigate an application-driven cross-layer approach for video transmission over OFDMA-based networks. The proposed schemes are named quality-fair and quality-maximizing, which are used to maximize the video quality with and without fairness constraints, respectively. The packet scheduling will be responsible for selecting the packets with the largest contribution to the video quality. The assumption in this design is that each video frame is partitioned into one or more slices, each slice header acts as a resynchronization marker to allow independent decoding of the slices, and each slice contains a number of macro-blocks. Hence, due to the variation of the content among different video streams, different packets make diverse contributions to the video quality. The parameters used are a quality contribution index for each packet (i.e., expressed by the decreased distortion value caused by the successful transmission of the packets), the size of the packets, maximum delay, real-time requirements for video applications, and the CSI fed back by the mobile station.

7.3.4 Application and Service-Specific QoE Management

As a final consideration, it is worth noting that different applications and services may have different definitions of quality of experience, as well as different requirements, and different methods may be used for QoE management. As an example, in medical applications the ability to use the received data to perform a diagnosis or a remote medical intervention is often the key QoE requirement [60–67]. In online gaming, interactivity is a main factor for QoE management, while in learning applications different strategies can be adopted to distribute the same content to a large number of users. The detailed study of the application-specific QoE requirements enables the most appropriate strategy for QoE management and control.

7.4 Conclusion

This chapter has presented a review of the most widely adopted strategies for QoE monitoring, control, and management, including solutions proposed by the authors. The focus was in particular on QoE provision for HTTP streaming and wireless transmission to multiple users. Designing current transmission systems with the goal of achieving QoE requirements for all involved users will enable better exploitation of resources, with a higher number of satisfied users in the system.

Acknowledgment

The authors acknowledge support from the European Union's Seventh Framework Program (FP7/2007-2013) under grant agreement no. 288502 (CONCERTO project).

References

[1] Martini, M.G., Chen, C.W., Chen, Z., Dagiuklas, T., Sun, L., and Zhu, X., 'Guest editorial: QoE-aware wireless multimedia systems. *IEEE Journal on Selected Areas in Communications*, **30**(7), 2012, 1153–1156.

[2] Chen, M. and Zakhor, A., 'Rate control for streaming video over wireless.' *IEEE Wireless Communications*, **12**(4), 2005, 32–41.

[3] Jurca, D. and Frossard, P., 'Video packet selection and scheduling for multipath streaming.' *IEEE Transactions on Multimedia*, **9**(3), 2007, 629–641.

[4] Pahalawatta, P.V., Berry, R., Pappas, T.N., and Katsaggelos, A.K., 'Content-aware resource allocation and packet scheduling for video transmission over wireless networks.' *IEEE Journal on Selected Areas in Communications*, **25**(4), 2007, 749–759.

[5] Singh, S., Andrews, J.G., and de Veciana, G., 'Interference shaping for improved quality of experience for real-time video streaming.' *IEEE Journal on Selected Areas in Communications*, **30**(7), 2012, 1259–1269.

[6] Le-Callet, P., Moeller, S., and Perkis, A., 'Qualinet white paper on definitions of quality of experience, version 1.1.' European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003), June 2012.

[7] Lassalle, J., Gros, L., and Coppin, G., 'Combination of physiological and subjective measures to assess quality of experience for audio visual technologies.' Third International Workshop on Quality of Multimedia Experience (QoMEX), 2011, pp. 13–18.

[8] Hewage, C.T. and Martini, M.G., 'Quality of experience for 3D video streaming.' *IEEE Communications Magazine*, **51**(5), 2013, 101–107.

[9] ITU-R Recommendation BT.500-13, 'Methodology for the subjective assessment of the quality of television pictures.' International Telecommunication Union, Geneva, 2012. Available at: www.itu.org.

[10] Hewage, C.T.E.R., Worrall, S.T., Dogan, S., and Kondoz, A.M., 'Quality evaluation of color plus depth map-based stereoscopic video.' *IEEE Journal of Selected Topics in Signal Processing*, **3**(2), 2009, 304–318.

[11] Hewage, C.T.E.R., Worrall, S.T., Dogan, S., and Kondoz, A.M., 'Prediction of stereoscopic video quality using objective quality models of 2-D video.' *Electronics Letters*, **44**(16), 2008, 963–965.

[12] Martini, M.G., Villarini, B., and Fiorucci, F., 'A reduced-reference perceptual image and video quality metric based on edge preservation.' *EURASIP Journal of Applied Signal Processing*, 2012, 2012, 66.

[13] Martini, M.G., Hewage, C.T., and Villarini, B., 'Image quality assessment based on edge preservation.' *Signal Processing: Image Communication*, **27**, 2012, 875–882.

[14] Hewage, C.T.E.R. and Martini, M.G., 'Edge based reduced-reference quality metric for 3D video compression and transmission.' *IEEE Journal of Selected Topics in Signal Processing*, **6**(5), 2012, 471–482.

[15] Martini, M.G., Mazzotti, M., Lamy-Bergot, C. Huusko, J., and Amon, P., 'Content adaptive network aware joint optimization of wireless video transmission.' *IEEE Communications Magazine*, **45**(1), 2007, 84–90.

[16] Atzori, L., Floris, A., Ginesu, G., and Giusto, D., 'Streaming video over wireless channels: Exploiting reduced-reference quality estimation at the user-side.' *Signal Processing: Image Communication*, **27**(10), 2012, 1049–1065.

[17] Friedman, T., Caceres, R., and Clark, A., RTP Control Protocol Extended Reports (RTCP XR). RFC 3611, November 2003, pp. 1–52.

[18] Sodagar, I., 'The MPEG-DASH standard for multimedia streaming over the Internet.' *IEEE Transactions on Multimedia*, **18**(4), 2011, 62–67.

[19] Ognenoski, O., Razaak, M., Martini, M.G., and Amon, P., 'Medical video streaming utilizing MPEG-DASH.' Proceedings of IEEE Healthcom 2013, Lisbon, Portugal.

[20] Ognenoski, O., Martini, M.G., and Amon, P., 'Segment-based teletraffic model for MPEG-DASH.' Proceedings of the IEEE 15th International Workshop on Multimedia Signal Processing (MMSP), October 2013, pp. 333–337.

[21] Mok, R.K.P., Chan, E.W.W., and Chang, R.K.C., 'Measuring the quality of experience of HTTP video streaming.' Proceedings of the IFIP/IEEE International Symposium on Integrated Network Management (IM), May 2011, pp. 485–492.

[22] ITU-T Recommendation P.911: Subjective audiovisual quality assessment methods for multimedia applications, December 1998.

[23] Mohamed, S. and Rubino, G., 'A study of real-time packet video quality using random neural networks.' *IEEE Transactions on Circuits and Systems for Video Technology*, **12**(12), 2002, 1071–1083.

[24] Okamoto, J., Watanabe, K., and Kurita, T., 'Objective video quality assessment method for evaluating effects of freeze distortion in arbitrary video scenes.' Proceedings of the IS&T/SPIE 19th Annual Symposium, January 2007.

[25] Dobrian, F., Sekar, V., Awan, A., *et al.*, 'Understanding the impact of video quality on user engagement.' *The SIGCOMM – Computer Communication Review*, **41**(4), 2011, 362–373.

[26] Bailey, C., Seyedebraihmi, M., and Xiao-Hong, P., 'Pause intensity: A no-reference quality assessment metric for video streaming in TCP networks.' Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), July 2012, pp. 818–823.

[27] Su, G.-M., Han, Z., and Liu, K., 'Multiuser cross-layer resource allocation for video transmission over wireless networks.' *IEEE Network*, March/April 2006.

[28] Dimic, G., Sidiropoulos, N.D., and Zhang, R., 'Medium access control–physical cross-layer design.' *IEEE Signal Processing Magazine*, September 2004.

[29] van der Schaar, M., Turag, D.S., and Wong, R., 'Classification-based system for cross-layer optimized wireless video transmission.' *IEEE Transactions on Multimedia*, **8**(5), 2006, 1082–1095.

[30] Saul, A., Khan, S., Auer, G., Kellerer, W., and Steinbach, E., 'Cross-layer optimization using model-based parameter exchange.' IEEE International Conference on Communications (ICC 2007), Glasgow, June 24–28, 2007.

[31] Liu, Q., Wang, X., and Giannakis, G.B., 'A cross-layer scheduling algorithm with QoS support in wireless networks.' *IEEE Transactions on Vehicular Technology*, **55**(3), 2007, 839–847.

[32] Martini, M.G. and Tralli, V., 'Video quality based adaptive wireless video streaming to multiple users.' IEEE International Symposium on Broadband Multimedia Systems and Broadcasting, 2008.

[33] Merino, P., Martini, M.G., Skorin-Kapov, L., Varela, M., and Schatz, R., 'Guest editorial: Improving QoS and QoE for mobile communications.' *Journal of Computer Networks and Communications*, 2013, 1–2.

[34] Stockhammer, T., 'Dynamic adaptive streaming over HTTP: Standards and design principles.' Second Annual ACM Conference on Multimedia Systems, MMSys'11, San-Jose, California, February 2011.

[35] Liu, C., Bouazizi, I., and Gabbouz, M., 'Rate adaptation for adaptive HTTP adaptive streaming.' Second Annual ACM Conference on Multimedia Systems, MMSys'11, San-Jose, California, February 2011.

[36] Singh, K.D., Hadjadj-Aoul, Y., and Rubino, G., 'Quality of experience estimation for adaptive HTTP/TCP video streaming using H.264/AVC.' Proceedings of the IEEE Consumer Communications and Networking Conference (CCNC), January 2012, pp. 127–131.

[37] Oyman, O. and Singh, S., 'Quality of experience for HTTP adaptive streaming services.' *IEEE Communications Magazine*, **50**(4), 2012, 20–27.

[38] Kim, T. and Ammar, M., 'Receiver buffer requirement for video streaming over TCP.' Proceedings of the 12th Annual ACM International Conference on Multimedia, October 2004, pp. 908–915.

[39] Wei, P., Zixuan, Z., and Chen, C.W., 'Video adaptation proxy for wireless dynamic adaptation streaming over HTTP.' Proceedings of the IEEE 19th International Packet Video Workshop (PV), May 2012, pp. 65–70.

[40] El Essaili, A., Schroeder, D., Staehle, D., Shehada, M., Kellerer, W., and Steinbach, E., 'Quality-of-experience driven adaptive HTTP media delivery.' Proceedings of the IEEE International Conference on Communications (ICC 2013), June 2013.

[41] Renzi, D., Lederer, S., Mattavelli, M., *et al.*, 'Automated QoE evaluation of dynamic adaptive streaming over HTTP.' Proceedings of the Fifth International Workshop on Quality of Multimedia Experience (QoMEX), July 2013.

[42] Mok, R.K.P., Chan, E.W.W., Luo, X., and Chang, R.K.C., 'Inferring the QoE of HTTP video streaming from user-viewing activities.' Proceedings of the First ACM SIGCOMM Workshop on Measurements up the Stack, August 2011, pp. 31–36.

[43] Pahalawatta, P.V., Berry, R., Pappas, T.N., and Katsaggelos, A.K., 'A content-aware scheduling scheme for video streaming to multiple users over wireless networks.' Proceedings of the European Signal Processing Conference, 2006.

[44] Maani, E., Pahalawatta, P.V., Berry, R., and Katsaggelos, A.K., 'Content-aware packet scheduling for multiuser scalable video delivery over wireless networks.' SPIE Optical Engineering and Applications, International Society for Optics and Photonics, 2009, pp. 74,430C–74,430C.

[45] Pahalawatta, P.V. and Katsaggelos, A.K., 'Review of content-aware resource allocation schemes for video streaming over wireless networks.' *Wireless Communications and Mobile Computing*, **7**(2), 2007, 131–142.

[46] Ji, X., Huang, J., Chiang, M., and Catthoor, F., 'Downlink OFDM scheduling and resource allocation for delay constraint SVC streaming.' IEEE International Conference on Communications (ICC), 2008, pp. 2512–2518.

[47] Ji, X., Huang, J., Chiang, M., Lafruit, G., and Catthoor, F., 'Scheduling and resource allocation for SVC streaming over OFDM downlink systems.' *IEEE Transactions on Circuits and Systems for Video Technology*, **19**(10), 2009, 1549–1555.

[48] Omiyi, P.E. and Martini, M.G., 'Cross-layer content/channel aware multi-user scheduling for downlink wireless video streaming.' Proceedings of 5th IEEE International Symposium on Wireless Pervasive Computing, Modena, Italy, May 5–7, 2010, pp. 412–417.

[49] Lu, Z., Wen, X., Zheng, W., Ju, Y., and Ling, D., 'Gradient projection based QoS driven cross-layer scheduling for video applications.' IEEE International Conference on Multimedia and Expo (ICME), 2011, pp. 1–6.

[50] Karachontzitis, S., Dagiuklas, T., and Dounis, L., 'Novel cross-layer scheme for video transmission over LTE-based wireless systems.' IEEE International Conference on Multimedia and Expo (ICME), 2011, pp. 1–6.

[51] Li, P., Chang, Y., Feng, N., and Yang, F., 'A cross-layer algorithm of packet scheduling and resource allocation for multi-user wireless video transmission.' *IEEE Transactions on Consumer Electronics*, **57**(3), 2011, 1128–1134.

[52] Li, F., Liu, G., Xu, J., and He, L., 'Packet scheduling and resource allocation for video transmission over downlink OFDMA networks.' Fourth International Conference on Communications and Networking in China (ChinaCOM), 2009, pp. 1–5.

[53] Khan, N., Martini, M.G., and Bharucha, Z., 'Quality-aware fair downlink scheduling for scalable video transmission over LTE systems.' IEEE SPAWC 2012, Cesme, Turkey, June 2012.

[54] Appuhami, H.D., Martini, M.G., and Hewage, C.T., 'Channel and content aware 3D video scheduling with prioritized queuing.' IEEE Wireless Advanced (WiAd), 2012, pp. 159–163.

[55] Ahmedin, A., Pandit, K., Ghosal, D., and Ghosh, A., 'Exploiting scalable video coding for content aware downlink video delivery over LTE.' International Conference on Distributed Computing and Networking (ICDCN), 2014.

[56] Kang, S.H. and Zakhor, A., 'Packet scheduling algorithm for wireless video streaming.' International Packet Video Workshop, 2002, pp. 1–11.

[57] Turaga, D.S. and van der Schaar, M., 'Cross-layer aware packetization strategies for optimized wireless multimedia transmission.' IEEE International Conference on Image Processing (ICIP), Vol. 1, 2005, pp. I-777–780.

[58] Juan, H.-H., Huang, H.-C., Huang, C., and Chiang, T., 'Scalable video streaming over mobile WiMAX.' IEEE International Symposium on Circuits and Systems (ISCAS), 2007, pp. 3463–3466.

[59] Li, F., Liu, G., and He, L., 'Application-driven cross-layer approaches to video transmission over downlink OFDMA networks.' IEEE GLOBECOM Workshops, 2009, pp. 1–6.

[60] Cosman, P.C., Davidson, H.C., Bergin, C.J., *et al.*, 'Thoracic CT images: Effect of lossy image compression on diagnostic accuracy.' *Radiology*, 1994, 514–524.

[61] Martini, M.G. and Papadopoulos, H. (eds), 'Health and inclusion.' Strategic Applications Agenda, eMobility European Technology Platform, 2009.

[62] Martini, M.G., 'Wireless broadband multimedia health services: Current status and emerging concepts.' IEEE Personal Indoor and Mobile Radio Communications (PIMRC 2008), Cannes, France, September 2008.

[63] Martini, M.G., Istepanian, R.S.H., Mazzotti, M., and Philip, N., 'Robust multi-layer control for enhanced wireless tele-medical video streaming.' *IEEE Transactions on Mobile Computing*, **9**(1), 2010, 5–16.

[64] Istepanian, R.S.H., Philip, N., and Martini, M.G., 'Medical QoS provision based on reinforcement learning in ultrasound streaming over 3.5G wireless systems.' *IEEE Journal on Selected Areas in Communications*, **27**(4), 2010, 566–574.

[65] Hewage, C.T.E.R., Martini, M.G., and Khan, N., '3D medical video transmission over 4G networks.' Fourth International Symposium on Applied Sciences in Biomedical and Communication Technologies, Barcelona, Spain, October 2011, pp. 26–29.

[66] Razaak, M. and Martini, M.G., 'Medical image and video quality assessment in e-health applications and services. IEEE Healthcom 2013 (Workshop on Service Science for e-Health), Lisbon, October 2013.

[67] Martini, M.G., Hewage, C.T.E.R., Nasralla, M.M., Smith, R., Jourdan, I., and Rockall, T., '3-D robotic telesurgery and training over next generation wireless networks.' IEEE EMBC 2013, Osaka, Japan, June 2013.

Further Reading

[68] 3GPP TS 26.234: Transparent end-to-end packet-switched streaming service (PSS); protocols and codecs, 2010.

[69] 3GPP TS 36.213: Physical layer procedures (Release 8). www.3gpp.org.

[70] Ameigeiras, P., Wigard, J., and Mogensen, P., 'Performance of the M-LWDF scheduling algorithm for streaming services in HSDPA.' IEEE Transactions on Vehicular Technology Conference, Vol. 2, September 2004, pp. 999–1003.

[71] Barakovic, S. and Skorin-Kapov, L., 'Survey and challenges of QoE management issues in wireless networks.' *Journal of Computer Networks and Communications*, 2013, 2013, 28.

[72] Basukala, R., Mohd Ramli, H., and Sandrasegaran, K., 'Performance analysis of EXP/PF and M-LWDF in downlink 3GPP LTE system.' Proceedings of IEEE First Asian Himalayas Conference, November 2009.

[73] Choi, J.G. and Bahk, S., 'Cell-throughput analysis of the proportional fair scheduler in the single-cell environment.' *IEEE Transactions on Vehicular Technology*, **56**(2), 2007, 766–778.

[74] Huusko, J., Vehkapera, J., Amon, P., *et al.*, 'Cross-layer architecture for scalable video transmission in wireless network.' *Signal Processing: Image Communication*, **22**(3), 2007, 317–330.

[75] Huynh-Thu, Q. and Ghanbari, M., 'No-reference temporal quality metric for video impaired by frame freezing artefacts.' Image Processing (ICIP), 16th IEEE International Conference, November 2009, pp. 2221–2224.

[76] ISO/IEC 23009-1, Information technology – Dynamic adaptive streaming over HTTP (DASH) – Part 1: Media presentation description and segment formats, 2012.

[77] ITU-T Recommendation G.1070: Opinion model for video-telephony applications, April 2007.

[78] Jarnikov, D. and Ozcelebi, T., 'Client intelligence for adaptive streaming solutions.' IEEE ICME, Singapore, July 2010.

[79] Khan, N., Martini, M.G., and Bharucha, Z., 'Quality-aware fair downlink scheduling for scalable video transmission over LTE systems.' IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), 2012, pp. 334–338.

[80] Lamy-Bergot, C., Martini, M.G., Hammes, P., *et al.*, 'Optimisation of multimedia over wireless IP links via X-layer design.' EUMOB 2008.

[81] Piro, G., Grieco, L.A., Boggia, G., *et al*., 'Simulating LTE cellular systems: An open source framework.' *IEEE Transactions on Vehicular Technology*, October 2010, pp. 1–16.

[82] Piro, G., Grieco, L.A., Boggia, G., Fortuna, R., and Camarda, P., 'Two-level downlink scheduling for real-time multimedia services in LTE networks.' *IEEE Transactions on Multimedia*, **13**(5), 2011, 1052–1065.

[83] Raisinghani, V.T. and Iyer, S., 'Cross-layer feedback architecture for mobile device protocol stacks.' *IEEE Communications Magazine*, January 2006.

[84] Ramli, H., Basukala, R., Sandrasegaran, K., and Patachaianand, R., 'Performance of well-known packet scheduling algorithms in the downlink 3GPP LTE system.' Proceedings of IEEE 9th Malaysia International Conference on Communications, 2009, pp. 815–820.

[85] Sarkar, M. and Sachdeva, H., 'A QoS aware packet scheduling scheme for WiMAX.' Proceedings of IAENG Conference on World Congress on Engineering and Computer Science, Berkeley, CA, 2009.

[86] Sesia, S., Toufic, I., and Baker, M., *LTE – The UMTS Long Term Evolution*. John Wiley & Sons, Chichester, 2009.

[87] Shakkottai, S., Rappaport, T.S., and Karlsson, P.C., 'Cross-layer design for wireless networks.' *IEEE Communication Magazine*, **41**(10), 2003, 74–80.

[88] Singh, K.D. and Rubino, G., 'No-reference quality of experience monitoring in DVB-H networks.' Proceedings of the IEEE Wireless Telecommunications Symposium (WTS), April 2010, pp. 1–6.

[89] Sun, K., Wang, Y., Wang, T., Chen, Z., and Hu, G., 'Joint channel-aware and queue-aware scheduling algorithm for multi-user MIMO-OFDMA systems with downlink beamforming.' IEEE 68th Vehicular Technology Conference (VTC2008-Fall), 2008, pp. 1–5.

[90] Svedman, P., Wilson, S.K., and Ottersten, B., 'A QoS-aware proportional fair scheduler for opportunistic OFDM.' IEEE 60th Vehicular Technology Conference (VTC2004-Fall), Vol. 1, 2004, pp. 558–562.

[91] Technical Specification Group Radio Access Network 3GPP. Physical layer aspect for evolved universal terrestrial radio access (UTRA) (Release 7). Technical report, 3GPP TS 25.814.

[92] Tong, W., Sich, E., Peiying, Z., and Costa, J.M., 'True broadband multimedia experience.' *IEEE Microwave Magazine*, **9**(4), 2008, 64–73.

[93] Wang, B., Kurose, J., Shenoy, P., and Towsley, D., 'Multimedia streaming via TCP: An analytic performance study.' Proceedings of the 12th Annual ACM International Conference on Multimedia, October 2004, pp. 908–915.

[94] Wang, Q. and Abu-Rgheff, M., 'Cross-layer signalling for next-generation wireless systems.' Proceedings of IEEE WCNC 2003.

[95] Yanhui, L., Chunming, W., Changchuan, Y., and Guangxin, Y., 'Downlink scheduling and radio resource allocation in adaptive OFDMA wireless communication systems for user-individual QoS.' Proceedings of the World Academy of Science, Engineering and Technology, Vol. 12, 2006, pp. 221–225.

Acronyms

AMC	Adaptive Modulation and Coding
APP	Application
BLER	Block Error Rate
CDMA	Code Division Multiple Access
CLD	Cross-Layer Design
CQI	Channel Quality Indicator
CSI	Channel State Information
DASH	Dynamic Adaptive Streaming over HTTP
DTQ	Dependency/Temporal/Quality
FR	Full Reference
GoP	Group of Pictures
HoL	Head of Line
LTE	Long-Term Evolution
MAC	Medium Access Control
MOS	Mean Opinion Score
MPEG	Motion Pictures Expert Group
MSE	Mean Square Error
NR	No Reference
OFDMA	Orthogonal Frequency Division Multiple Access
PLR	Packet Loss Ratio
PRB	Physical Resource Block
PSNR	Peak Signal-to-Noise Ratio
QoE	Quality of Experience
QoS	Quality of Service

RR	Reduced Reference
RTCP	RTP Control Protocol
SINR	Signal-to-Interference Noise Ratio
STB	Set-Top Box
SVC	Scalable Video Coding
UE	User Equipment
VoIP	Voice over IP
WiMAX	Worldwide Interoperability for Microwave Access

8

Conclusions

Tasos Dagiuklas[1], Luigi Atzori[2], Chang Wen Chen[3]
and Periklis Chatzimisios[4]
[1]Hellenic Open University, Greece
[2]University of Cagliari, Italy
[3]State University of New York at Buffalo, USA
[4]Alexander Technological Educational Institute, Thessaloniki, Greece

It is a matter of fact that QoE is central in the design, creation, provisioning, and management of current and future multimedia services, especially when these are provided over the Internet, which is the most common case. The time it takes for a video-on-demand service to start playout, the number of stalling events experienced during a video streaming service, the resolution of the pictures shown on a social network site, how different images are combined in a news portal and the type of content they convey, as well as many other system factors, all impact perceived quality. However, not only are system factors important but also those belonging to the human, context, and business domains. Indeed, the mood of the user, the expected quality, and the place where the service is consumed are some elements that should be taken into account when estimating user perception and inferring consumer willingness to keep paying for a service or move to another provider.

Unfortunately for the service providers and fortunately for the researchers working in the field, there is a lot to be done to reach a satisfactory level of understanding on how to perform the estimation and manage it. Even the definition of this novel concept is not agreed upon by the relevant research and industrial communities, as highlighted in Chapter 2. Indeed, QoE is a complex concept that lies at the junction of several, mostly unrelated, scientific, technical, and human disciplines. It is probably for this reason that research in this domain is evolving, and mostly toward the user. This poses conceptual and practical difficulties, but it is a necessary step to take if QoE is to establish itself as a mature field of study. QoE is, after all, all about the

Multimedia Quality of Experience (QoE): Current Status and Future Requirements, First Edition.
Edited by Chang Wen Chen, Periklis Chatzimisios, Tasos Dagiuklas and Luigi Atzori.

user, and the user is the most important element of the end-to-end chains of service creation and provisioning.

In the QoE arena, one of the areas where significant results have already been obtained is the definition of perceptual quality metrics for quality assessment for various types of content, as highlighted in Chapter 3. Indeed, during the past 10 years, some perceptual quality metrics have gained popularity and been used in various signal processing applications, such as the Structural Similarity Measure (SSIM). In the past, a lot of effort has been focused on designing FR metrics for audio or video. It is not easy to obtain good evaluation performance with RR or NR quality metrics. However, effective NR metrics are greatly desired, with more and more multimedia content (such as image, video, or music files) being distributed over the Internet today. The widely used Internet transmission and new compression standards bring many new challenges for multimedia quality evaluation, such as new types of transmission loss and compression distortion. Additionally, various emerging applications of 3D systems and displays require new quality metrics. Depth perception in particular should be investigated further for 3D quality evaluation. Other substantial quality evaluation topics include the quality assessment for super-resolution images/video and High Dynamic Range (HDR) images/video. All these emerging content types and their corresponding processing methods bring about many challenges for multimedia quality evaluation.

In recent years, MPEG-DASH has proved to be one of the most important standardization activities related to multimedia content streaming, whose impact on QoE needs to be carefully investigated. Indeed, the correct configuration of the DASH systems needs to take into account the impact of each packet on end-user perceived quality; additionally, information must be exchanged between the server and the client. This introduces several issues that require further work, as highlighted in Chapter 4: (i) development of evaluation methodologies and performance metrics to accurately assess user QoE for DASH services; (ii) DASH-specific QoS delivery and service adaptation at the network level, involving the development of new policy and charging control guidelines, QoS mapping rules, and resource management techniques over radio access network and core IP network architectures; (iii) QoE/QoS-based adaptation schemes for DASH at the client, network, and server (potentially assisted by QoE feedback reporting from clients), to jointly determine the best video, transport, network, and radio configurations, toward realizing the highest possible service capacity and end-user QoE; (iv) DASH-specific transport optimizations over heterogeneous network environments, where content is delivered over multiple access networks such as WWAN (Wireless Wide Area Networks) unicast. Accordingly, in the future, we expect to see major work in this area.

No-reference or blind image and video quality assessment are reviewed in detail in Chapter 5. The existing algorithms mostly differ in the amount of information available on the types of artifacts that affect the visual data. Even in the face of almost no information regarding the distortion type or human opinion on quality, certain algorithms exist that predict visual quality with a fair degree of accuracy. However, while great progress has been observed in the field of image quality assessment, no-reference video quality assessment has received less attention. Indeed, this is due to the complexity of motion and its interaction with distortion.

Methodologies for assessing the QoE are of fundamental importance and should consider, among other aspects, the following issues. First and foremost, the picture quality perceived; thereafter, the ratings given by human observers are affected by what they have seen or experienced prior to a specific subjective test. Second, it has long been acknowledged that human perception and judgment in a psychophysical measurement task usually perform better in comparison tasks than casting an absolute rating. Third, an issue not altogether disassociated with the previous one is HVS (Human Visual System) response under two distinctive picture quality assessment conditions – that is, where artifacts and distortions are at visibility sub-threshold or around the threshold (usually found in high-quality pictures) and at supra-threshold (commonly associated with medium and low-quality pictures). Finally, it is becoming increasingly clear that assessment of QoE requires more than evaluation of picture quality alone, with the need to differentiate measurement of the perceived resemblance of a picture at hand to the original and that of usefulness of the picture to an intended task. All these issues are discussed in Chapter 6, together with the available approaches and models that consider all or part of these.

One of the major activities in QoE management is dynamic rate control, which can be performed at the application layer in order to allocate the available resources according to user requirements and transmission conditions. Whereas originally this approach was used with the intention of achieving a constant bit rate, now it adapts dynamically the transmission to the available bandwidth as well as to the variable channel and network conditions. Additionally, recent research has proposed a QoE-based control scheme which relies on a cross-layer approach to optimize the overall system configuration according to the quality as perceived by the end-user, as highlighted in Chapter 7. Clearly, this requires constant monitoring of the played content and how it is affected by the experience of the user. For this purpose, only RR and NR metrics can be used, with the former allowing us to achieve better results at the expense of an increase in the information that is sent to the receiver.

Summarizing, this book has addressed several aspects related to QoE: its definition, how to establish appropriate metrics, how the evaluation methodology works, which control and management aspects are to be considered. These represent only some of the key aspects of the subject, which is still in its infancy, according to the editors' opinion. Indeed, QoE is something that is destined to evolve over time, as ICT services are continuously changing – introducing new challenges and benefits for society. Accordingly, much research activity is expected in the years to come, for which we hope this book will form a foundation.

Index

Multimedia Quality of Experience (QoE): Current Status and Future Requirements, First Edition.
Edited by Chang Wen Chen, Periklis Chatzimisios, Tasos Dagiuklas and Luigi Atzori.