

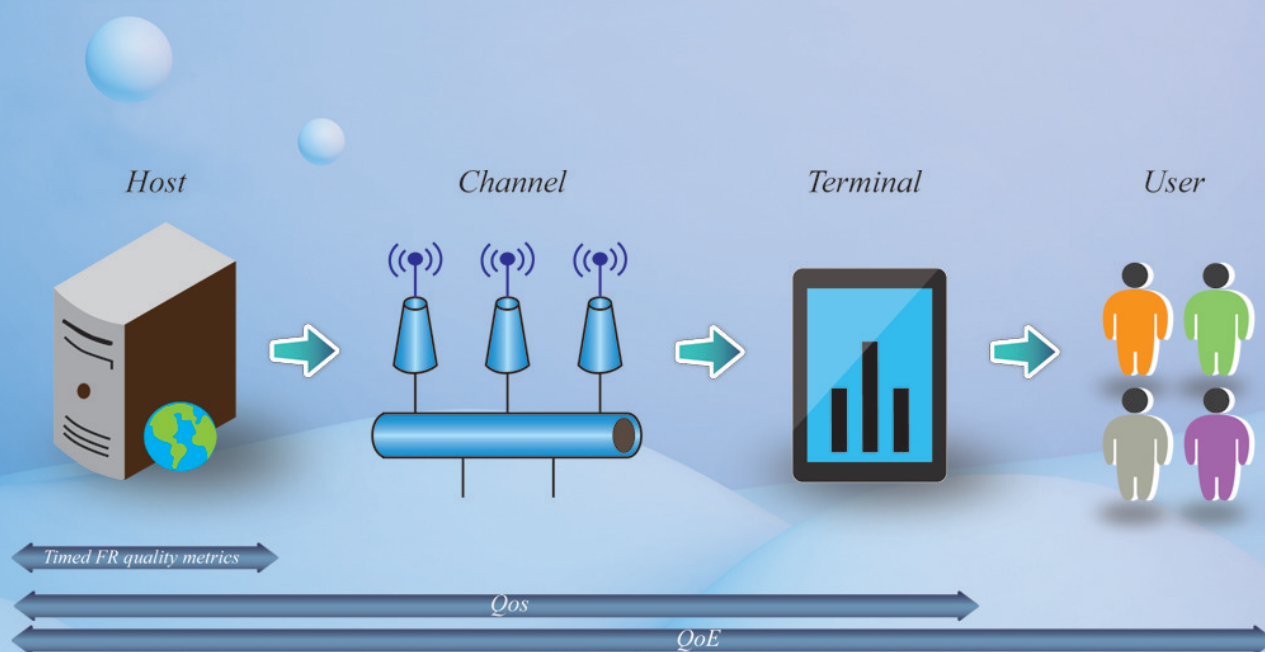
# ZTE COMMUNICATIONS

中兴通讯技术(英文版)

<http://tech.zte.com.cn>

March 2019, Vol. 17 No. 1

## SPECIAL TOPIC: Quality of Experience for Emerging Video Communications



# The 8th Editorial Board of ZTE Communications

## Chairman

**GAO Wen:** Peking University (China)

## Vice Chairmen

**XU Ziyang:** ZTE Corporation (China) | **XU Chengzhong:** University of Macau (China)

### Members (in Alphabetical Order):

**CAO Jiannong**

**Hong Kong Polytechnic University (China)**

**CHEN Changwen**

**University at Buffalo, The State University of New York (USA)**

**CHEN Yan**

**Northwestern University (USA)**

**CHI Nan**

**Fudan University (China)**

**CUI Shuguang**

**University of California, Davis (USA); The Chinese University of Hong Kong, Shenzhen (China)**

**GAO Wen**

**Peking University (China)**

**HWANG Jenq-Neng**

**University of Washington (USA)**

**Victor C. M. Leung**

**The University of British Columbia (Canada)**

**LI Guifang**

**University of Central Florida (USA)**

**LIN Xiaodong**

**ZTE Corporation (China)**

**LIU Jian**

**ZTE Corporation (China)**

**LIU Ming**

**Institute of Microelectronics of the Chinese Academy of Sciences (China)**

**MA Jianhua**

**Hosei University (Japan)**

**PAN Yi**

**Georgia State University (USA)**

**REN Fuji**

**Tokushima University (Japan)**

**SONG Wenzhan**

**University of Georgia (USA)**

**SUN Huifang**

**Mitsubishi Electric Research Laboratories (USA)**

**SUN Zhili**

**University of Surrey (UK)**

**TAO Meixia**

**Shanghai Jiao Tong University (China)**

**WANG Xiang**

**ZTE Corporation (China)**

**WANG Xiaodong**

**Columbia University (USA)**

**WANG Xiyu**

**ZTE Corporation (China)**

**WANG Zhengdao**

**Iowa State University (USA)**

**XU Chengzhong**

**University of Macau (China)**

**XU Ziyang**

**ZTE Corporation (China)**

**YANG Kun**

**University of Essex (UK)**

**YUAN Jinhong**

**University of New South Wales (Australia)**

**ZENG Wenjun**

**Microsoft Research Asia (China)**

**ZHANG Chengqi**

**University of Technology Sydney (Australia)**

**ZHANG Honggang**

**Zhejiang University (China)**

**ZHANG Yueping**

**Nanyang Technological University (Singapore)**

**ZHOU Wanlei**

**Deakin University (Australia)**

**ZHUANG Weihua**

**University of Waterloo (Canada)**

# CONTENTS

ZTE COMMUNICATIONS March 2019 Vol. 17 No. 1 (Issue 65)

## Special Topic

### Quality of Experience for Emerging Video Communications

#### Editorial 01

*CHEN Changwen, ZHAO Tiesong, and CHEN Zhibo*

#### Recent Advances and Challenges in Video Quality Assessment 03

Video quality assessment plays a vital role in the field of video processing and has gained much attention in recent years. This paper gives an up-to-date review of VQA research and highlights the challenges to conduct VQA research. Both subjective study and common VQA databases, as well as various objective VQA methods are reviewed. The authors pointed out several challenges in VQA, including the impact of video content, the memory effects, the computational efficiency, and the personalized video quality prediction.

*LI Dingquan, JIANG Tingting, and JIANG Ming*

#### Quality Assessment and Measurement for Internet Video Streaming 12

In this paper, the authors point out that conventional video quality assessment methods have been designed for broadcasting mode of operations. Emerging Internet-based video services are fundamentally different from broadcasting mode and different assessment strategies must be adopted. Both subjective and objective metrics should be implemented and the measurement may be carried out at client side, server side and in-network to ensure an overall picture of the video service quality.

*ZHANG Xinggong, XIE Lan, and GUO Zongming*

#### 18 Automating QoS and QoE Evaluation of HTTP Adaptive Streaming Systems

For the HTTP streaming systems, the adaptation of video bitrate and possibly even the video resolution makes the assessment of the overall quality much more challenging. This paper presents a flexible and comprehensive framework to conduct objective and subjective evaluations of HAS systems in a fully automated and scalable way. Main features of the proposed approach include end-to-end evaluation of video streaming players deployed in industry, collection and analysis of objective streaming performance metrics, and subjective quality assessment utilizing crowdsourcing for QoE evaluation.

*Christian Timmerer and Anatoliy Zabrovskiy*

#### 25 Quality of Experience Effects in Video Delivery

This paper discusses the quality-of-experience effects in video delivery eco-system from the source, via complex networks, to the destination. The authors investigate the significant differences between QoS and QoE, summarize the end-to-end QoE effects in video delivery, and present their classification based on the deployment. The authors also specifically analyze the impacts of different kinds of factors on QoE in video transmission systems.

*CHEN Jinling, XU Yiwen, LIU Yisang, HUANG Huiwen, and ZHUANG Zhongwen*

Submission of a manuscript implies that the submitted work has not been published before (except as part of a thesis or lecture note or report or in the form of an abstract); that it is not under consideration for publication elsewhere; that its publication has been approved by all co-authors as well as by the authorities at the institute where the work has been carried out; that, if and when the manuscript is accepted for publication, the authors hand over the transferable copyrights of the accepted manuscript to *ZTE Communications*; and that the manuscript or parts thereof will not be published elsewhere in any language without the consent of the copyright holder. Copyrights include, without spatial or timely limitation, the mechanical, electronic and visual reproduction and distribution; electronic storage and retrieval; and all other forms of electronic publication or any other types of publication including all subsidiary rights.

Responsibility for content rests on authors of signed articles and not on the editorial board of *ZTE Communications* or its sponsors.

All rights reserved.

# CONTENTS

ZTE COMMUNICATIONS March 2019 Vol. 17 No. 1 (Issue 65)

## Visual Attention Modeling in Compressed Domain: From Image Saliency Detection to Video Saliency Detection **31**

This paper explores visual attention modeling in the compressed domain for both image and video saliency detection. Saliency regions in compressed image or video play a significant role in the perception of compressed image and video and therefore are closely related to the quality of experience when viewing the received image and video. In particular, this paper introduces fusion strategies to combine spatial and temporal saliency maps to obtain the consistent result for video saliency map.

*FANG Yuming and ZHANG Xiaoqiang*

## Perceptual Quality Assessment of Omnidirectional Images: Subjective Experiment and Objective Model Evaluation **38**

Under the VR environment, providing a good quality of experience is extremely important. In this paper, the authors address the quality assessment of one emerging type of media, omnidirectional images and videos. They first build an omnidirectional image quality assessment database, and then conduct a subjective quality evaluation study in the VR environment based on this database.

Some insightful observations have been obtained through this interesting study.

*DUAN Huiyu, ZHAI Guangtao, MIN Xiongkuo, ZHU Yucheng, FANG Yi, and YANG Xiaokang*

## 48 Quality-of-Experience in Human-in-the-Loop Haptic Communications

With the worldwide rapid development of 5G networks, haptic communications, a key use case of the 5G, has attracted increasing attentions nowadays. Its human-in-the-loop nature makes QoE the leading performance indicator of the system design. A vast number of high quality works were published on user-level, application-level and network-level QoE-oriented designs in haptic communications. In this paper, the authors present an overview of the recent research activities in this progressive research area.

*LIU Qian and ZHAO Tiesong*

## Review

## 56 Comparison Analysis on Feasible Solutions for LTE Based Next-Generation Railway Mobile Communication System

In this paper, the authors introduce the development of railway mobile communications and LTE technology. They also discuss the user requirements of future railway mobile communication system. The two feasible broadband trunking communication solutions for LTE based Next-Generation Railway Mobile Communication System are then analyzed from different aspects.

*SUN Bin, DING Jianwen, LIN Siyu, WANG Wei, CHEN Qiang, and ZHONG Zhangdui*

Serial parameters: CN 34-1294/TN\*2003\*Q\*16\*64\*en\*P\* ¥20.00\*5000\*09\*2019-03

### Statement

This magazine is a free publication for you. If you do not want to receive it in the future, you can send the "TD unsubscribe" mail to magazine@zte.com.cn. We will not send you this magazine again after receiving your email. Thank you for your support.



# Editorial on Special Topic: Quality of Experience for Emerging Video Communications



Guest Editor

**CHEN Changwen** is currently Dean of School of Science and Engineering at the Chinese University of Hong Kong, Shenzhen, China. He also serves as Deputy Director of Peng Cheng Laboratory. He continues to serve as an Empire Innovation Professor of Computer Science and Engineering at the University at Buffalo, State University of New York, USA. He was Allen Henry Endow Chair Professor at the Florida Institute of Technology, USA from July 2003 to December 2007. He was on the faculty of Electrical and Computer Engineering at the University of Rochester, USA from 1992 to 1996 and on the faculty of Electrical and Computer Engineering at the University of Missouri-Columbia, USA from 1996 to 2003.

He was the Editor-in-Chief for *IEEE Trans. Multimedia* from January 2014 to December 2016. He also served as the Editor-in-Chief for *IEEE Trans. Circuits and Systems for Video Technology* from January 2006 to December 2009. He has been an Editor for several other major IEEE Transactions and Journals, including the *Proceedings of IEEE*, *IEEE Journal of Selected Areas in Communications*, and *IEEE Journal of Emerging and Selected Topics in Circuits and Systems*. He has served as Conference Chair for several major IEEE, ACM and SPIE conferences related to multimedia video communications and signal processing. His research has been supported by NSF, DARPA, Air Force, NASA, Whitaker Foundation, Microsoft, Intel, Kodak, Huawei, and Technicolor.

He received his B.S. from University of Science and Technology of China in 1983, M.S.E.E. from University of Southern California, USA in 1986, and Ph.D. from University of Illinois at Urbana-Champaign, USA in 1992. He and his students have received nine 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, Alexander von Humboldt Research Award in 2009, the University at Buffalo Exceptional Scholar—Sustained Achievement Award in 2012, and the State University of New York System Chancellor's Award for Excellence in Scholarship and Creative Activities in 2016. He has been an IEEE Fellow since 2004 and an SPIE Fellow since 2007.



Guest Editor

**ZHAO Tiesong** is currently a Minjiang Distinguished Professor with Fuzhou University, China. He received the B.S. degree in electrical engineering from the University of Science and Technology of China in 2006, and the Ph.D. degree in computer science from the City University of Hong Kong, China in 2011. From 2012 to 2015, he served as postdoc researchers in City University of Hong Kong, University of Waterloo, Canada and the University at Buffalo, State University of New York, USA, respectively. He has joined Fuzhou University since 2015. In 2017, he received the Fujian Science & Technology Award for Young Scientists. Since 2019, he

has also been served as the AE of *IET Electronics Letters*. His research interests include multimedia signal processing, coding and transmission.



Guest Editor

**CHEN Zhibo** received the B.Sc. and Ph.D. degrees from Department of Electrical Engineering, Tsinghua University, China in 1998 and 2003, respectively. He is now a professor in University of Science and Technology of China. Before that he worked with SONY and Thomson from 2003 to 2012. He was a principal scientist and research manager at Thomson Research & Innovation Department. His research interests include image and video compression, visual quality of experience assessment, immersive media computing, and intelligent media computing. He has more than 50 granted and over 100 filed EU and US patent applications, more than 80 publications. He is an IEEE senior member, member of IEEE Visual Signal Processing and Communications Committee, and member of IEEE Multimedia Communication Committee. He was an organization committee member of ICIP 2017 and ICME 2013, and served as a TPC member in IEEE ISCAS and IEEE VCIP.

As 5G mobile communication is making its powerful progress towards full deployment in the near future, we have witnessed tremendous growth of smart mobile devices capable of various video streaming and sharing services. Mobile video services account for majority of the current Internet and wireless data services. Unlike other type of data services, the quality of video service is primarily governed by the end users who are watching videos on the receiving display terminals. The perception and experience of the end users should be the true criteria to assess the quality of the video services. For emerging video communication services, it is the quality of experience, or QoE in short, of the users that should be the most important measure for systematic design for next generation mobile communications.

To examine the state-of-the-art QoE for video communication and networking, we invited a distinguished group of researchers worldwide to present their most recent researches in this special issue. A wide range of topics related to QoE for vid-

eo communications, from fundamental techniques in video quality assessment, to quality assessment and measurement strategies, to automating quality of service (QoS) and QoE evaluations, to QoE issues related to visual attention modeling, omnidirectional video, and haptic communications, have all been explored in this special issue. We hope such diverse topics related to QoE for video communications can bring the readers some fresh perspectives about how important the issue of QoE is and how the video communication users are best served with enhanced QoE through innovative design.

This special issue begins with the paper entitled "Recent Advances and Challenges in Video Quality Assessment." This paper gives an up-to-date review of video quality assessment (VQA) research and highlights the challenges to conduct VQA research. Both subjective study and common VQA databases, as well as various objective VQA methods are reviewed. The authors pointed out several challenges in VQA, including the impact of video content, the memory effects, the computational efficiency, and the personalized video quality prediction.

The second paper is entitled "Quality Assessment and Measurement for Internet Video Streaming." The authors point out that conventional video quality assessment methods have been



designed for broadcasting mode of operations. Emerging Internet-based video services are fundamentally different from broadcasting mode and different assessment strategies must be adopted. Both subjective and objective metrics should be implemented and the measurement may be carried out at client side, server side and in-network to ensure an overall picture of the video service quality.

The third paper entitled “Automating QoS and QoE Evaluation of HTTP Adaptive Streaming Systems” presents a novel strategy of automating QoS and QoE evaluations for the emerging HTTP video streaming systems. For the HTTP streaming systems, the adaptation of video bitrate and possibly even the video resolution makes the assessment of the overall quality much more challenging. This paper presents a flexible and comprehensive framework to conduct objective and subjective evaluations of HAS systems in a fully automated and scalable way. Main features of the proposed approach include end-to-end evaluation of video streaming players deployed in industry, collection and analysis of objective streaming performance metrics, and subjective quality assessment utilizing crowdsourcing for QoE evaluation.

The next paper entitled “Quality of Experience Effects in Video Delivery” discusses the quality-of-experience effects in video delivery eco-system from the source, via complex networks, to the destination. One interesting aspect of this paper is its report on the investigation of the significant differences between the conventional QoS and QoE. Based on the investigation, end-to-end QoE effects have been studied and main conclusions are summarized. In particular, this paper presents the analysis of different types of impacting factors on the overall QoE of the current video communication systems.

The next three papers address service quality issues from different perspectives and for different applications. The first paper in this group is entitled “Visual Attention Modeling in Compressed Domain: From Image Saliency Detection to Video Saliency Detection.” This paper explores the visual attention modeling in the compressed domain for both image and video

saliency detection. Saliency regions in compressed image or video play a significant role in the perception of compressed image and video and therefore are closely related to the quality of experience when viewing the received image and video. In particular, this paper introduces fusion strategies to combine spatial and temporal saliency maps to obtain the consistent result for video saliency map.

The second paper in this group entitled “Perceptual Quality Assessment of Omnidirectional Images: Subjective Experiment and Objective Model Evaluation” addresses the quality assessment of one emerging type of media, omnidirectional images and videos. This new class of media provides immersive experience of real-world scenes in virtual reality environments and special evaluation strategies are very much needed. The authors have established the first database of omnidirectional images for the study how such a new type of media data is different from conventional image quality assessment. Some insightful observations have been obtained through this interesting study.

Finally, we present a paper entitled “Quality-of-Experience in Human-in-the-Loop Haptic Communications” that addresses futuristic media application in haptic communication, one of the key use scenarios for 5G. One unique feature of this type of media application is its human-in-the-loop nature which makes the QoE more important than other 5G use scenarios. The QoE for haptic communications can be observed at user level, or at application level, and even at network level. This paper not only provides comprehensive review of the state-of-the-art QoE management strategies in haptic communications, but also shows technical challenges and research opportunities for seamless haptic communications in the future.

The Guest Editors would like to thank the Editorial Office of *ZTE Communications* for their continuous support throughout the submission and review process. The Guest Editors would also like to thank all the authors for accepting our invitation to contribute to this special issue and to the reviewers for their timely and professional review of these papers.



# Recent Advances and Challenges in Video Quality Assessment

LI Dingquan, JIANG Tingting, and JIANG Ming

(Peking University, Beijing 100871, China)

**Abstract:** Video quality assessment (VQA) plays a vital role in the field of video processing, including areas of video acquisition, video filtering in retrieval, video compression, video restoration, and video enhancement. Since VQA has gained much attention in recent years, this paper gives an up-to-date review of VQA research and highlights current challenges in this field. The subjective study and common VQA databases are first reviewed. Then, a survey on the objective VQA methods, including full-reference, reduced-reference, and no-reference VQA, is reported. Last but not most importantly, the key limitations of current research and several challenges in the field of VQA are discussed, which include the impact of video content, memory effects, computational efficiency, personalized video quality prediction, and quality assessment of newly emerged videos.

**Keywords:** databases; perceptual optimization; personalization; video content; VQA

DOI: 10.12142/ZTECOM.201901002

<http://kns.cnki.net/kcms/detail/34.1294.TN.20190305.1714.002.html>, published online March 5, 2019

Manuscript received: 2018-06-09

## 1 Introduction

In recent years, video-based applications (e.g., video conferencing, video surveillance, and digital television) are growing rapidly in all walks of life. Especially, with the evolution of network and video technologies, people can capture videos to record their daily life with portable mobile devices wherever and whenever they like, and share the interesting ones with other people through social networking services. There is no doubt that video traffic has been the largest part of Internet traffic. However, videos pass through several processing stages before they finally reach the end users of the videos (typically the human consumers). Most of these stages impair the perceived video quality, while some of them try to improve the perceived video quality. Therefore, to provide a satisfying end-user experience, video quality as-

essment (VQA) is a crucial step in many video-based applications. VQA has many practical applications, including quality monitoring in real time; performance evaluation of video systems for video acquisition, compression, transmission, enhancement, display, and so on; and perceptual optimization of video systems.

VQA can be achieved by subjective VQA or objective VQA. The most reliable way to assess the perceived video quality is subjective VQA, which asks the subjects to rate the perceived video quality and processes the subjective ratings to obtain the overall video quality score. However, it is hard to carry out the subjective study in real-time video-based applications, since the subjective experiments are inconvenient, expensive, and inefficient. To automatically predict perceived video quality in real-time video-based applications, we need efficient and effective objective VQA methods.

Subjective VQA is still necessary since we need to benchmark the objective VQA methods with the “ground truth” provided by the subjective VQA, although it has so many drawbacks. Many researchers throw themselves into subjective VQA to construct benchmarking databases. In short, one constructs a video database that can reflect the variety of video

This work was partially supported by National Basic Research Program of China (“973” Program) (2015CB351803), the National Natural Science Foundation of China (61390514, 61527804, 61572042, 61520106004), Sino-German Center (GZ 1025). We also acknowledge the High-Performance Computing Platform of Peking University for providing computational resources.

content and distortions in the considered applications, and the conducted subjective study enables the constructed video database to be a benchmarking VQA database.

Developing objective VQA methods that correlate well with subjective VQA is the main goal of VQA research. According to the availability of reference videos, objective VQA methods include three types: full-reference VQA (FR-VQA), reduced-reference VQA (RR-VQA), and no-reference VQA (NR-VQA). FR-VQA methods, such as motion-based video integrity evaluation (MOVIE) index [1], require the distorted video and the corresponding pristine reference video as well. The complete access to the reference video accounts for the excellent performance of FR-VQA methods since FR-VQA can be seen as the fidelity measure. RR-VQA methods, such as spatio-temporal reduced reference entropic differences (ST - RRED) [2], lie somewhere between FR-VQA and NR-VQA, and only need partial information of the reference video in addition to the distorted one. Compared to FR-VQA, RR-VQA can achieve a good tradeoff between bandwidth occupation and superb performance. NR-VQA methods, such as video intrinsic integrity and distortion evaluation oracle (VVIDEO) [3], predict the perceived video quality without any access to the reference video. Since the reference videos are unavailable in most practical applications, NR-VQA is preferable but also more challenging.

The research field of VQA is in a rapid growth, with the fact that more and more works on new VQA methods, extensions of existing ones, and applications of these VQA methods to other disciplines are put forward every year. The goal of this paper is to provide an up-to-date review of the recent advances of VQA research as a complement to the previous reviews in [4] and [5], and more importantly to highlight the current challenges in this research field. Based on the overview of recent VQA methods, we discuss key limitations of the current VQA research and highlight some challenges in the field of VQA research that we are facing nowadays, including but not limited to the impact of video content, the memory effects and long-term dependencies, the computational efficiency and memory efficiency, the personalized video quality prediction, and the quality assessment of newly emerged videos (e.g., high dynamic range (HDR) panoramic videos) as well as quality assessment guided perceptual optimization of video systems.

This paper is organized as follows. A briefly review on the subjective VQA and public benchmarking VQA databases is presented in Section 2. Section 3 reviews the recent proposed objective VQA methods including FR-VQA, RR-VQA and NR-VQA methods. The key limitations of current VQA research and the challenges in developing effective and efficient VQA methods are discussed in Section 4. Finally, we have a concluding remark in Section 5.

## 2 Subjective Study and VQA Databases

Subjective video quality, i.e., the video quality perceived by

humans, is the most accurate estimation of video quality since humans are the ultimate video receivers. To collect the subjective video quality scores, one must first construct a video database that can reflect the “real distribution” of videos in the application, ensuring the content diversity and distortion (level and type) diversity. Then he can select a suitable method to conduct the subjective study on the database.

The ITU [6] provides the standard settings for the subjective study of video quality. There are many subjective study methods to collect the subjective ratings, including the single-stimulus (SS) and absolute category rating (ACR) method; ACR with hidden reference (ACR - HR); double stimulus impairment scale (DSIS); double stimulus continuous quality scale (DSCQS); pair comparison (PC); subjective assessment of multimedia video quality (SAMVIQ); single stimulus continuous quality evaluation (SSCQE). PC can provide more reliable subjective quality. However, in terms of the number of videos  $n$ , its time complexity is  $O(n^2)$ , while the complexity of other methods is only  $O(n)$ . So some researchers have devoted themselves to improve the PC method by HodgeRank on random graphs [7], active sampling [8], etc.

**Table 1** summarizes some common VQA databases [9]–[18] with the information about the number of reference/distorted videos, distortion types, score types, and the chosen subjective study methods. More VQA databases can be found in a collection of image and video resources on the Winkler’s website [19]. The distorted videos in the first six VQA databases are all obtained by applying compression and transmission errors to the reference videos, and we refer the distortions in these videos as simulated distortions, since we can reproduce exactly the same distorted videos. However, the last four VQA databases contain no reference videos, and the distorted videos in them are authentically distorted, by which we mean that we cannot easily reproduce the same distorted videos. Actually, the simulated distortions are induced by post-processing, while the authentic ones are already induced during the video capture process. The traditional VQA databases have been analyzed in previous literatures, such as [20]. Here, we give more information about the last four VQA databases that include authentic distortions.

Camera Video Database (CVD2014) [15] includes complex authentic distortions induced during the video acquisition process. It contains 234 videos of resolution  $640 \times 480$  or  $1280 \times 720$  recorded by 78 different cameras. In addition to the video quality, the conductors also ask the subjects to give ratings about sharpness, graininess, color balance, darkness, and jerkiness. One should know that, unlike previous databases, CVD2014 enables the audios in the videos. The database provides the raw subjective ratings, which means all the ratings from different subjects are available. The realigned MOS ranges from  $-6.50$  to  $93.38$ .

LIVE-Qualcomm Subjective Mobile In-Capture Video Quality Database [16] aims at authentic, in-capture video distortions



**▼Table 1. VQA databases with the subjective study methods, numbers of (#) reference/distorted videos and score types**

VQA Database	Subjective Study Method	#Reference/Distorted Videos	Score Type**
VQEG FR-TV Phase I [9]	DSCQS	22/352	DMOS+ $\sigma$
VQEG HDTV [10]	ACR-HR	49/740	Raw
EPFL-PoliMI [11]	SS/ACR	12/156	Raw
LIVE [12]	ACR-HR	10/150	DMOS+ $\sigma$
LIVE Mobile [13]	SSCQE-HR	10/200	DMOS+ $\sigma$
CSIQ [14]	SAMVIQ	12/216	DMOS+ $\sigma$
CVD2014 [15]	SS/ACR	None/234	Raw
LIVE-Qualcomm [16]	SS/ACR	None/208	MOS+ $\sigma$
KoNViD-1k* [17]	SS/ACR	None/1 200	Raw
LIVE-VQC* [18]	SS/ACR	None/585	MOS+ $\sigma$

\*The subjective study of KoNViD-1k and LIVE-VQC is conducted on the crowdsourcing platform.

\*\* $\sigma$  indicates the standard deviation of subjective rating and “raw” means that all subjective data are available.

ACR: absolute category rating  
ACR-HR: ACR with hidden reference  
CSIQ: Computational and Subjective Image Quality  
CVD2014: Camera Video Database  
DMOS: difference mean opinion score  
DSCQS: double stimulus continuous quality scale  
EPFL-PoliMI: École Polytechnique Fédérale de Lausanne and Politecnico di Milano  
FR-TV: full-reference television  
HDTV: high definition television  
KoNViD-1k: Konstanz Natural Video Database  
LIVE: Laboratory for Image & Video Engineering  
LIVE-VQC: LIVE Video Quality Challenge Database  
MOS: mean opinion score  
SAMVIQ: subjective assessment of multimedia video quality  
SS: single stimulus  
SSCQE: single stimulus continuous quality evaluation  
VQA: video quality assessment  
VQEG: Video Quality Experts Group

since the simulated distortions in previous databases cannot reflect these in-capture distortions. It consists of 208 videos of resolution  $1\,920 \times 1\,080$  captured by eight different smart-phones and models six in-capture distortions (artifacts, color, exposure, focus, sharpness, and stabilization). The subjective study is carried out on 39 subjects, and the realigned MOS ranges from 16.5621 to 73.6428.

Konstanz Natural Video Database (KoNViD-1k) [17] focuses on authentic distortions “in the wild”. To guarantee the video content diversity, it comprises a total of 1200 videos of resolution  $960 \times 540$  that are fairly sampled from a large public video dataset, YFCC100M. In terms of the video content diversity, KoNViD-1k is now the largest VQA database in the community. The large scale subjective study is not suitable to be conducted in the laboratory environments, so the crowdsourcing platform is chosen. KoNViD-1k also provides the raw data of the subjective study, and the MOS ranges from 1.22 to 4.64.

LIVE Video Quality Challenge Database (LIVE-VQC) [18] is another VQA database including authentic distortions “in the wild”. Same as KoNViD-1k, the large-scale study of LIVE-VQC is also conducted on the crowdsourcing platform. The subjective study has 4 776 unique participants, yielding more than 205 000 opinion scores on the 585 videos.

### 3 Objective Video Quality Assessment

In 2011, FR-VQA and RR-VQA methods were classified

and reviewed [4], while Shahid et al. [5] classified and reviewed NR-VQA methods three years later. The research of VQA is in a rapid growth, and it has gained more attention in recent years. There have been a lot of newly proposed VQA methods since the two review articles published, thus an up-to-date review of the recent progress in VQA research is needed. Here, we give an overview of the recent advances of FR-VQA, RR-VQA, and NR-VQA methods in the following three subsections.

#### 3.1 Full-Reference Video Quality Assessment

The research of FR-VQA methods has a long history. Since the FR-VQA methods have full access to the reference information, they can usually achieve an acceptable performance. Structural information is proved to be essential for image quality assessment (IQA), so it should be also useful for VQA. Different from images, videos have one more dimension over the time axis. So motion information should also be crucial for VQA. Furthermore, to develop an FR-VQA method that correlates well with human perception, investigating the knowledge of human visual system (HVS) is very helpful. We roughly classify the FR-VQA methods into three categories, i.e., structural information guided methods, motion information tuned methods, and HVS inspired perceptual hybrid methods.

(1) Structural information guided methods: Due to the success of structural similarity (SSIM) [21] in the field of IQA, some works in the field of VQA exploit the structural information. The most direct work that extends SSIM to video domain is proposed in [22]. Wang and Li [23] consider the frame-wise SSIM with motion associated weighting, where the motion information is obtained from a statistical model of human visual speed perception. With a novel concept of motion vector reuse, Moorthy and Bovik propose an efficient FR-VQA method, called the motion compensated SSIM (MC-SSIM) [24]. In [25], hysteresis effect is found in the subjective study, so temporal hysteresis pooling is applied to frame-wise SSIM, which is proved to be better than simply taking an average. Wang et al. extract structural information from local spatial-temporal regions [26]. More specifically, the structural information in the local space-time region is represented by the largest eigenvalue and its corresponding eigenvector of the 3D structure tensor. Besides luminance, contrast, structure similarity, Xu et al. consider the space-temporal texture by a rotation sensitive 3D texture pattern [27]. Motivated by the contrast effect, they refine the frame quality score based on the score of the previous frame. In [28], Park et al. propose a video quality pooling method to pool the frame-wise SSIM scores, which emphasizes the “worst” scores in the space-time regions.

(2) Motion information tuned methods: Motion information is very important in the videos, and this encourages developing VQA methods that utilize motion information. Seshadrinathan and Bovik put forward the MOVIE index, an FR-VQA method that considers motion perception [1]. It captures spatial distortions by spatial MOVIE maps and temporal distortions by temporal MOVIE maps, where the temporal MOVIE index is calculated with the guide of additional motion vector information. Vu et al. extend the most apparent distortion (MAD) index [29] to the video domain by taking into account of human perception on motion distortions, resulting the spatial-temporal MAD (ST-MAD) method for FR-VQA [30]. Finding that distortions can affect local optical flow statistics, Manasa and Channappayya measure the amount of distortions by the deviations of these statistics from the pristine optical flow statistics [31]. Yan and Mou [33] decompose the spatiotemporal slice images into simple motion areas and complex motion areas, and then use gradient magnitude standard deviation (GMSD) [32] to estimate the distortions in these two parts.

(3) HVS inspired perceptual hybrid methods: The goal of objective VQA is to predict video quality that correlates well with human perception, so HVS mechanism can inspire new ideas on VQA. Aydin et al. [34] propose an FR-VQA method that considers luminance adaptation, spatiotemporal contrast sensitivity and visual masking. Taking distortion detection and visual masking effects into account, Zhang and Bull [35] exploit noticeable distortion and blurring artifacts, and predict video quality by adaptively combining these two terms through a non-linear model. Visual attention is also an important part of HVS, so some works have tried to investigate the impact of visual saliency or its implications in the field of VQA [36]–[38]. Based on the fact that HVS has the property of energy compaction representation, He et al. [39] propose an FR-VQA method by transforming the videos into the 3D discrete cosine transform (3D-DCT) domain and exploiting the energy and frequency distribution with statistical models. In [40] and [41], several perceptual-related features and methods are combined to boost the performance. Recently, in [42], video multi-method assessment fusion (VMAF) [41] is extended to embedding effective temporal features, and the resulting two methods, called spatiotemporal VMAF (ST-VMAF) and ensemble VMAF (E-VMAF), show further improvement over the VMAF method.

### 3.2 Reduced-Reference Video Quality Assessment

Although FR-VQA methods have the most promising performance, they have limited applications since the original videos are usually unavailable in many real-world video applications. On the other hand, NR-VQA is an extremely difficult task since it does not have access to the reference information at all. These call for a tradeoff between FR-VQA and NR-VQA tasks, and RR-VQA aims to provide this compromise.

The goal of RR-VQA methods is to reduce the issue of high bandwidth occupation in FR-VQA with minor sacrifice of per-

formance. Video quality model (VQM) is an RR-VQA method that first calibrates the reference video and the distorted video then extract low-bandwidth spatial and temporal features to predict video quality [43]. It only requires reference data of around 4% of the size of the uncompressed video sequence, which makes it possible to perform real-time in-service quality measurements. For video quality monitoring applications, Masry et al. [44] exploit the multichannel decomposition of videos using wavelet transform with a coefficient selection mechanism that allows to adjust the bitrate of the reference video decomposition. The reference bitrates can be as low as 10 kbit/s while the proposed method keeps a good performance [44]. Gunawan and Ghanbari [45] propose an RR-VQA for compressed videos based on harmonics gain and loss information created by a discriminative analysis of harmonic strength computed from edge-detected images. Without explicit motion estimation process, Zeng and Wang [46] directly examine temporal variations of local phase structures for RR-VQA in the complex wavelet transform domain. The resulting method is very easy to be adopted by real-world video communication systems since it has only five features with very low rate of reference information. Based on the analysis of contrast and motion sensitivity characteristics of HVS, Wang et al. [47] propose a spatiotemporal information selection mechanism for RR-VQA to reduce the rate of reference information needed.

It is an issue for RR-VQA how to integrate features over time axis. To predict video quality, Le Callet et al. [48] combine three types of perceptual features (frequency content, temporal content, and blocking effects) using a time-delay neural network. It should be noted that the proposed method requires the subjective scores provided by the SSCQE method. Zhu et al. [49] propose a practical strategy for optimizing feature integration, which includes a linear model for local alignment and a non-linear model for quality calibration.

In recent years, the research of RR-VQA has been considering natural scene statistics (NSS) since distortions can alter the statistical regularities related to scene, i.e., change the NSS, evidenced by IQA methods, e.g. naturalness image quality evaluator (NIQE) [50]. Ma et al. [51] develop an RR-VQA method that exploits spatial information loss with an energy variation descriptor and exploits temporal information loss with temporal characteristics of the inter-frame histogram modeled by a statistical model. Soundararajan and Bovik [2] consider using Gaussian scale mixture model to model the wavelet coefficients of frames and frame differences, and then the measured spatial and temporal information differences between the reference and distorted videos are combined to predict video quality. ST-RRED [2] is shown to have robust performance over a wide range of VQA datasets. To further reduce complexity without sacrificing performance, Bampis et al. [52] propose the spatial efficient entropic differencing for quality assessment (SpEED-QA). Like NIQE but unlike ST-RRED, SpEED-QA applies NSS model in the spatial domain, and calculates local entropic

differencing between reference and distorted videos. Since it does not need to wavelet transform, SpEED-QA is much faster than ST-RRED.

### 3.3 No-Reference Video Quality Assessment

In most practical video applications, the pristine videos are unavailable. For example, during the video capture process, it is incapable to capture “perfect” videos which are totally free of distortions. The additional information of the reference video also leads to high bandwidth occupation during video transmission. Moreover, people can perceive the video quality without a reference video. Therefore, NR-VQA is a more natural and preferable way to assess the perceived video quality. Over the years, numerous efforts have been put into studying distortion-specific NR-VQA methods which make assumptions on the distortion type. These methods focus on estimating the perceived quality of videos with specific distortions, such as H. 264/AVC compression [53], transmission error [54], [55], exposure distortion [56], channel-induced distortion [57], shakiness [58], spatially correlated noise [59], and scaling artifacts [60]. However, less efforts have been put into developing non-distortion-specific NR-VQA methods. This is because non-distortion-specific NR-VQA is more general and challenging since it is unaware of distortion types. With the development and applications of machine learning in the field of VQA, non-distortion-specific NR-VQA has gained much attention in recent years. Here, we give an overview of the recent advances in developing non-distortion-specific NR-VQA methods.

Some works extract frame-wise features and pool them over the time axis to obtain the video-level features for quality assessment. Xu et al. propose an NR-VQA method, called V-CORNIA, which is based on unsupervised feature learning [61]. Spatial features are first extracted in a frame-wise way based on a modification of CORNIA [62] with the max-min pooling strategy. Then, a support vector regression (SVR) model taking these features as inputs is trained for approximating frame quality to GMSD [32]. Hysteresis pooling [25] is finally employed to pool the frame quality over temporal axis. Men et al. use contrast, blurriness [63], colorfulness, spatial information (SI) and temporal information (TI) for quality assessment since they are important attributes that are related to the perceived video quality [64]. The video-level features are represented by the average of these frame-level attributes over temporal axis, and a feature combination model is proposed to map the five attributes to video quality.

Some works further consider the information contained in two adjacent frames, e.g., statistics of frame differences and optical flow. Saad et al. develop an NR-VQA method, known as VBLIINDS, which makes use of three types of features: spatial-temporal features based on a natural video statistics (NVS) model of frame differences in DCT domain, spatial naturalness index using NIQE [50], and motion-related features, i.e., motion coherency and ego-motion [65]. Finally, these features are

mapped to video quality predictions by training an SVR model with linear kernel. Manasa and Channappayya propose an NR-VQA method, named FLOSIM-FR, which is based on the optical flow irregularities induced by distortions [66]. Besides, the intra-patch and inter-patch irregularities are measured in a frame-wise way while the distortion-induced flow randomness and frame irregularities are measured based on consecutive frames. The mapping between the extracted features and the video quality score is also achieved by an SVR model. Unlike the above methods, [3] and [67] are free of both distortion types and subjective ratings, which belong to “opinion-free” methods. Mittal et al. develop an efficient NR-VQA method, named VIIDEO, which is based on quantifying the intrinsic statistical irregularities due to the existence of distortions and examining the inter-subband correlations for quality assessment in the frame-difference domains [3]. By considering internal generative mechanism of HVS, Zhu et al. propose a complete blind VQA method based on spatio-temporal internal generative mechanism (ST-IGM) [67]. This method first decomposes the video content into the predicted part and the uncertain part by applying a spatio-temporal autoregressive prediction model on adjacent frames, then employs an improved NVS model to evaluate the quality of these two parts, and finally combines the two quality scores with a weighted geometric mean.

The other works directly consider cubes of video slices to exploit the spatial, temporal, and spatio-temporal information simultaneously. Li et al. develop an NR-VQA method based on an NVS model in the 3D-DCT domain, where the NVS features of short video clips (of size  $4 \times 4 \times 4$ ) are extracted according to the statistical analysis on basic spectral behavior, NVS shape parameter, energy fluctuation, and distribution variation [68]. These NVS features are then pooled over temporal axis to get the video-level features, and the principal components of the video-level features are fed into a linear SVR model for quality prediction. Li et al. [69] propose shearlet- and CNN-based NR VQA (SACONVA), an NR-VQA method based on 3D shearlet transform and 1D convolutional neural network (CNN). 3D shearlet transform is first employed to extract primary spatio-temporal features for video clips of size  $128 \times 128 \times 128$ . 1D CNN is used for exaggerating and integrating discriminative parts of the primary features, followed by a logistic regression for video quality prediction as well as a softmax classification layer for video distortion classification. Shabeer et al. [70] extract spatio-temporal features by modelling the coefficients of sparse representation of video slices, where the spatio-temporal dictionaries are first constructed by the popular k-singular value decomposition (K-SVD) algorithm.

## 4 Challenges

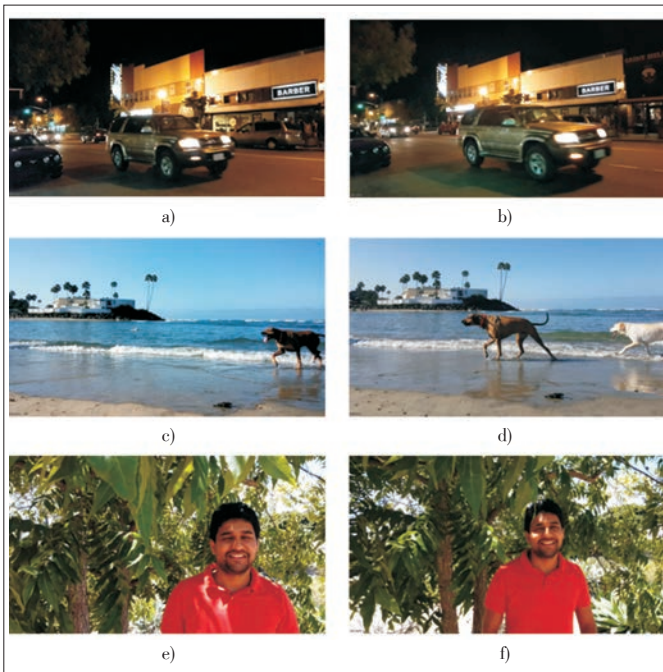
Although the previous two sections show that great progress has been made in the field of VQA research, there still remains some challenges in bridging the gap between human per-

ception and objective VQA. In this section, we discuss several challenging issues, all of which are important aspects for overcoming barriers on the road of developing objective VQA methods that correlate well with human perception.

#### 4.1 Impact of Video Content

The video content diversity has a strong impact on the estimation of perceived video quality since the occurrence probability of distortions and the human tolerance thresholds for distortions vary in different video content/scenes. **Fig. 1** shows an example, where the six videos suffer from almost the same levels of distortion artifacts. However, the two different videos with the same video content (“NightScene”, “DogsOnBeach”, or “ManUnderTree”) have similar perceived quality, while the two videos with different video content have very different perceived quality. Specifically, comparing Figs. 1a/1b to Figs. 1c/1d, we can see that “DogsOnBeach” has significantly higher MOS values than “NightScene”. This is because humans tend to give a higher rating for day scene videos, compared to night scene videos. Comparing Figs. 1e/1f to Figs. 1c/1d, we can see that “DogsOnBeach” has higher MOS values than “ManUnderTree”. This is because humans are more sensitive to distortions occurred in human videos than in landscape videos. The depicted examples support that video content can affect human perception on the perceived quality of distorted videos.

Most of the existing objective VQA methods do not fully



▲ **Figure 1.** The six images are the first frame of six different distorted videos in LIVE-Qualcomm [16]. The distortions in these videos are at similar levels: a) NightScene, mean opinion score (MOS)=35.7252; b) NightScene, MOS=32.2755; c) DogsOnBeach, MOS=65.9423; d) DogsOnBeach, MOS=61.2497; e) ManUnderTree, MOS=52.5666; f) ManUnderTree, MOS=57.2888. It can be seen that the perceived video quality does strongly depend on video content.

take the video content information into account, which may cause the performance decline when the VQA methods are tested on cross-content videos and thus cannot meet the requirements of real-world video applications that contain abundant video content information. In the FR-VQA tasks, the reference video contains the true video content information, therefore, the FR-VQA methods usually have better generalization capability on cross-content videos than the NR-VQA ones. The impact of video content on FR-VQA methods depends on how these methods utilize the reference information. Some works [36]–[38] focus on integrating visual saliency information into the VQA methods. These methods somehow further take the video content into account, since responses of human visual attention rely on “salient” video content and other salient information. NR-VQA tasks do not have the information of reference videos, thus they suffer a lot from the impact of video content. To bridge the gap between FR-VQA and NR-VQA, the first problem to be solved is finding a solution of embedding the video content information into NR-VQA.

Only NR-VQA methods are applicable to quality assessment of videos with authentic distortions. The impact of video content on quality assessment of authentically distorted videos is stronger than quality assessment of simulated distorted videos, which is evidenced by the poor performance of state-of-the-art NR-VQA methods on authentically distorted videos [15]–[18]. To bridge the gap between NR-VQA and human perception, the video content effects must be considered.

#### 4.2 Memory Effects and Long-Term Dependencies

There exist memory effects of subjects during subjective VQA experiments, i.e., the memory of poor quality frames in the past causes subjects to provide lower quality scores for the following frames, even when the frame quality returns to acceptable levels after a time [25]. This is the evidence that long-term dependencies should be considered in the field of VQA. The existing methods consider relationships in limited adjacent frames and cannot handle the long-term dependencies well. From IQA to VQA, it is an open problem on how to deal with the memory effects and long-term dependencies in objective VQA methods.

#### 4.3 Efficiency

Objective VQA methods can be used in real-world video applications only when they are effective and efficient. Most works focus on developing effective methods that have high performance, but less works aim at developing efficient methods which can run fast and even can be deployed in real-time video applications. Even with a C++ implementation, MOVIE [1] spends 11 438 s (more than three hours) for estimating quality of videos in LIVE [12], where the running environment is Windows 7 with 16 GB RAM and a 3.40 GHz Intel Core i7 processor [31]. This computational speed is far from the requirements in real-time applications.



Besides the computational efficiency, the memory efficiency is also a problem. RR-VQA is a way to improve memory efficiency and reduce bandwidth occupation. However, one should also pay attention to improving memory efficiency in the algorithm level if he wants his VQA method to be deployed in memory-limited applications.

#### 4.4 Personalized Video Quality Prediction

MOS, representing the video quality given by the “average user”, is not a suitable representation of video quality, since there is no “average user” in reality [71]. The standard deviation of subjective ratings may be large due to different users’ preferences. Although the quality distribution of subjects can give more information about subjective quality of the video perceived by humans, it still cannot reflect the personalized video quality, which is very important for the next generation of multimedia services. The perceived video quality varies from subjects to subjects. To provide a satisfying quality of experience (QoE) for each user, personalized video quality prediction is required for guiding the user-based video delivery optimization.

The subjective studies conducted in the laboratory environments only include limited number of subjects, which is not suitable for studying the personalized video quality, and it calls for crowdsourcing platforms to collect subjective ratings from various subjects. The subjective studies should collect the user factors of each subject, including physiological factors (e.g., visual acuity and color blindness), socio-cultural factors (e.g., educational and socio-cultural background), demographics (e.g., age, sex and nationality), and psychological factors (e.g., mood and interest). Besides the environments and subjects of subjective studies, the materials, i.e., the constructed video databases, are required to contain enough video content to reflect the real distribution of videos in the applications.

The ultimate goal of personalized video quality prediction is to achieve user-centered optimization and adaptation of video applications. Quantifying the individual differences/preferences and embedding them into the VQA methods to reflect the personalized video quality are challenging but will be desired in the next generation multimedia services.

#### 4.5 Quality Assessment of Newly Emerged Videos and Its Applications

With the development of digital devices and multimedia services, there are many emerging videos. These new videos have some new characteristics, which may raise new challenges for quality assessment research. Stereoscopic 3D video quality assessment needs to further consider depth perception [72]; VQA methods for low/standard dynamic range videos cannot directly be used for HDR videos due to different dynamic ranges [73]; quality assessment of panoramic videos/first-person videos/free-view point videos are needed with the development and popularization of virtual reality technology [74]–[76]; etc. The emerging videos become more and more popular, and they call

for new VQA methods. At the meantime, the progress on quality assessment of these new videos will encourage the development of the new videos themselves. The developed quality assessment methods can also guide the perceptual optimization of video systems, e.g., the video restoration/enhancement/compression systems of both traditional and newly emerged videos. There are a good deal of challenges and opportunities in assessing the quality of newly emerged videos as well as the quality assessment guided perceptual optimization of video systems.

### 5 Conclusions

In this paper, we have reviewed previous works on VQA. Remarkable progress has been made in the past decade, evidenced by a number of state-of-the-art methods (especially the full-reference ones) correlating better with subjective evaluations than traditional PSNR on synthetic distorted videos. However, FR-VQA and RR-VQA methods are not applicable to authentic distorted videos since there is no way to access the reference video, thus we need NR-VQA methods in this case. Existing NR-VQA methods fail to estimate the perceived quality of authentic distorted videos, which is the evidence that the VQA research is far from mature. Then, we discuss the key limitations and five challenges of the current VQA research. We have the following statements. First, good objective VQA methods should not only consider the distortions, but also take the video content information into account. Second, memory effects and long-term dependencies are observed in the subjective studies of VQA databases, and they should be examined in developing objective VQA methods. Third, computational efficiency and memory efficiency are still big issues of quality assessment in real-time video-based applications. Fourth, by accounting for user factors, more practical VQA methods should consider predicting the personalized video quality instead of the “average video quality” of all users. At the meantime, VQA databases should provide raw data that include the user factors of subjects, and the diversity of these databases (including video content diversity and video distortion type and level diversity) should be large enough to reflect the real video distribution in the considered applications. Fifth, it is needed to develop new VQA methods for newly emerged videos (e.g., HDR panoramic videos). We also point out that how to apply the VQA methods in the perceptual optimization of video systems remains many challenges as well as great opportunities.

#### References

- [1] SESHADRINATHAN K, BOVIK A C. Motion Tuned Spatio-Temporal Quality Assessment of Natural Videos [J]. *IEEE Transactions on Image Processing*, 2010, 19(2): 335–350. DOI: 10.1109/tip.2009.2034992
- [2] SOUNDARARAJAN R, BOVIK A C. Video Quality Assessment by Reduced Reference Spatio-Temporal Entropic Differencing [J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2013, 23(4): 684–694. DOI: 10.1109/tcsvt.2012.2214933



- [3] MITTAL A, SAAD M A, BOVIK A C. A Completely Blind Video Integrity Oracle [J]. *IEEE Transactions on Image Processing*, 2016, 25(1): 289–300. DOI: 10.1109/tip.2015.2502725
- [4] CHIKKERUR S, SUNDARAM V, REISSLEIN M, et al. Objective Video Quality Assessment Methods: A Classification, Review, and Performance Comparison [J]. *IEEE Transactions on Broadcasting*, 2011, 57(2): 165–182. DOI: 10.1109/tbc.2011.2104671
- [5] SHAHID M, ROSSHOLM A, LÖVSTRÖM B, et al. No-Reference Image and Video Quality Assessment: A Classification and Review of Recent Approaches [J]. *EURASIP Journal on Image and Video Processing*, 2014. DOI: 10.1186/1687-5281-2014-40
- [6] ITU. Radiocommunication Sector & Telecommunication Standardization Sector [EB/OL]. [2018-05-27]. <https://www.itu.int/en/Pages/default.aspx>
- [7] XU Q Q, HUANG Q M, JIANG T T, et al. HodgeRank on Random Graphs for Subjective Video Quality Assessment [J]. *IEEE Transactions on Multimedia*, 2012, 14(3): 844–857. DOI: 10.1109/tmm.2012.2190924
- [8] FAN Z W, JIANG T T, HUANG T J. Active Sampling Exploiting Reliable Informativeness for Subjective Image Quality Assessment Based on Pairwise Comparison [J]. *IEEE Transactions on Multimedia*, 2017, 19(12): 2720–2735. DOI: 10.1109/tmm.2017.2711860
- [9] VQEG. VQEG FR-TV Phase I Database [EB/OL]. (2000)[2018-05-27]. <http://www.its.bldrdoc.gov/vqeg/projects/frtv-phase-i/frtv-phase-i.aspx>
- [10] VQEG HDTV Group. VQEG HDTV Database [EB/OL]. (2009)[2018-05-27]. <http://www.its.bldrdoc.gov/vqeg/projects/hdtv/hdtv.aspx>
- [11] DE SIMONE F, NACCARI M, TAGLIASACCHI M, et al. Subjective Assessment of H.264/AVC Video Sequences Transmitted over a Noisy Channel [C]// *International Workshop on Quality of Multimedia Experience*. San Diego, USA, 2009: 204–209. DOI: 10.1109/QoMEX.2009.5246952
- [12] SESHADRINATHAN K, SOUNDARARAJAN R, BOVIK A C, et al. Study of Subjective and Objective Quality Assessment of Video [J]. *IEEE Transactions on Image Processing*, 2010, 19(6): 1427–1441. DOI: 10.1109/tip.2010.2042111
- [13] MOORTHY A K, CHOI L K, BOVIK A C, et al. Video Quality Assessment on Mobile Devices: Subjective, Behavioral and Objective Studies [J]. *IEEE Journal of Selected Topics in Signal Processing*, 2012, 6(6): 652–671. DOI: 10.1109/jstsp.2012.2212417
- [14] VU P V, CHANDLER D M. ViS3: An Algorithm for Video Quality Assessment via Analysis of Spatial and Spatiotemporal Slices [J]. *Journal of Electronic Imaging*, 2014, 23(1): 013016. DOI: 10.1117/1.jei.23.1.013016
- [15] NUUTINEN M, VIRTANEN T, VAAHTERANOKSA M, et al. CVD2014: A Database for Evaluating No-Reference Video Quality Assessment Algorithms [J]. *IEEE Transactions on Image Processing*, 2016, 25(7): 3073–3086. DOI: 10.1109/tip.2016.2562513
- [16] GHADIYARAM D, PAN J, BOVIK A C, et al. In-Capture Mobile Video Distortions: A Study of Subjective Behavior and Objective Algorithms [J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2018, 28(9): 2061–2077. DOI: 10.1109/tcsvt.2017.2707479
- [17] HOSU V, HAHN F, JENADELM H, et al. The Konstanz Natural Video Database (KoNViD-1k) [C]// *Ninth International Conference on Quality of Multimedia Experience (QoMEX)*. Erfurt, Germany, 2017: 1–6. DOI: 10.1109/QoMEX.2017.7965673
- [18] SINNO Z, BOVIK A C. Large Scale Study of Perceptual Video Quality [EB/OL]. [2018-03-05][2018-05-27]. <https://arxiv.org/abs/1803.01761>
- [19] WINKLER S. Image and Video Quality Resources [EB/OL]. [2018-05-27]. <http://stefan.winkler.site/resources.html>
- [20] WINKLER S. Analysis of Public Image and Video Databases for Quality Assessment [J]. *IEEE Journal of Selected Topics in Signal Processing*, 2012, 6(6): 616–625. DOI: 10.1109/jstsp.2012.2215007
- [21] WANG Z, BOVIK A C, SHEIKH H R, et al. Image Quality Assessment: From Error Visibility to Structural Similarity [J]. *IEEE Transactions on Image Processing*, 2004, 13(4): 600–612. DOI: 10.1109/tip.2003.819861
- [22] WANG Z, LU L G, BOVIK A C. Video Quality Assessment Based on Structural Distortion Measurement [J]. *Signal Processing: Image Communication*, 2004, 19(2): 121–132. DOI: 10.1016/s0923-5965(03)00076-6
- [23] WANG Z, LI Q. Video Quality Assessment Using a Statistical Model of Human Visual Speed Perception [J]. *Journal of the Optical Society of America A*, 2007, 24(12): B61. DOI: 10.1364/josaa.24.000b61
- [24] MOORTHY A K, BOVIK A C. Efficient Video Quality Assessment along Temporal Trajectories [J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2010, 20(11): 1653–1658. DOI: 10.1109/tcsvt.2010.2087470
- [25] SESHADRINATHAN K, BOVIK A C. Temporal Hysteresis Model of Time Varying Subjective Video Quality [C]// *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Prague, Czech Republic, 2011: 1153–1156. DOI: 10.1109/ICASSP.2011.5946613
- [26] WANG Y, JIANG T T, MA S W, et al. Novel Spatio-Temporal Structural Information Based Video Quality Metric [J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2012, 22(7): 989–998. DOI: 10.1109/tcsvt.2012.2186745
- [27] XU Q Q, WU Z P, SU L, et al. Bridging the Gap between Objective Score and Subjective Preference in Video Quality Assessment [C]// *IEEE International Conference on Multimedia and Expo*. Suntec City, Singapore, 2010: 908–913. DOI: 10.1109/ICME.2010.5583853
- [28] PARK J, SESHADRINATHAN K, LEE S, et al. Video Quality Pooling Adaptive to Perceptual Distortion Severity [J]. *IEEE Transactions on Image Processing*, 2013, 22(2): 610–620. DOI: 10.1109/tip.2012.2219551
- [29] CHANDLER D M. Most Apparent Distortion: Full-Reference Image Quality Assessment and the Role of Strategy [J]. *Journal of Electronic Imaging*, 2010, 19(1): 011006. DOI: 10.1117/1.3267105
- [30] VU P V, VU C T, CHANDLER D M. A Spatiotemporal Most-Apparent-Distortion Model for Video Quality Assessment [C]// *18th IEEE International Conference on Image Processing*. Brussels, Belgium, 2011: 2505–2508. DOI: 10.1109/ICIP.2011.6116171
- [31] MANASA K, CHANNAPPAYYA S S. An Optical Flow-Based Full Reference Video Quality Assessment Algorithm [J]. *IEEE Transactions on Image Processing*, 2016, 25(6): 2480–2492. DOI: 10.1109/tip.2016.2548247
- [32] XUE W F, ZHANG L, MOU X Q, et al. Gradient Magnitude Similarity Deviation: A Highly Efficient Perceptual Image Quality Index [J]. *IEEE Transactions on Image Processing*, 2014, 23(2): 684–695. DOI: 10.1109/tip.2013.2293423
- [33] YAN P, MOU X Q. Video Quality Assessment Based on Motion Structure Partition Similarity of Spatiotemporal Slice Images [J]. *Journal of Electronic Imaging*, 2018, 27(3): 1. DOI: 10.1117/1.jei.27.3.033019
- [34] AYDIN T O, CADÍK M, MYŠKOWSKI K, et al. Video Quality Assessment for Computer Graphics Applications [J]. *ACM Transactions on Graphics*, 2010, 29(6): 1. DOI: 10.1145/1882261.1866187
- [35] ZHANG F, BULL D R. A Perception-Based Hybrid Model for Video Quality Assessment [J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2016, 26(6): 1017–1028. DOI: 10.1109/tcsvt.2015.2428551
- [36] YOU J Y, EBRAHIMI T, PERKIS A. Attention Driven Foveated Video Quality Assessment [J]. *IEEE Transactions on Image Processing*, 2014, 23(1): 200–213. DOI: 10.1109/tip.2013.2287611
- [37] PENG P, LIAO D P, LI Z N. An Efficient Temporal Distortion Measure of Videos Based on Spacetime Texture [J]. *Pattern Recognition*, 2017, 70: 1–11. DOI: 10.1016/j.patcog.2017.04.031
- [38] ZHANG W, LIU H T. Study of Saliency in Objective Video Quality Assessment [J]. *IEEE Transactions on Image Processing*, 2017, 26(3): 1275–1288. DOI: 10.1109/tip.2017.2651410
- [39] HE L H, LU W, JIA C C, et al. Video Quality Assessment by Compact Representation of Energy in 3D-DCT Domain [J]. *Neurocomputing*, 2017, 269: 108–116. DOI: 10.1016/j.neucom.2016.08.143
- [40] FREITAS P G, AKAMINE W Y L, FARIAS M C Q. Using Multiple Spatio-Temporal Features to Estimate Video Quality [J]. *Signal Processing: Image Communication*, 2018, 64: 1–10. DOI: 10.1016/j.image.2018.02.010
- [41] LI Z, AARON A, KATSAVOUNIDIS I, MOORTHY A, MANOHARA M. Toward A Practical Perceptual Video Quality Metric [EB/OL]. (2016-06)[2018-05-27]. <https://medium.com/netflix-techblog/toward-a-practical-perceptual-video-quality-metric-653f208b9652>
- [42] BAMPIS C G, LI Z, BOVIK A C. SpatioTemporal Feature Integration and Model Fusion for Full Reference Video Quality Assessment [J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2018: 1. DOI: 10.1109/tcsvt.2018.2868262
- [43] PINSON M H, WOLF S. A New Standardized Method for Objectively Measuring Video Quality [J]. *IEEE Transactions on Broadcasting*, 2004, 50(3): 312–322. DOI: 10.1109/tbc.2004.834028
- [44] MASRY M, HEMAMI S S, SERMADEVI Y. A Scalable Wavelet-Based Video Distortion Metric and Applications [J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2006, 16(2): 260–273. DOI: 10.1109/tcsvt.2005.861946
- [45] GUNAWAN I P, GHANBARI M. Reduced-Reference Video Quality Assessment Using Discriminative Local Harmonic Strength with Motion Consideration [J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2008, 18(1): 71–83. DOI: 10.1109/tcsvt.2007.913755
- [46] ZENG K, WANG Z. Temporal Motion Smoothness Measurement for Reduced-Reference Video Quality Assessment [C]// *IEEE International Conference on Acoustics, Speech and Signal Processing*. Dallas, USA, 2010: 1010–1013. DOI:

- 10.1109/ICASSP.2010.5495316
- [47] WANG M M, ZHANG F, AGRAFIOTIS D. A very Low Complexity Reduced Reference Video Quality Metric Based on Spatio-Temporal Information Selection [C]/IEEE International Conference on Image Processing (ICIP). Quebec City, Canada, 2015: 571–575. DOI: 10.1109/ICIP.2015.7350863
- [48] LE CALLET P, VIARD-GAUDIN C, BARBA D. A Convolutional Neural Network Approach for Objective Video Quality Assessment [J]. IEEE Transactions on Neural Networks, 2006, 17(5): 1316–1327. DOI: 10.1109/tnn.2006.879766
- [49] ZHU K F, BARKOWSKY M, SHEN M M, et al. Optimizing Feature Pooling and Prediction Models of VQA Algorithms [C]/IEEE International Conference on Image Processing (ICIP). Paris, France, 2014: 541–545. DOI: 10.1109/ICIP.2014.7025108
- [50] MITTAL A, SOUNDARARAJAN R, BOVIK A C. Making a “Completely Blind” Image Quality Analyzer [J]. IEEE Signal Processing Letters, 2013, 20(3): 209–212. DOI: 10.1109/lsp.2012.2227726
- [51] MA L, LI S N, NGAN K N. Reduced-Reference Video Quality Assessment of Compressed Video Sequences [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2012, 22(10): 1441–1456. DOI: 10.1109/tcsvt.2012.2202049
- [52] BAMPIS C G, GUPTA P, SOUNDARARAJAN R, et al. SpEED-QA: Spatial Efficient Entropic Differencing for Image and Video Quality [J]. IEEE Signal Processing Letters, 2017, 24(9): 1333–1337. DOI: 10.1109/lsp.2017.2726542
- [53] ZHU K F, LI C Q, ASARI V, et al. No-Reference Video Quality Assessment Based on Artifact Measurement and Statistical Analysis [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2015, 25(4): 533–546. DOI: 10.1109/tcsvt.2014.2363737
- [54] ZHANG F, LIN W S, CHEN Z B, et al. Additive Log-Logistic Model for Networked Video Quality Assessment [J]. IEEE Transactions on Image Processing, 2013, 22(4): 1536–1547. DOI: 10.1109/tip.2012.2233486
- [55] ZHAO T S, LIU Q, CHEN C W. QoE in Video Transmission: A User Experience-Driven Strategy [J]. IEEE Communications Surveys & Tutorials, 2017, 19(1): 285–302. DOI: 10.1109/comst.2016.2619982
- [56] ROMANIUK P, JANOWSKI L, LESZCZUK M, et al. A no Reference Metric for the Quality Assessment of Videos Affected by Exposure Distortion [C]/IEEE International Conference on Multimedia and Expo. Barcelona, Spain, 2011: 1–6. DOI: 10.1109/ICME.2011.6011903
- [57] VALENZISE G, MAGNI S, TAGLIASACCHI M, et al. No-Reference Pixel Video Quality Monitoring of Channel-Induced Distortion [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2012, 22(4): 605–618. DOI: 10.1109/tcsvt.2011.2171211
- [58] CUI Z X, JIANG T T. No-Reference Video Shakiness Quality Assessment [M]/CUI Z X, JIANG T T. eds. Computer Vision—ACCV 2016. Cham: Springer International Publishing, 2017: 396–411. DOI: 10.1007/978-3-319-54193-8\_25
- [59] CHEN C, IZADI M, KOKARAM A. A Perceptual Quality Metric for Videos Distorted by Spatially Correlated Noise [C]/ACM on Multimedia Conference. Amsterdam, The Netherlands, 2016: 1277–1285. DOI: 10.1145/2964284.2964302
- [60] GHADIYARAM D, CHEN C, INGUVA S, et al. A No-Reference Video Quality Predictor for Compression and Scaling Artifacts [C]/IEEE International Conference on Image Processing (ICIP). Beijing, China, 2017: 3445–3449. DOI: 10.1109/ICIP.2017.8296922
- [61] XU J T, YE P, LIU Y, et al. No-Reference Video Quality Assessment via Feature Learning [C]/IEEE International Conference on Image Processing (ICIP). Paris, France, 2014: 491–495. DOI: 10.1109/ICIP.2014.7025098
- [62] YE P, KUMAR J, KANG L, et al. Unsupervised Feature Learning Framework for No-Reference Image Quality Assessment [C]/IEEE Conference on Computer Vision and Pattern Recognition. Providence, USA, 2012: 1098–1105. DOI: 10.1109/CVPR.2012.6247789
- [63] MEN H, LIN H H, SAUPE D. Empirical Evaluation of No-Reference VQA Methods on a Natural Video Quality Database [C]/Ninth International Conference on Quality of Multimedia Experience (QoMEX). Erfurt, Germany, 2017: 1–3. DOI: 10.1109/QoMEX.2017.7965644
- [64] NARVEKAR N D, KARAM L J. A No-Reference Image Blur Metric Based on the Cumulative Probability of Blur Detection (CPBD) [J]. IEEE Transactions on Image Processing, 2011, 20(9): 2678–2683. DOI: 10.1109/tip.2011.2131660
- [65] SAAD M A, BOVIK A C, CHARRIER C. Blind Prediction of Natural Video Quality [J]. IEEE Transactions on Image Processing, 2014, 23(3): 1352–1365. DOI: 10.1109/tip.2014.2299154
- [66] MANASA K, CHANNAPPAYYA S S. An Optical Flow-Based No-Reference Video Quality Assessment Algorithm [C]/IEEE International Conference on Image Processing (ICIP). Phoenix, USA, 2016: 2400–2404. DOI: 10.1109/ICIP.2016.7532789
- [67] ZHU Y, WANG Y F, SHUAI Y. Blind Video Quality Assessment Based on Spatio-Temporal Internal Generative Mechanism [C]/IEEE International Conference on Image Processing (ICIP). Beijing, China, 2017: 305–309. DOI: 10.1109/ICIP.2017.8296292
- [68] LI X L, GUO Q, LU X Q. Spatiotemporal Statistics for Video Quality Assessment [J]. IEEE Transactions on Image Processing, 2016, 25(7): 3329–3342. DOI: 10.1109/tip.2016.2568752
- [69] LI Y M, PO L M, CHEUNG C H, et al. No-Reference Video Quality Assessment with 3D Shearlet Transform and Convolutional Neural Networks [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2016, 26(6): 1044–1057. DOI: 10.1109/tcsvt.2015.2430711
- [70] SHABEER P M, BHATI S, CHANNAPPAYYA S S. Modeling Sparse Spatio-Temporal Representations for No-Reference Video Quality Assessment [C]/IEEE Global Conference on Signal and Information Processing (GlobalSIP). Montreal, Canada, 2017: 1220–1224. DOI: 10.1109/GlobalSIP.2017.8309155
- [71] ZHU Y, GUNTUKU S C, LIN W, GHINEA G, REDI J A. Measuring Individual Video QoE: A Survey, and Proposal for Future Directions Using Social Media [J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2018, 14(2s): 30. DOI: 10.1145/3183512
- [72] CHEN Z B, ZHOU W, LI W P. Blind Stereoscopic Video Quality Assessment: From Depth Perception to Overall Experience [J]. IEEE Transactions on Image Processing, 2018, 27(2): 721–734. DOI: 10.1109/tip.2017.2766780
- [73] NARWARIA M, PERREIRA DA SILVA M, LE CALLET P. HDR-VQM: An Objective Quality Measure for High Dynamic Range Video [J]. Signal Processing: Image Communication, 2015, 35: 46–60. DOI: 10.1016/j.image.2015.04.009
- [74] ZHANG Y X, WANG Y B, LIU F Y, et al. Subjective Panoramic Video Quality Assessment Database for Coding Applications [J]. IEEE Transactions on Broadcasting, 2018, 64(2): 461–473. DOI: 10.1109/tbc.2018.2811627
- [75] BAI C, REIBMAN A R. Image Quality Assessment in First-Person Videos [J]. Journal of Visual Communication and Image Representation, 2018, 54: 123–132. DOI: 10.1016/j.jvcir.2018.05.005
- [76] LING S Y, LE CALLET P. Image Quality Assessment for Free Viewpoint Video Based on Mid-Level Contours Feature [C]/IEEE International Conference on Multimedia and Expo (ICME). Hong Kong, China, 2017: 79–84. DOI: 10.1109/ICME.2017.8019431

### Biographies

**LI Dingquan** received the double B.S. degrees in electronic science and technology & applied mathematics from Nankai University, China in 2015, and he is currently working toward the Ph.D. degree in applied mathematics at Peking University, China. He is a member of National Engineering Lab for Video Technology. His research interests include image/video quality assessment, perceptual optimization, and machine learning. He has published papers in *IEEE Transactions on Multimedia* and ACM Multimedia Conference.

**JIANG Tingting** (ttjiang@pku.edu.cn) received the B.S. degree in computer science from University of Science and Technology of China in 2001 and the Ph.D. degree in computer science from Duke University, USA in 2007. She is currently an associate professor of computer science at Peking University, China. Her research interests include computer vision and image/video quality assessment. She has published more than 40 papers in journals and conferences.

**JIANG Ming** received the B.Sc. and Ph.D. degrees in mathematics from Peking University, China in 1984 and 1989, respectively. He is a professor with Department of Information Science, School of Mathematical Science, Peking University since 2002. His research interests are mathematical and technical innovations in biomedical imaging and image processing.

# Quality Assessment and Measurement for Internet Video Streaming

ZHANG Xinggong, XIE Lan, and GUO Zongming

(Peking University, Beijing 100871, China)



**Abstract:** Benefiting from the improvements of Internet infrastructure and video coding technology, online video services are becoming a new favorite form of video entertainment. However, most of the existing video quality assessment methods are designed for broadcasting/cable televisions and it is still an open issue how to assess and measure the quality of online video services. In this paper, we survey the state-of-the-art video streaming technologies, and present a framework of quality assessment and measurement for Internet video streaming. This paper introduces several metrics for user's quality of experience (QoE). These QoE metrics are classified into two categories: objective metrics and subjective metrics. It is different for service participators to measure objective and subjective metrics. The QoE measurement methodologies consist of client-side, server-side, and in-network measurement.

**Keywords:** Internet video streaming; QoE; QoE assessment and measurement; HTTP adaptive streaming

DOI: 10.12142/ZTECOM.201901003

<http://kns.cnki.net/kcms/detail/34.1294.tn.20190314.0916.002.html>, published online March 14, 2019

Manuscript received: 2018-09-04

## 1 Introduction

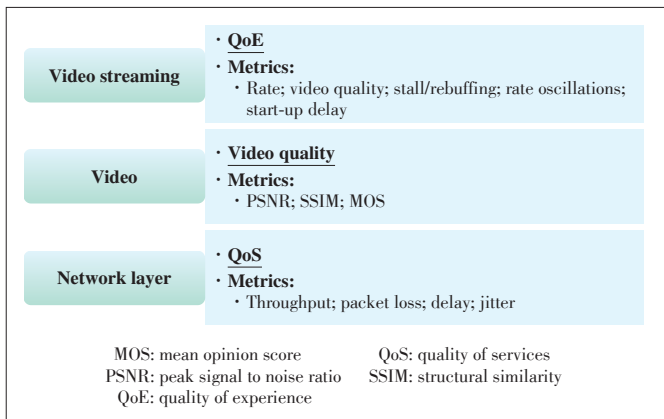
**P**ast few years have witnessed the booms of Internet video services. The Internet unicorns, such as Youku, Tencent, Toutiao from China and YouTube, Amazon, Hulu from US, are becoming main players in the video entertainment market. Mobile phones, over-the-top (OTT) devices, and online streaming are replacing broadcast/cable TVs as new favorable video entertainment for the generation born after 1980. According to the data from China Internet Network Information Center (CNNIC) [1], the total number of online video users in China is about 751 million, more than the population of Australia. At the same time, the users of broadcasting/cable are going steadily downhill. In 2018, the total number of cable TV users was about 295 million, dropping

19% from that of 2017. Online video services have totally changed the status quo of video transmission. Video streaming for delivering and playing multimedia at the same time emerges as one of the main technologies for Internet video transmission.

Although online video services have been widely deployed, they have not been standardized on the assessment and measurement of the quality of services. Unlike broadcast/cable televisions with dedicated infrastructure, online video streaming systems have to compete for network resources over the Internet. They provide services without quality guarantee. The existing methods of quality assessment are mainly designed for legacy broadcast/cable TVs, which is no more applicable to online video services. It is needed to propose a new framework for video streaming quality assessment.

As shown in **Fig. 1**, the quality metrics are different for network layer, video layer and streaming layer. Quality of service (QoS) is defined by ITU [2] to measure the performance of network, not the actual experience of user. The common QoS met-

This work was supported by National Key R&D Program of China No. 2018YFB0803702, Beijing Culture Development Funding under Grant No.2016-288 and Toutiao Funding No. ZN20171224003.



▲ Figure 1. Quality metrics for network, video, and streaming.

rics are throughput, packet losses, delay, jitter, etc. The video quality is assessed by comparing original videos with outcome content, pixel by pixel. The metrics of video quality are mainly designed for video coding or legacy broadcast/cable TVs, such as the peak signal to noise ratio (PSNR), structural similarity (SSIM), and subjective metric of the mean opinion score (MOS).

However, video streaming is an end-to-end video delivery and playback. Its quality depends on video coding and on network conditions as well. Mostly, its quality is assessed by quality-of-experience (QoE), a user-centric metric that measures the performance subjectively perceived by the user.

The QoE of video streaming is influenced by the following factors:

- Video level: video quality (PSNR), frame rate, and resolution
- Network level: start-up delay, bitrate, stall/rebuffering, and rate oscillations
- Application-level: video buffering, browser/player, and screen size.

Due to the various factors that affect the QoE, it is needed to standardize the quality assessment for video streaming. This paper presents a framework of quality assessment and measurement, and introduces it from three perspectives: video streaming technologies, QoE metrics, and measurement methodology. This survey paper tries to present an overall framework of quality assessment and measurement, and provide tools to quantify QoE of Internet video streaming.

The paper is organized as follows. The framework of QoE is illustrated in Section 2. Then, three of the most used video streaming technologies are introduced in Section 3. The subjective and objective QoE metrics are given in Sections 4 and 5. The measurement methods are introduced in Section 6. At last, conclusions are given in Section 7.

## 2 Quality Assessment Framework for Video Streaming

The quality of video streaming is impacted by several factors as video coding, network, and video streaming technolo-

gies. It is needed to capture user's QoE to assess the quality of video streaming. A framework of quality assessment and measurement for video streaming is illustrated in Fig. 2. It mainly consists of three parts: video streaming technologies, quality metrics, and measurement methods.

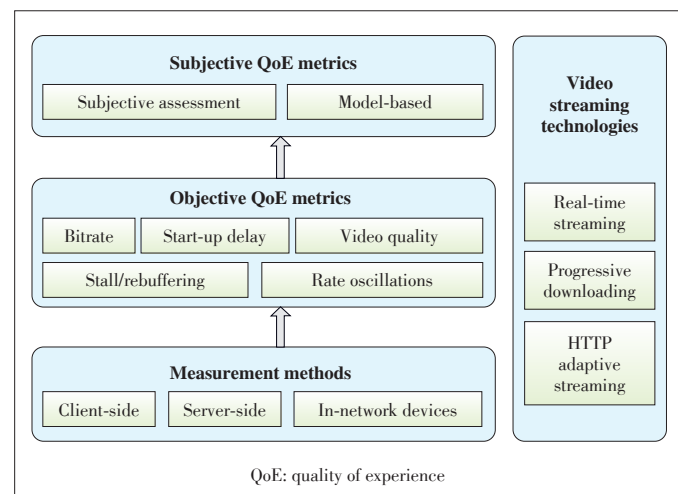
There are various video streaming technologies, which may result in different quality impairments. The streaming technology is one of the most important factors affecting QoE. The widely used streaming technologies includes real-time streaming (RTS), HTTP progressive downloading (HPD), and HTTP adaptive streaming (HAS). All of them are able to enable users to start the playback once the part of the video is downloaded. However, due to their different transmission technologies, their quality impairments are not same. For example, RTS is mainly used in low-latency interactive applications, such as live streaming and video chatting. It is not only sensitive to video quality, but also to round-trip delay. The quality impairments have different impacts on various streaming technologies.

The QoE metrics are used to assess the quality of video streaming. It can be classified into two categories: objective metrics and subjective metrics.

Objective metrics are the QoE metrics which can be quantified with a measurement tool, such as bitrate and delay. These metrics are objective and easy to be measured. However, they have only indirect impacts on users' experience with the service.

Subjective metrics are the direct QoE feedbacks from users. Users rate the video service on a standard measuring way. However, subjective metrics are susceptible to bias because the users' QoE could be varied from one subject to another.

The techniques to measure QoE are also important for video streaming. Video streaming is an end-to-end service. There are multiple parties participating in it, such as content providers, content-distribution-network (CDN) providers, network operators, and users. They view the end-to-end streaming from differ-



▲ Figure 2. Framework of quality assessment and measurement for video streaming.



ent perspectives, thus the measurement methodologies and tools are also different.

### 3 Video Streaming Technologies

Video streaming is a delivery technology, which enables users to playback the video while it is being downloaded. For on-line video services, there are mainly three video streaming technologies: RTS, HPD, and HAS.

#### 3.1 Real-Time Video Streaming

RTS is mainly used for low-latency video applications such as video chat, video conferences, and live video. RTS achieves low latency by a stateful protocol through User Datagram Protocol (UDP) or Transmission Control Protocol (TCP). The streaming server maintains the status of each connection and feeds back the status to clients.

The implementation of RTS depends on public standardization and proprietary protocols. Real-Time Streaming Protocol (RTSP) was developed by RealNetworks, Netscape and Columbia University. It was standardized as the IETF RFC 2326 standard in 1998. It works with Real-Time Transport Protocol (RTP) and Real-Time Control Protocol (RTCP) together to transmit video data. Real-Time Messaging Protocol (RTMP) was initially a proprietary protocol developed by Macromedia (Adobe). It is a stateful protocol which streams audio/video between a Flash Player and Flash Server. RTMP runs on the TCP protocol and supports the parallel transport of video, audio, data, user commands, and control information.

#### 3.2 HTTP Progressive Downloading

In HPD, a video file is downloaded as a regular file using HTTP from a web server. A client can playback the video while the downloading is going on. HPD is a stateless transmission. The server need not maintain session status. The use of HTTP greatly simplifies the traversal of firewalls and proxy server. Current Internet infrastructure and CDN are fully reusable for HPD. Thus, its deployment cost is relatively low.

However, HPD video playback may be interrupted under poor bandwidth or high packet loss situations. This leads to playback re-buffering or stall. Even more, HPD downloads video files at the fastest speed and stores them on the local hard disk, therefore, once the user exits early, the data that has been downloaded but not watched are wasted.

Many websites using Flash Player, such as Youku, use HPD for the streaming. However, in recent

years, more and more websites have given it up for it is not adaptable to bandwidth variation.

#### 3.3 HTTP Adaptive Streaming

HAS technique is proposed to support adaptive streaming over HTTP. An HAS server does not maintain any state information during the streaming. The rate adaptation is done at the client side. This provides scalability with better QoE experience to users. The diagram of HAS is illustrated in Fig. 3.

In HAS, media files are divided into “segments”, which can be encoded into multiple bitrate versions and assigned to a unique URL. Different versions may have different bitrates, resolutions, formats, languages, and other characteristics. An HAS client requests the proper bitrate version to adapt to bandwidth variation.

Many online video services have already supported HAS, such as Netflix, Hulu, and Amazon. Some products, such as Adobe’s HTTP Dynamic Streaming (HDS), Microsoft’s HTTP Smooth Streaming (HSS), and Apple’s HTTP Live Streaming (HLS), also provide HAS functions.

### 4 Objective Quality Assessment

The quality of video streaming can be quantified by some tools and objective metrics. The most used objective QoE metrics are listed as follows.

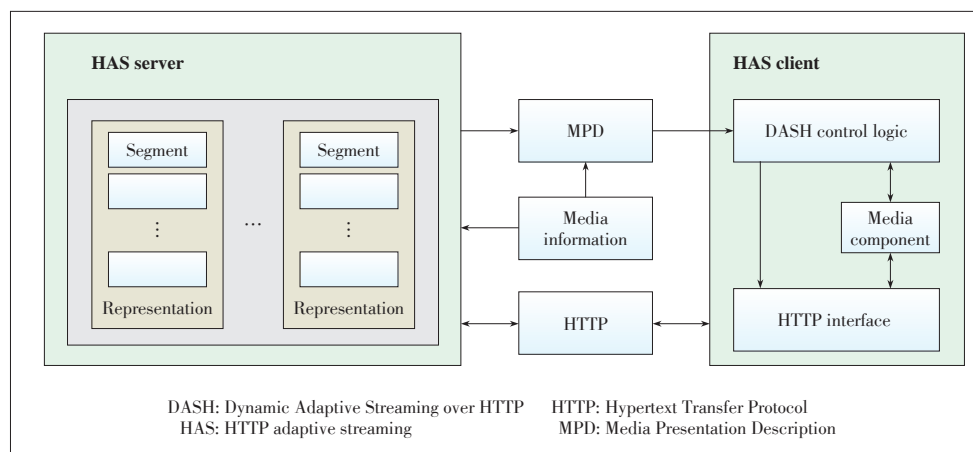
#### 4.1 Video Quality

Video quality in streaming refers to the distortion caused by encoding and transmission compared with the original video. It is often measured with the metrics of PSNR, SSIM, and video quality metric (VQM).

Bitrate is one of the simplest ways to assess the video quality without reference. It is another metric of video quality.

#### 4.2 Start-up Delay

Start-up delay is the time of users’ clicking a video and



▲ Figure 3. Diagram of HTTP adaptive streaming.



waiting before the video starts playing. The start-up delay includes the time of HTML page loading, script loading, video clip buffering, etc.

The start-up delay is an important factor which affects QoE. Some online video services (such as YouTube) tend to initially download data faster and fill the play buffer as soon as possible. In [3], a large-scale user study shows that the start-up delay has a significant impact on a user's online time, and if the start-up delay exceeds two seconds, the user may be stopping watching video.

#### 4.3 Playback Rebuffering or Stall

Stall occurs temporarily when the playback buffer is exhausted. The stall duration is the time that a player waits for the buffer to be filled. In addition, the frequency of stall is also an indicator of video streaming performance.

The rebuffering events during playback will result in a poor user experience. In [3], the authors found that the users with four or more video interruptions were more likely to watch short videos. Also, when the stall duration was more than three seconds, the dissatisfaction increased [4]–[6].

#### 4.4 Bitrate Fluctuation

Frequent bitrate switching will drop users' QoE [7]. Bitrate switching events occur during dynamic adaptive streaming. When network bandwidth deteriorates, a player will reduce the video bitrate and ensure continuous playback. Vice versa, the player increases the bitrate when network becomes better. The bitrate switching can improve bandwidth utilization, but with a bad impact on users' QoE.

### 5 Subjective QoE Assessment

The other way to assess the quality of video streaming is using subjective QoE metrics. Subjective QoE metrics are used to measure the satisfaction of users in video streaming sessions. The subjective assessment methods are divided into two categories: QoE feedback and model-based QoE.

#### 5.1 QoE Feedback

The QoE score is decided by the feedback scores collected from the human subjects based on their experience of video playback. However, the feedback score can be biased across human subjects, since they are different in physical and psychological confounding factors. To obtain an unbiased and general QoE score, the introduction of statistical analysis techniques is necessary.

One of the most popular subjective QoE metrics is Mean Opinion Score (MOS). For getting MOS, limited sets of human subjects are exposed to watch a video under a controlled test-bed and are asked to rate the experience of streaming session. The MOS is a five-point discrete value (Excellent, Good, Fair, Poor, and Bad). And the QoE score is calculated by averaging

the MOS given by the users.

#### 5.2 Model-Based Subjective QoE

Collecting feedbacks from the users is time consuming and has limitation on real practice. Therefore, it is feasible to establish a QoE model to estimate the subjective QoE scores from objective metrics. This is an automatic, quantitative and repeatable manner.

There are two model-based methods: (1) learning-based, which uses learning techniques to map the objective metrics to MOS; (2) heuristic methods, which estimate the subjective QoE scores by some manual functions.

##### 5.2.1 Learning-Based QoE Models

The learning-based QoE model uses machine learning and regression analysis to estimate users' MOS. Meanwhile, some objective metrics such as rebuffering and video quality are recorded as well. The subjective ratings and objective metrics are used to train predictive models to estimate the subjective QoE.

In [8], the authors use Random Neural Network (RNN) to map objective metrics to MOS and train a predictive model. In [9], the authors model the correlations between MOS and objective metrics, including video quality level  $Q_k$ , rebuffering times  $F_{freq}$ , and average rebuffering duration  $F_{avg}$ . They use regression analysis to obtain the weights of each term. Furthermore, Maxim et al. [10] define the influence of the average quality level  $\mu$ , quality variation  $\sigma$  and rebuffering event  $\phi$  on the estimated MOS:

$$\begin{aligned}\mu &= \frac{\sum_{k=1}^K Q_k}{K}, \\ \sigma &= \sqrt{\frac{\sum_{k=1}^K \left( \frac{Q_k}{N} - \mu \right)^2}{K-1}}, \\ \phi &= \frac{7 \times \max \left( \frac{\ln(F_{freq})}{6} + 1, 0 \right) + \left( \frac{\min(F_{avg}, 15)}{15} \right)}{8}, \\ eMOS &= \max(5.67 \times \mu - 6.72 \times \sigma - 4.95 \times \phi + 0.17, 0),\end{aligned}\quad (1)$$

where  $N$  is the number of video bitrate levels and  $K$  is the number of video segments.

##### 5.2.2 Heuristic-Based Predictive Models

Heuristic-based predictive models manually establish relationship between QoE and objective metrics. Yin et al. [11] consider video quality, quality variation, rebuffering, and start-up delay as the objective factors. They define a QoE function

between the objective factors and MOS :

$$QoE = \sum_{k=1}^K q(R_k) - \lambda \sum_{k=1}^{K-1} |q(R_{k+1}) - q(R_k)| - \mu \sum_{k=1}^K \left( \frac{d_k(R_k)}{C_k} - B_k \right) - \mu_s T_s, \quad (2)$$

where  $R_k$  is the bitrate of  $k$ -th segment,  $q(*)$  is the relationship between video bitrate and video quality,  $d_k(R_k)/C_k$  is the download time of  $k$ -th segment,  $B_k$  is the buffer occupancy, and  $T_s$  is the startup delay. Therefore, the estimated MOS is a linear increasing function of the average video quality, and it is a linear decreasing function of the video variation, the rebuffering times, and the startup delay. Besides,  $l$ ,  $m$  and  $ms$  are the weights on the objective factors.

## 6 QoE Measurement Methodologies

Using QoE to represent user satisfaction has been widely recognized by the industry, but there is no unified standard for measuring and obtaining the QoE for online video streaming services. According to the methodology and location of data-collection in the network, we classify QoE measurement methodologies for online video services into the following three categories: client-side, in-network, and server-side measurement.

### 6.1 Client-Side Measurement

There are passive measurement and proactive measurement in the client-side, where some tools are used to measure users' QoE directly.

Passive measurement tools [12], [13] collect the objective QoE metrics when users are watching videos. In this case, the measurement is completely depended on the users and the tools have no control on the video content or duration. Such QoE monitoring tools have been developed for YouTube [13] and Windows Media Player [14] users. By collecting information such as buffer status, TCP rates, and packet loss, they predict QoE metrics like start-up delays and stall times.

Proactive measurements typically use crawlers or bots that crawl through the websites and collect the QoE metrics for a large number of videos. The advantage of using such tools is that they can avoid user participation, thus eliminating any subjective bias. In [15], the authors used a tool called Pytomo to crawl video data on YouTube websites, collecting the network latency, startup-delay, number of stall, and the CDN information.

### 6.2 In-Network Measurement

Measurements of QoE within network [16] do not require modifying client or server software. It just overhears IP packets passing through links, and estimates the QoE of video streaming in the application layer. It is easy for network operators to

deploy these measurement tools.

According to the type of data, in-network measurement can be divided into two categories: TCP layer measurement and HTTP layer measurement.

TCP layer measurement collects on-line or off-line packet information from the TCP layer or lower layer, such as throughput and Round-trip Time (RTT). By tracking the packet-level information of each session, the objective QoE can be estimated, including stall duration, start-up delay, etc.

HTTP layer measurement tracks the performance of HTTP sessions on the application layer. By analyzing the HTTP requests and responses of video data packets, the objective QoE can be obtained.

### 6.3 Server-Side Measurement

Server-side measurements [17] collect each HTTP data packet on server side and rebuild the HTTP session, by which they can obtain information such as rebuffering frequency, start-up delay, stall time, and bitrate switching frequency.

## 7 Conclusions

With the boom of online video services, it has attracted more and more interests from industry and academia how to measure and assess the service quality. This paper presents a whole image of the state-of-the-art quality assessment methods of online video streaming. It introduces three streaming technologies and the corresponding quality assessment methods: subjective quality assessment, objective quality assessment, and quality measurement. There still exists a large gap between the industry requirements and the existing academic works. More studies on the QoE modeling and measurement should be carried out in future.

## References

- [1] China Internet Network Information Center. The 40th CNNIC Statistical Survey Report on Internet Development in China [R]. Aug. 2017
- [2] ITU - T. Vocabulary for Performance and Quality of Service: Recommendation P.10/G.100-Amendment 3 [S]. 2012
- [3] KRISHNAN S S, SITARAMAN R K. Video Stream Quality Impacts Viewer Behavior: Inferring Causality Using Quasi-Experimental Designs [J]. ACM Transactions on Networking, 2013, 21(6): 2001–2014. DOI: 10.1109/tnet.2013.2281542
- [4] HOSSFELD T, EGGER S, SCHATZ R, et al. Initial Delay vs. Interruptions: Between the Devil and the Deep Blue Sea [C]//Fourth International Workshop on Quality of Multimedia Experience, Yarra Valley, Australia, 2012: 1–6. DOI: 10.1109/QoMEX.2012.6263849
- [5] HOSSFELD T, SEUFERT M, HIRTH M, et al. Quantification of YouTube QoE Via Crowdsourcing [C]//IEEE International Symposium on Multimedia, Dana Point, USA, 2011: 494–499. DOI: 10.1109/ISM.2011.87
- [6] QI Y N, DAI M Y. The Effect of Frame Freezing and Frame Skipping on Video Quality [C]//International Conference on Intelligent Information Hiding and Multimedia, Pasadena, CA, USA, 2006: 423–426. DOI: 10.1109/IIH-MSP.2006.265032

- [7] LIU Y T, YUN S, MAO Y N, et al. A Study on Quality of Experience for Adaptive Streaming Service [C]//IEEE International Conference on Communications Workshops (ICC), Budapest, Hungary, 2013: 682–686. DOI: 10.1109/IC-CW.2013.6649320
- [8] CHERIF W, KSENTINI A, NÉGRU D, et al. A-SQA: Efficient Real-Time Video Streaming QoE Tool in a Future Media Internet Context [C]//IEEE International Conference on Multimedia and Expo, Barcelona, Spain, 2011: 1–6. DOI: 10.1109/ICME.2011.6011993
- [9] MOK R K P, CHAN E W W, CHANG R K C. Measuring the Quality of Experience of HTTP Video Streaming [C]//12th IFIP/IEEE International Symposium on Integrated Network Management (IM 2011) and Workshops, Dublin, Ireland, 2011: 485–492. DOI: 10.1109/INM.2011.5990550
- [10] CLAEYS M, LATRE S, FAMAËY J, et al. Design and Evaluation of a Self-Learning HTTP Adaptive Video Streaming Client [J]. IEEE Communications Letters, 2014, 18(4): 716–719. DOI: 10.1109/lcomm.2014.020414.132649
- [11] YIN X Q, JINDAL A, SEKAR V, et al. A Control-Theoretic Approach for Dynamic Adaptive Video Streaming over HTTP [J]. ACM SIGCOMM Computer Communication Review, 2015, 45(5): 325–338. DOI: 10.1145/2829988.2787486
- [12] YANG F Z, WAN S. Bitstream-Based Quality Assessment for Networked Video: A Review [J]. IEEE Communications Magazine, 2012, 50(11): 203–209. DOI: 10.1109/mcom.2012.6353702
- [13] STAEHLE B, HIRTH M, PRIES R, et al. YoM: A YouTube Application Comfort Monitoring Tool [R]. Berlin, Germany: Inst. Informatik, Tech. Rep. 467, 2010
- [14] DALAL A C and PERRY E. A New Architecture for Measuring and Assessing Streaming Media Quality [C]//3rd Workshop on Passive and Active Measurement Workshop (PAM), La Jolla, USA, 2003: 223–231
- [15] JULURI P, PLISSONNEAU L, ZENG Y, et al. Viewing YouTube from a Metropolitan Area: What do Users Accessing from Residential ISPs Experience? [C]//IFIP/IEEE International Symposium on Integrated Network Management (IM 2013), Ghent, Belgium, 2013: 589–595
- [16] MAZHAR M H, SHAFIQ Z. Real-Time Video Quality of Experience Monitoring for HTTPS and QUIC [C]//IEEE INFOCOM 2018—IEEE Conference on Computer Communications, Honolulu, USA, 2018: 1331–1339. DOI: 10.1109/INFOCOM.2018.8486321
- [17] MANGLA T, HALEPOVIC E, AMMAR M, et al. MIMIC: Using Passive Network Measurements to Estimate HTTP-Based Adaptive Video QoE Metrics [C]//Network Traffic Measurement and Analysis Conference (TMA), Dublin, Ireland, 2017: 1–6. DOI: 10.23919/TMA.2017.8002920

### Biographies

**ZHANG Xinggong** (zhangxg@pku.edu.cn) received the Ph.D. degree from Department of Computer Science, Peking University, China in 2011. He has been an associate professor with the Institute of Computer Science and Technology of Peking University since 2012. His research interests lie in multimedia networks, video communications, information - centric network, and dynamic adaptive HTTP streaming over HTTP (DASH). He was a senior researcher at Founder R&D Center, Beijing, China from 1998 to 2007, and a visiting scholar at the Polytechnic Institute of New York University, USA from 2010 to 2011.

**XIE Lan** received her master's degree from Department of Computer Science, Peking University, China, and currently is a researcher at Hulu, China. Her research interests lie in multimedia streaming, adaptive video streaming, and panoramic video streaming.

**GUO Zongming** received the bachelor, master and Ph.D. degrees from Department of Computer Science, Peking University, China in 1987, 1990 and 1994, respectively. He has been a professor with the Institute of Computer Science and Technology of Peking University since 2002. His research interests lie in multimedia streaming, image/video compression, image and video retrieval, watermarking, IPTV, and mobile multimedia. He is a member of the Institute of Electrical and Electronics Engineers (IEEE), a senior member of The Society of Motion Picture and Television Engineers (SMPTE) for China, and a senior member of China Computer Federation (CCF).

# Automating QoS and QoE Evaluation of HTTP Adaptive Streaming Systems



Christian Timmerer<sup>1</sup> and Anatoliy Zabrovskiy<sup>2</sup>

(1. Alpen-Adria-Universität Klagenfurt/Bitmovin Inc., Klagenfurt 9020, Austria;

2. Petrozavodsk State University & Alpen-Adria-Universität Klagenfurt Petrozavodsk, Petrozavodsk 185910, Russia)

**Abstract:** Streaming audio and video content currently accounts for the majority of the Internet traffic and is typically deployed over the top of the existing infrastructure. We are facing the challenge of a plethora of media players and adaptation algorithms showing different behavior but lacking a common framework for both objective and subjective evaluation of such systems. This paper aims to close this gap by proposing such a framework, describing its architecture, providing an example evaluation, and discussing open issues.

**Keywords:** HTTP adaptive streaming; DASH; QoE; performance evaluation

DOI: 10.12142/ZTECOM.201901004

<http://kns.cnki.net/kcms/detail/34.1294.TN.20190319.1713.004.html>, published online March 19, 2019

Manuscript received: 2018-08-16

## 1 Introduction

Universal access to and provisioning of multimedia content is now reality. It is easy to generate, distribute, share, and consume any media content, anywhere, anytime, on any device. Interestingly, most of these services adopt a streaming paradigm, are typically deployed over the open, unmanaged Internet, and account for the majority of today's Internet traffic. Current estimations expect that the global video traffic will be about 82 percent of all Internet traffic by 2021 [1]. Additionally, Nielsen's law of Internet bandwidth states that the users' bandwidth grows by 50 percent per year, which roughly fits data from 1983 to 2018 [2]. Thus, the users' bandwidth will reach approximately 1 Gbit/s by 2021.

Similarly, like programs and their data expand to fill the memory available in a computer system, network applications will grow and utilize the bandwidth provided. The majority of the available bandwidth is consumed by video applications and the amount of data is further increasing due to already established and emerging applications, e.g., ultra high-definition,

virtual, augmented, and mixed realities. A major technical breakthrough and enabler was certainly HTTP adaptive streaming (HAS), which provides multimedia assets in multiple versions—referred to as representations—and chops each version into short-duration segments (e.g., 2–10 s) for dynamic adaptive streaming over HTTP (MPEG-DASH or just DASH) [3] and HTTP live streaming (HLS) [4], which are both compatible with MPEG's Common Media Application Format (CMAF) [5]. Independent of the representation format, the media is provided in multiple versions (e.g., different resolutions and bitrates) and each version is divided into chunks of a few seconds (typically 2–10 s). A client first receives a manifest describing the available content on a server, and then, the client requests chunks based on its context (e.g., observed available bandwidth, buffer status, and decoding capabilities). Thus, it is able to adapt the media presentation in a dynamic, adaptive way. In Dynamic Adaptive Streaming over HTTP (DASH), the chunks are referred to as segments and the manifest is called a media presentation description (MPD). In this paper, we use the terminology of DASH, however, this work can be also applied to any other format sharing the same principles.

In the past, we witnessed a plethora of research papers in this area (e.g., [6] and [7]), however, we still lack a comprehensive evaluation framework for HAS systems in terms of both the objective metric, i.e., quality of service (QoS), and the sub-

This work was supported in part by the Austrian Research Promotion Agency (FFG) under the next generation video streaming project "PROMETHEUS".

jective metric, i.e., quality of experience (QoE). Initial evaluations have been based on simple traffic shaping and network emulation tools [8] or means to rapidly prototype the adaptation algorithms [9]. Recently, we have seen various evaluation frameworks in this domain focusing on adaptation algorithms proposed both in academia and industry [8]–[10]. However, the main focus has been on QoS rather than QoE. The latter typically requires user studies, which are mainly conducted within controlled laboratory environments. Yet, nowadays crowdsourcing is also considered as a reliable tool [11] and various platforms have been proposed [12] for this purpose.

In this paper, we propose a flexible and comprehensive framework to conduct objective and subjective evaluations of HAS systems in a fully automated and scalable way. It provides the following features:

- End-to-end HAS evaluation of players deployed in industry and algorithms proposed in academia under various conditions and use cases (e.g., codecs/representations, network configurations, end user devices, and player competition).
- Collection and analysis of objective streaming performance metrics (e.g., startup time, stalls, quality switches, and average bitrate).
- Subjective quality assessment utilizing crowdsourcing for QoE evaluation of HAS systems and QoE model testing/verification (e.g., testing or verifying a proposed QoE model using subjective user studies).

The remainder of this paper is as follows. Section 2 comprises a detailed description of the architecture of the proposed framework. Section 3 presents example evaluation results to demonstrate the capabilities of the framework. A discussion and open research issues are provided in Section 4 and Section 5 concludes the paper.

## 2 System Architecture

### 2.1 Overview

Our framework (**Fig. 1**) supports both objective and subjective evaluation of HAS systems and is composed of Adaptive

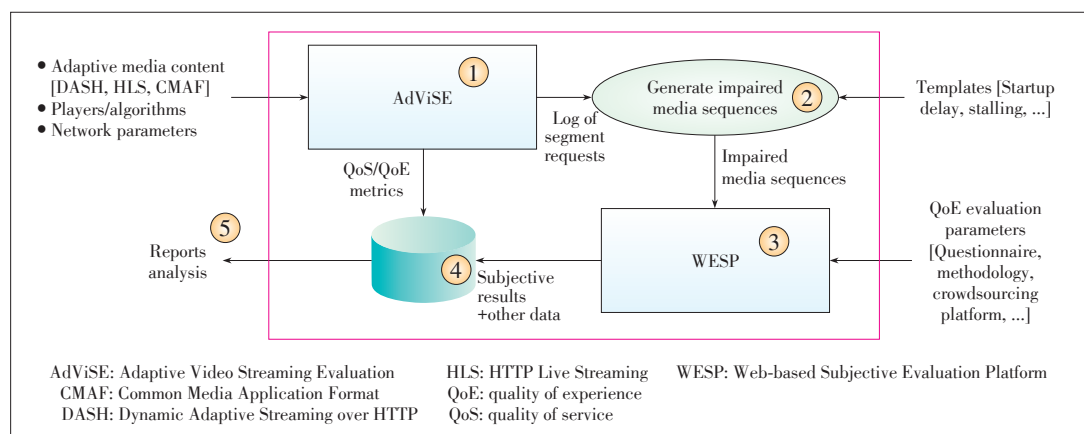
Video Streaming Evaluation (AdViSE) [13] and Web-based Subjective Evaluation Platform (WESP) [14] plus extensions. AdViSE is an adaptive video streaming evaluation framework for the automated testing of web-based media players and adaptation algorithms. It has been designed in an extensible way to support (1) different adaptive media content formats (e.g., DASH, HLS, and CMAF), (2) commercially deployed media players as well as implementations of adaptation algorithms proposed in the research literature, and (3) various networking parameters (e.g., bandwidth and delay) through network emulation. The output of AdViSE comprises a set of QoS and (objective) QoE metrics gathered and calculated during the adaptive streaming evaluation as well as a log of segment requests, which are used to generate the impaired media sequences used for the subjective evaluation.

The subjective evaluation is based on WESP [14], which is a web-based subjective evaluation platform using existing crowdsourcing platforms for subject recruitment implementing best practices according to [15]. WESP takes the impaired media sequences as an input and allows for a flexible configuration of various QoE evaluation parameters such as (1) typical questionnaire assets (e.g., drop-down menus, radio buttons, and free text fields), (2) subjective quality assessment methodology based on ITU recommendations (e.g., absolute category rating), and (3) different crowdsourcing platforms (e.g., Microworkers and Mechanical Turk). The output of WESP comprises the subjective results, including mean opinion scores (MOS) and any other data gathered during the subjective quality assessment, which are stored in a MySQL database. Together with the output of AdViSE, it is used to generate fully automated reports and data export functions, which are eventually used for further analysis.

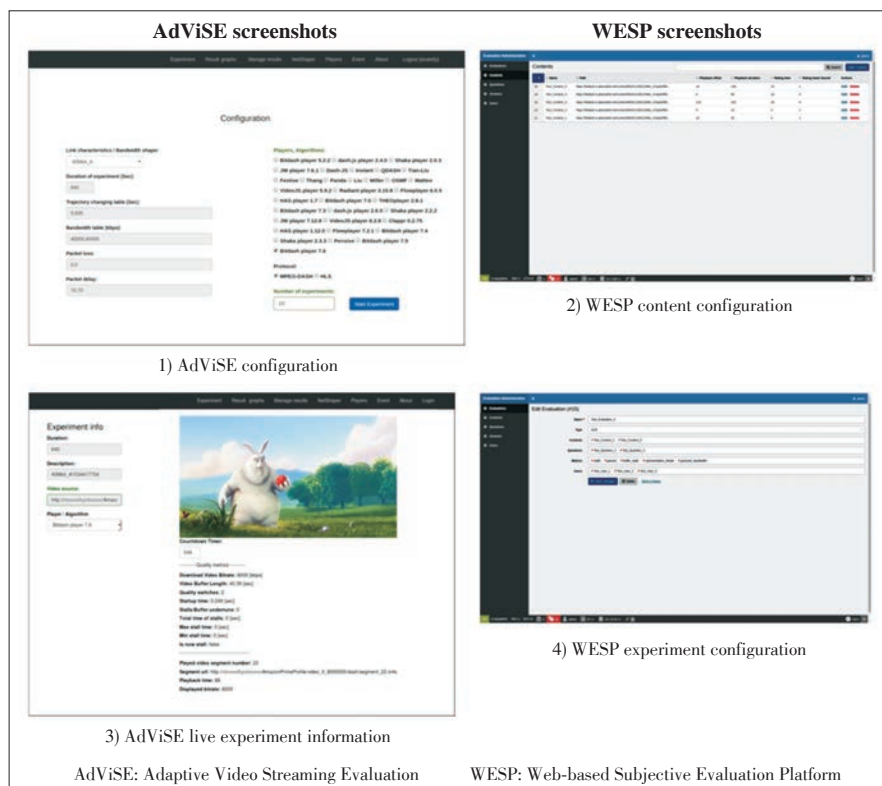
**Fig. 2** shows screenshots of both AdViSE and WESP configuration interfaces to demonstrate easy setup of HAS evaluations.

In the following we provide a detailed description of AdViSE and WESP focusing on how they connect with each other leading to a fully automated objective and subjective evaluation of HAS systems. Further details about the individual build-

**Figure 1. ▶**  
**General framework architecture: AdViSE and WESP framework for the automated testing of web-based media players and adaptation algorithms.**







▲ Figure 2. Example screenshots of AdViSE and WESP to demonstrate easy setup of HTTP Adaptive Streaming (HAS) evaluations.

ing blocks can be found in [10], [11], [13], and [14].

## 2.2 AdViSE: Adaptive Video Streaming Evaluation

AdViSE includes the following components (Fig. 3):

- Web server with standard HTTP hosting the media content and a MySQL database
- Network emulation server with a customized Mininet<sup>1</sup> environment for, e.g., bandwidth shaping
- Selenium<sup>2</sup> servers for running adaptive media players/algorithms on various platforms. Note there might be multiple physical servers, each of which hosts a limited set of players/algorithms.
- Web management interface for conducting the experiments and running the adaptive media players.

AdViSE defines a flexible system that allows adding new adaptive media players/algorithms relatively fast. The Web management interface provides two functions, (1) for configuring and conducting the experiments, and (2) including the actual player/algorithm to provide real-time information about the currently conducted experiment. Thus, the proposed framework in this paper provides means for a comprehensive end-to-

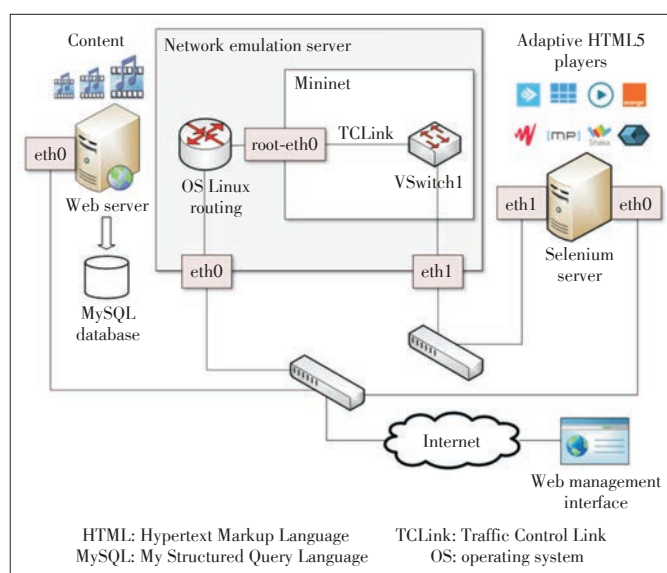
end evaluation of adaptive streaming services over HTTP including the possibility for subjective quality testing. The interface allows to define the following items and parameters:

- Configuration of network emulation profiles including the bandwidth trajectory, packet loss, and packet delay
- Specification of the number of runs of an experiment
- Selection of one or more adaptive HTML5 player (or adaptation algorithm) and the adaptive streaming format used (e.g., DASH, HLS, CMAF).

The result page provides a list of conducted experiments and the analytics section contains various metrics of the conducted experiments. It is possible to generate graphs for the results by using Highcharts<sup>3</sup> and export the raw values for further offline analysis. The following quality parameters and metrics are currently available: (1) startup time; (2) stalls (or buffer underruns); (3) number of quality switches; (4) download bitrate; (5) buffer length; (6) average bitrate; (7) instability and inefficiency; (8) simple QoE models specially designed for HAS. Further metrics

can be easily added based on what the application programming interfaces (APIs) of players actually offer, as new metrics or QoE models become available.

Finally, AdViSE provides the log of the segment requests, which are used—together with metrics such as startup time



▲ Figure 3. Architecture of adaptive video streaming evaluation framework for the automated testing of media players and adaptation algorithms.

<sup>1</sup> <http://mininet.org/>, accessed July 28, 2018.

<sup>2</sup> <http://www.seleniumhq.org/>, accessed July 28, 2018.

<sup>3</sup> <https://www.highcharts.com/>, accessed July 28, 2018.

and stalls—to generate a media sequence as received by the player, and consequently, perceived by the user. The request log is used to concatenate the segments according to the request schedule of the player, thus, reflecting the media bitrate and quality switches. Other impairments such as startup time or stalls are automatically inserted based on the corresponding metrics gathered during the evaluation and by using predefined templates (e.g., stalls displayed as spinning wheel). This impaired media sequence is used in the subsequent step for the subjective QoE evaluation using WESP, which could also include the unimpaired media presentation depending on the employed evaluation method.

In summary, AdViSE provides scalable, end-to-end HAS evaluation through emulation with a plenty of configuration possibilities regarding content configuration, players/algorithms (including for player competition), and network parameters. With AdViSE, it is possible to utilize actual content and network settings with actual dynamic, adaptive streaming including rendering. We collect various metrics from players based on their API (i.e., when access to source code is restricted) or from the algorithms/HTML5 directly. Additionally, we implemented so-called derived metrics and utilize QoE models proposed in the literature. Finally, the segment request log is used to generate impaired media sequence as perceived by end users for subjective quality testing.

### 2.3 WESP: Web-Based Subjective Evaluation Platform

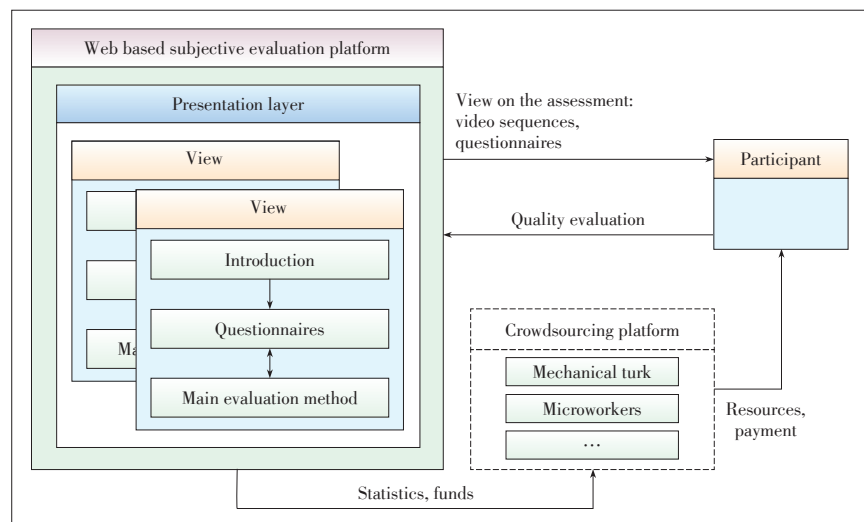
Subjective quality assessments (SQAs) are used as a vital tool for evaluating QoE. SQAs provide reliable results but is considered as cost-intensive and SQAs are typically conducted within controlled laboratory environments. Crowdsourcing has been proposed as an alternative to reduce the cost, however, various aspects need to be considered in order to get similar reliable results [15]. In the past, several frameworks have been proposed leveraging crowdsourcing platforms to conduct SQAs with each providing different features [16]. However, a common shortcoming of these frameworks is that they required tedious configuration and setup for each SQA, which made it difficult to use. Therefore, we propose to use a web-based management platform, which shall (1) enable easy and simple configuration of SQAs including possible integration of third-party tools for online surveys, (2) provide means to conduct SQAs using the

existing crowdsourcing platforms considering best practice as discussed in [15], and (3) allow for the result analysis.

The goal of WESP is not only to provide a framework, which fulfills the recommendations of the ITU for subjective evaluations of multimedia applications (e.g., BT.500<sup>4</sup>, P.910<sup>5</sup>, and P.911<sup>6</sup>), but also provide the possibility to select and to configure the preferred evaluation method via a web interface. The conceptual WESP architecture (Fig. 4) is implemented using HTML/PHP with MySQL database.

The introduction and questionnaires can be configured separately from the test methodology and may include control questions during the main evaluation. The voting possibility can be configured independently from the test methodology, providing more flexibility in selecting the appropriate voting mechanism and rating scale. The predefined voting mechanisms include the common HTML interface elements and some custom controls like a slider in different variations. The platform consists of a management layer and a presentation layer. The management layer allows for maintaining the user study such as adding new questions or multimedia content and setting up the test method to be used (including single stimulus, double stimulus, pair comparison, continuous quality evaluation, etc.). The presentation layer is responsible for presenting the content to the participants. This allows providing different views on the user study, and thus, one can define groups to which the participants may be randomly (or in a predefined way) assigned. After a participant finishes the user study, the gathered data is stored in a MySQL database. Furthermore, the platform offers methods of tracking the participant's behavior during an SQA (e.g., focus of web browser's window/tab, time for consuming each stimuli presentation, and time it takes for the voting phase) and data provided by the web player API.

The stimuli presentation can be configured independently from the test method and may be combined with the voting possibility to support continuous quality evaluations. The media



▲ Figure 4. A Web-Based Subjective Evaluation Platform (WESP).

<sup>4</sup> <https://www.itu.int/rec/R-REC-BT.500>, accessed July 28, 2018.

<sup>5</sup> <https://www.itu.int/rec/T-REC-P.910>, accessed July 28, 2018.

<sup>6</sup> <https://www.itu.int/rec/T-REC-P.911>, accessed July 28, 2018.

content can be fully downloaded and cached on the evaluation device prior starting the actual media presentation to avoid glitches during the evaluation, e.g., due to network issues. However, it also supports streaming evaluation in real-world environments where various metrics (e.g., startup time and stalls) are collected and stored for analysis.

In summary, WESP provides an extensible, web-based QoE evaluation platform utilizing crowdsourcing. It supports a plenty of evaluation methodologies and configuration possibilities. Although it has been specifically designed to implement SQAs for HAS systems using crowdsourcing (including support for real-world environments), it can also be used for SQAs within laboratory environments.

### 3 Example Evaluation Results

In this section, we provide example evaluation results of selected industry players and adaptation algorithms proposed in the research literature: Bitmovin v7.0<sup>7</sup>, dash.js v2.4.0<sup>8</sup>, Flowplayer v6.0.5<sup>9</sup>, FESTIVE [17], Instant [18], and Thang [19]. Note that we show only a small selection and the results presented here should be only seen as an example of what the framework provides rather than a full-fledged player comparison sheet. Additional further results using the tools described in this paper can be found in [10], [11], and [20].

For the evaluation, we used the Big Buck Bunny sequence<sup>10</sup> and encoded it according to the Amazon Prime video service, which offers 15 different representations as follows: 400×224 (100 kbit/s), 400×224 (150 kbit/s), 512×288 (200 kbit/s), 512×288 (300 kbit/s), 512×288 (500 kbit/s), 640×360 (800 kbit/s), 704×396 (1 200 kbit/s), 704×396 (1 800 kbit/s), 720×404 (2 400 kbit/s), 720×404 (2 500 kbit/s), 960×540 (2 995 kbit/s), 1 280×720 (3 000 kbit/s), 1 280×720 (4 500 kbit/s), 1 920×1 080 (8 000 kbit/s), and 1 920×1 080 (15 000 kbit/s). The segment length was 4 s and one audio representation at 128 kbit/s was used. We adopted the bandwidth trajectory from [8] providing both step-wise and abrupt changes in the available bandwidth, i.e., 750 kbit/s (65 s), 350 kbit/s (90 s), 2 500 kbit/s (120 s), 500 kbit/s (90 s), 700 kbit/s (30 s), 1 500 kbit/s (30 s), 2 500 kbit/s (30 s), 3 500 kbit/s (30 s), 2 000 kbit/s (30 s), 1 000 kbit/s (30 s) and 500 kbit/s (85 s). The network delay was set to 70 ms.

**Fig. 5** shows the download bitrate for the players and algorithms in question, and **Table 1** provides an overview of all metrics. Metrics a.–e. are directly retrieved from the player/HTML5 API and algorithm implementation, respectively. Metrics f.–g. utilize simple QoE models [21], [22] to calculate MOS values ranging from one to five based on a subset of other

metrics. Interestingly, industry players and research algorithms provide different performance behavior under the same conditions but can be directly compared among each other.

### 4 Discussion and Challenges

In this section, we provide a discussion about our framework for the automated objective and subjective evaluation of HAS systems. It allows for an easy setup of various configurations and running multiple evaluations in parallel. New players and algorithms can be added easily as they appear in the market and research literature. Over time it is possible to build up a repository of players and algorithms for comprehensive performance evaluation. As it is possible to run multiple Selenium servers in parallel, our framework is capable to evaluate when players/algorithms compete for bandwidth in various configurations (e.g., n player A vs. m player B).

The framework is quite flexible, and thus, comes with a high number of degrees of freedom. Hence, it is important to design the evaluation carefully. Here we provide a brief list of the aspects to consider:

- (1) Content assets: content type, codec/coding parameters (including High Dynamic Range, White Color Gamut), representations (bitrate/resolution pairs, also referred to as bitrate ladder), segment length (including GOP size), representation format (i.e., DASH, HLS, CMAF), etc.
- (2) Network parameters: bandwidth trajectory (i.e., pre-defined and network traces), delay, loss, and other networking aspects (see below for further details)
- (3) End user device environment: device type, operating system, browser, etc.
- (4) Streaming performance metrics: average bitrate, startup time, stalls (frequency, duration), quality switches (frequency, amplitude), etc.
- (5) Quantitative QoE models based on audio-video quality and/or streaming performance metrics
- (6) General HAS evaluation setup: live vs. on-demand content, single player vs. multiple players competing for bandwidth, etc.
- (7) Templates for generating the impaired media sequence (i.e., how to realize startup delay and stalls)
- (8) Questionnaire for SQA including control questions for crowdsourcing
- (9) SQA method (e.g., single stimulus, double stimulus, pairwise comparison) and its parametrization
- (10) J. Collection of all results and further (offline) analysis.

All these aspects are important to consider any a potential source of risk when conducting such experiments.

Based on our experience of conducting multiple evaluations and performance comparisons, we identified the following research challenges, possibly subject to future work:

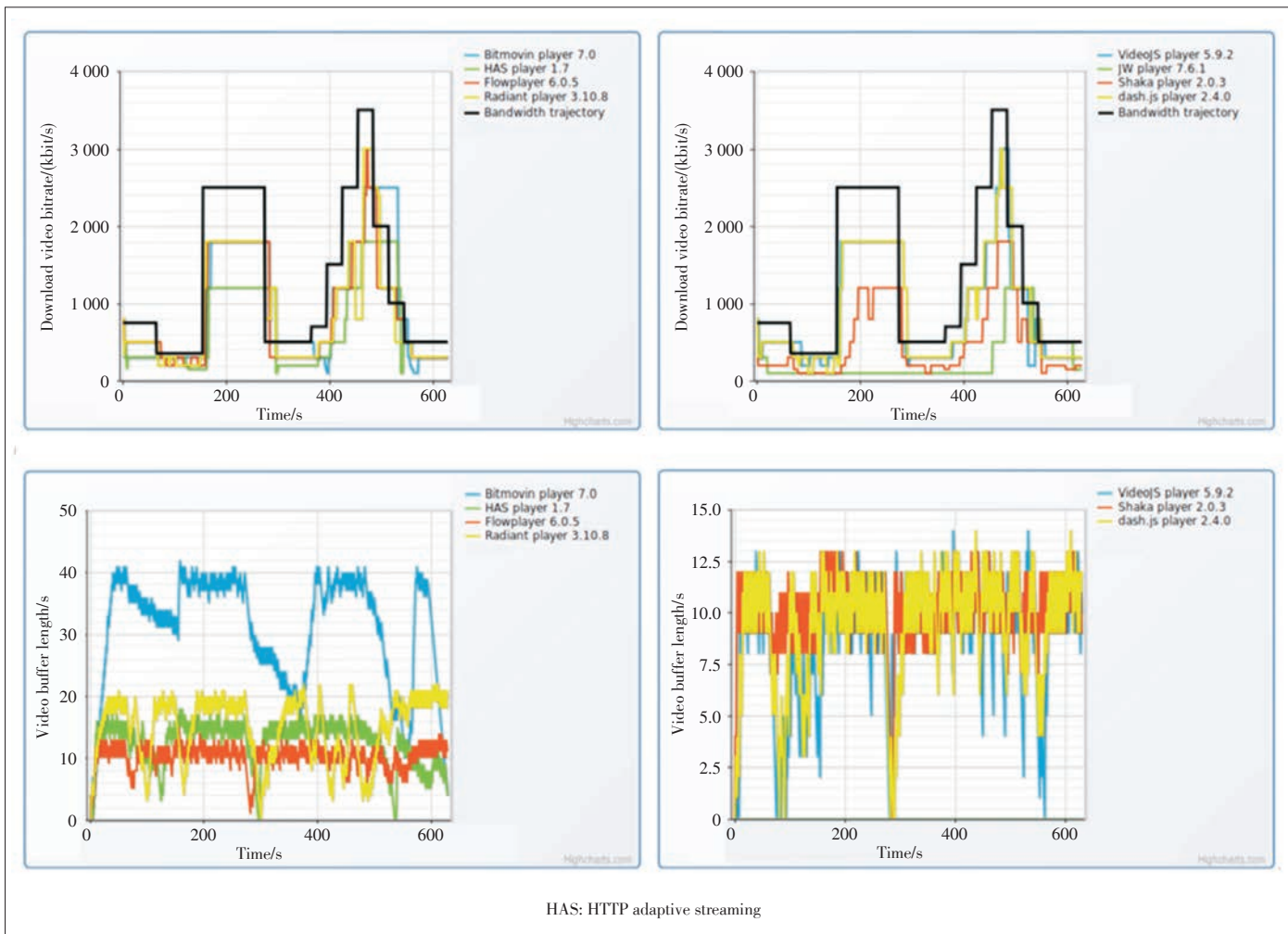
- (1) The reliability of results requires cross-validation, specifically those from SQAs, which typically call for SQAs in con-

<sup>7</sup> <https://bitmovin.com/>, accessed July 28, 2018.

<sup>8</sup> <http://dashif.org/>, accessed July 28, 2018.

<sup>9</sup> <https://flowplayer.com/>, accessed July 28, 2018.

<sup>10</sup> <https://peach.blender.org/>, accessed July 28, 2018.



▲ Figure 5. Download video bitrate (top) and video buffer length (bottom) for the selected industry players (left) and adaptation algorithms proposed in the research literature (right).

▼ Table 1. Overview of example results

Metrics	Bitmovin		dash.js		Flowplayer		Festive		Instant		Thang	
	$\mu$	$\sigma$	$\mu$	$\sigma$	$\mu$	$\sigma$	$\mu$	$\sigma$	$\mu$	$\sigma$	$\mu$	$\sigma$
a. Startup time/s	1.8	0.2	3.5	0.3	3.2	0.1	3.2	0.2	9.0	1.4	9.7	0.8
b. Stalls [#]	0	0	4	1.6	7	1.7	1	0.8	0	0	0	0
c. Stall duration/s	0	0	5.4	3.4	14.2	3.5	1.0	1.0	0	0	0	0
d. Quality switches [#]	17	1	29	4	23	1	65	5	45	3	43	9
e. Bitrate/(kbit/s)	982	22	981	9	908	10	664	10	916	19	617	14
f. QoE/MOS [Maki] [21]	4.56	0.0	4.38	0.08	4.2	0.09	4.53	0.04	4.56	0.0	4.56	0.0
g. QoE/MOS [Mok] [22]	4.1	0.0	3.84	0.05	3.79	0.02	3.93	0.1	3.62	0.09	3.58	0.06

MOS: mean opinion score    QoE: quality of experience

trolled laboratory environments.

(2) The network is a key aspect within HAS systems but is

<sup>11</sup> <https://www.planet-lab.org/>, accessed July 28, 2018.

often neglected. Network emulation is a vital tool but with limitations. For HAS systems, we also need to consider content distribution networks (CDNs), software-defined networking (SDN), information-centric networking (ICN), and next-generation (mobile) networks (e.g., 5G). Detailed analysis and evaluations of these aspects in the context of HAS are currently missing. However, recent standardization and research contributions have showed benefits for HAS systems when combined them with SDN [23].

(3) Reproducibility of such a framework can be achieved by providing containerized versions of the modules as done in [12].

This is considered critical for industry players, which often require licenses. Additionally, it could be interesting to connect to large-scale research networks (such as PlanetLab<sup>11</sup>, Virtual



Internet Routing Lab<sup>12</sup>, and GENI<sup>13</sup>).

## 5 Conclusions

This paper describes how AdViSE and WESP can be combined to perform objective and subjective evaluations of HAS systems in a fully automated and scalable way. For example, it can be used to test and compare new players/algorithms under various context conditions or research new QoE models with practically instant verification through subjective tests. The main finding of this work is that a comprehensive objective and subjective evaluation of HAS systems is feasible for both industry players and adaptation algorithms proposed in the research literature. Hence, we recommend adopting it when proposing new features in this area and evaluating the state of the art of these features.

## References

- [1] Cisco Systems, Inc. Cisco Visual Networking Index: Forecast and Methodology, 2016–2021 (White Paper). [R/OL]. (2017-09-15)[2018-07-28]. <http://bit.ly/2wm-dZJb>
- [2] NIELSEN J. Nielsen's Law of Internet Bandwidth (updated 2018) [EB/OL]. (1998-04)[2018-03-03]. <https://www.nngroup.com/articles/law-of-bandwidth>
- [3] Sodagar, I. The MPEG-DASH Standard for Multimedia Streaming Over the Internet [J]. *IEEE Multimedia*, 2011, 18(4): 62–67. DOI: 10.1109/MMUL.2011.71
- [4] PANTOS R, MAY W. HTTP Live Streaming [EB/OL]. (2017)[2018-07-28]. <https://www.ietf.org/rfc/rfc8216.txt>.
- [5] ISO/IEC. Information Technology—Multimedia Application Format (MPEG-A)—Part 19: Common Media Application Format (CMAF) for Segmented Media: ISO/IEC 23000-19 [S]. 2017.
- [6] SEUFERT M, EGGER S, SLANINA M, et al. A Survey on Quality of Experience of HTTP Adaptive Streaming [J]. *IEEE Communications Surveys & Tutorials*, 2015, 17(1): 469–492. DOI: 10.1109/comst.2014.2360940
- [7] BENTALEB A, TAANI B, BEGEN A C, et al. A Survey on Bitrate Adaptation Schemes for Streaming Media over HTTP [J]. *IEEE Communications Surveys Tutorials*, 2019, 21(1): 562–585. DOI: 10.1109/COMST.2018.2862938
- [8] MÜLLER C, LEDERER S, TIMMERER C. An Evaluation of Dynamic Adaptive Streaming over HTTP in Vehicular Environments [C]//*Proceedings of the 4th Workshop on Mobile Video*, ser. MoVid'12, New York, USA: ACM, 2012: 37–42. DOI: 10.1145/2151677.2151686
- [9] CICCIO De L, CALDARALO V, PALMISANO V, et al. TAPAS: A Tool for rApid Prototyping of Adaptive Streaming Algorithms [C]//*Proceedings of the 2014 Workshop on Design, Quality and Deployment of Adaptive Video Streaming*, ser. VideoNext'14, New York, USA: ACM, 2014: 1–6. DOI: 10.1145/2676652.2676654
- [10] ZABROVSKIY A, PETROV E, KUZMIN E, et al. Evaluation of the Performance of Adaptive HTTP Streaming Systems [EB/OL]. *CoRR*, vol. abs/1710.02459 [2017]. <http://arxiv.org/abs/1710.02459>
- [11] TIMMERER C, ZABROVSKIY A, KUZMIN E, et al. Quality of Experience of Commercially Deployed Adaptive Media Players [C]//*21st Conference of Open Innovations Association (FRUCT)*, Helsinki, Finland, 2017: 330–335
- [12] STOHR D, FRÖMMGEN A, RIZK A, et al. Where are the Sweet Spots? A Systematic Approach to Reproducible DASH Player Comparisons [C]//*Proceedings of the 2017 ACM on Multimedia Conference*, ser. MM'17, New York, USA: ACM, 2017: 1113–1121. DOI: 10.1145/3123266.3123426
- [13] ZABROVSKIY A, KUZMIN E, PETROV E, et al. AdViSE: Adaptive Video Streaming Evaluation Framework for the Automated Testing of Media Players [C]//*Proceedings of the 8th ACM on Multimedia Systems Conference*, ser. MM-Sys'17, New York, USA: ACM, 2017: pp. 217–220. DOI: 10.1145/3083187.3083221
- [14] RAINER B, WATTL M, TIMMERER C. A Web Based Subjective Evaluation Platform [C]//*Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*, Klagenfurt am Wörthersee, Austria, 2013: 24–25. DOI: 10.1109/QoMEX.2013.6603196
- [15] HOSSFELD T, KEIMEL C, HIRTH M, et al. Best Practices for QoE Crowdstesting: QoE Assessment with Crowdsourcing [J]. *IEEE Transactions on Multimedia*, 2014, 16(2): 541–558. DOI: 10.1109/tmm.2013.2291663
- [16] HOSSFELD T, HIRTH M, KORSHUNOV P, et al. Survey of Web-Based Crowdsourcing Frameworks for Subjective Quality Assessment [C]//*IEEE 16th International Workshop on Multimedia Signal Processing (MMSP)*, Jakarta, Indonesia, 2014: 1–6. DOI: 10.1109/MMSP.2014.6958831
- [17] JIANG J, SEKAR V, ZHANG H. Improving Fairness, Efficiency, and Stability in HTTP-based Adaptive Video Streaming with FESTIVE [C]//*Proceedings of the 8th International Conference on Emerging Networking Experiments and Technologies*, ser. CoNEXT '12, New York, USA: ACM, 2012: 97–108. DOI: 10.1145/2413176.2413189
- [18] ROMERO L R. A Dynamic Adaptive HTTP Streaming Video Service for Google Android [D]. Master of Science Thesis, Stockholm, Sweden: Royal Institute of Technology (KTH) Stockholm, 2011.
- [19] THANG T, HO Q D, KANG J, et al. Adaptive Streaming of Audiovisual Content Using MPEG DASH [J]. *IEEE Transactions on Consumer Electronics*, 2012, 58(1): 78–85. DOI: 10.1109/tce.2012.6170058
- [20] TIMMERER C, MAIERO M, RAINER B. Which Adaptation Logic? An Objective and Subjective Performance Evaluation of HTTP-based Adaptive Media Streaming Systems [EB/OL]. *arXiv:1606.00341* (2016)[2018-07-28]. <http://arxiv.org/abs/1606.00341>
- [21] MÄKI T, VARELA M, AMMAR D. A Layered Model for Quality Estimation of HTTP Video from QoS Measurements [C]//*11th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, Bangkok, Thailand, 2015: 591–598. DOI: 10.1109/SITIS.2015.41
- [22] MOK R K P, CHAN E W W, CHANG R K C. Measuring the Quality of Experience of HTTP Video Streaming [C]//*12th IFIP/IEEE International Symposium on Integrated Network Management (IM 2011) and Workshops*, Dublin, Ireland, 2011: 485–492. DOI: 10.1109/INM.2011.5990550
- [23] BENTALEB A, BEGEN A C, ZIMMERMANN R, et al. SDNHAS: An SDN-Enabled Architecture to Optimize QoE in HTTP Adaptive Streaming [J]. *IEEE Transactions on Multimedia*, 2017, 19(10): 2136–2151. DOI: 10.1109/tmm.2017.2733344

## Biographies

**Christian Timmerer** (christian.timmerer@itec.aau.at) is an associate professor with Alpen-Adria-Universität Klagenfurt, Austria. He is a Co-Founder of Bitmovin Inc., San Francisco, USA, as well as the CIO and the Head of Research and Standardization. He has co-authored seven patents and over 200 publications in workshops, conferences, journals, and book chapters. He participated in several EC-funded projects, notably DANA, ENTHRON, P2P-Next, ALICANTE, SocialSensor, ICOSOLE, and the COST Action IC1003 QUALINET. He also participated in ISO/MPEG work for several years, notably in the areas of MPEG-21, MPEG-M, MPEG-V, and MPEG-DASH. His research interests include immersive multimedia communications, streaming, adaptation, and quality of experience. He was the General Chair of WIAMIS 2008, QoMEX 2013, ACM MMSys 2016, and Packet Video 2018. Further information can be found at <http://blog.timmerer.com>.

**Anatoliy Zabrovskiy** received his B.S. and M.S. degrees in information and computer technology from Petrozavodsk State University, Russia in 2006 and 2008 respectively, and a Ph.D. degree in engineering from the same university in 2013. He has been working in the field of network and multimedia communication technologies for over ten years. He was a Cisco certified academy instructor for CCNA. He was award winner of two international programs: Scholarships of the Scholarship Foundation of the Republic of Austria for Postdocs and Erasmus Mundus External Cooperation Window program for doctorate students. He was a prize winner of Sun Microsystems contest "Idea2Project". He is currently a post-doctoral researcher at the Department of Information Technology (ITEC), Alpen-Adria-Universität Klagenfurt, Austria. He is a member of the Technical Program Committee of ACM MMSys 2019. His research interests include video streaming, network technologies, quality of experience, and machine learning.

<sup>12</sup> <http://virl.cisco.com/getvirl/>, accessed July 28, 2018.

<sup>13</sup> <http://www.geni.net/>, accessed July 28, 2018.





# Quality of Experience Effects in Video Delivery

CHEN Jinling, XU Yiwen, LIU Yisang, HUANG Huiwen, and ZHUANG Zhongwen

(Fuzhou University, Fuzhou, Fujian 350108, China)

**Abstract:** With the popularization of smartphones and high-speed networks, a larger number of users are getting used to watching videos online and have increasing requirements of video quality. Therefore, the video content delivery has become a progressively challenging task, especially for ultra-high-definition (UHD) videos and heterogonous networks. Recently, quality of experience (QoE), which represents the true visual experience of users, has shown its advantages in management of video delivery and thus attracted increasing attention. In a video delivery system, the user QoE can be greatly influenced by numerous effects from video sources to display terminals. In this paper, we first investigate the significant differences between quality of service (QoS) and QoE. In addition, we summarize the end-to-end QoE effects in video delivery and present their classification based on the deployment. We also specifically analyze the impacts of different kinds of factors on QoE in video transmission systems.

**Keywords:** QoE; QoS; video delivery; video quality

DOI: 10.12142/ZTECOM.201901005

<http://kns.cnki.net/kcms/detail/34.1294.TN.20190221.1044.002.html>, published online February 21, 2019

Manuscript received: 2018-08-13

## 1 Introduction

The recent development of high-speed networks and smart devices have brought a great need of multimedia services. As a result, it is necessary to develop quality metrics to measure the performance of video services. During the past decades, an increasing number of conventional quality metrics have been proposed to predict the quality of videos. The peak signal to noise ratio (PSNR) and structural similarity (SSIM) index [1], as the most widely used signal fidelity metrics, evaluate the quality of videos by the similarity between the reference and distorted video frames. In addition, quality of service (QoS) [2] has been developed to estimate video quality at system perspective and become the most suitable one for the measurement of performance and reliability of network elements.

All above-mentioned quality metrics are limited to evaluate video quality from the perspective of signals and systems and do not take users' true visual experience into account. Therefore, quality of experience (QoE) [3] has been proposed to represent the true experience of a user, which has overtaken the traditionally used objective measures. It is defined as "the overall acceptability of an application or service, as perceived subjectively by the end user" [4]. It includes the complete end-to-end system effects and may be affected by user expectations and context. Then, to mitigate some of the problems related with the above definition, the following definition of QoE was developed: "Degree of delight of the user of a service. In the context of communication services, it is influenced by content, network, device, application, user expectations and goals, and context of use." [5] However, these definitions seem to only reflect the user's acceptance. Taking the limitations of the above definitions into account, a more accurate definition was proposed in 2013: "QoE is the degree of delight or annoyance of the user of an application or service. It results from the fulfill-

This work is supported by the National Natural Science Foundation of China (Grant 61671152).

ment of his or her expectations with respect to the utility and/or enjoyment of the application or service in the light of the users' personality and current state." [6]

The new definition of QoE emphasizes the subjective experience of users compared to the objective indicators. Due to its advantage, QoE has been widely used in video delivery. Hoßfeld et al. [7] studied YouTube video streaming in terms of the QoE impact of Internet delivery. Different QoE monitoring approaches were qualitatively compared and estimated considering the accuracy of QoE estimation. Rehman et al. [8] performed a subjective experiment to investigate the impacts of display device properties and viewing conditions on perceptual video QoE. They also proposed a full-reference (FR) video QoE metric, named SSIMplus, to predict the perceptual quality of a video. Maia et al. [9] analyzed the subjective, objective and hybrid QoE approaches in video streaming services. Zhao et al. [10] described the main QoE factors of video transmission and the modeling approaches of these factors, and surveyed the QoE assessment approaches, including subjective test and objective QoE monitoring. Li et al. [11] proposed a novel QoE-driven centralized scheduling framework for multiuser downlink networks.

As introduced in the above studies, the increasing dominance of video traffic has driven the widespread use of QoE in video transmission. As a result, it is necessary to survey the end-to-end QoE effects in video delivery. In order to achieve the purpose, we first investigate the differences between QoS and QoE in Section 2. Then, the classification of QoE effects is provided in Section 3. Furthermore, we analyze the influencing factors of QoE in Section 4. Finally, Section 5 concludes the paper.

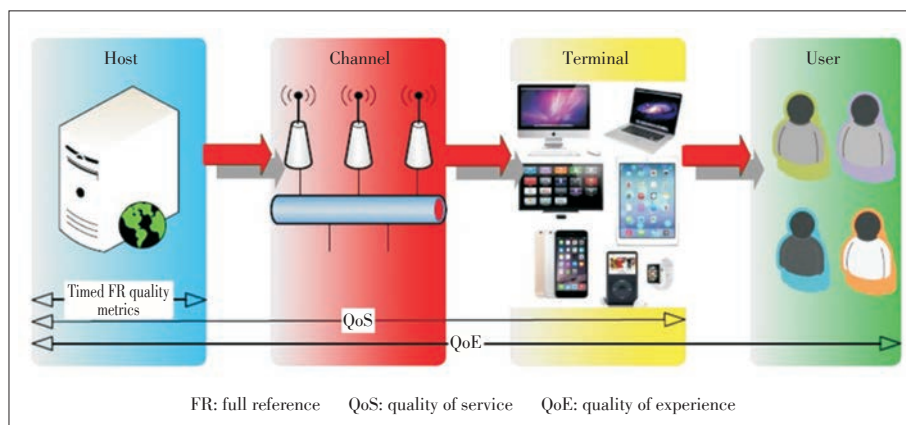
## 2 From QoS to QoE

In general, the conventional QoS metrics have been used to study the performance of online services and networked elements. QoS reflects the reliability of the network and its components, which was described by ITU as: "totality of characteristics of a telecommunications service that bear on its ability to satisfy stated and implied needs of the user of the service." [2] This definition implies several obvious differences from the concept of QoE. First, QoS handles the performance aspects of physical systems. Thus, it is a network-centric metric. The commonly used QoS metrics are throughput, bandwidth, packet loss, delay, and jitter. However, QoE is a user-centric metric that deals with the users' assessment of system performance, such as context, culture, user-specific characteristics, delivered content, and psychological profiles,

among other factors. The second difference resides in the fact that QoS and QoE have different scopes. The QoS is usually focused on telecommunications and network services, while QoE covers more extensive areas, which is not limited to telecommunications and networks. QoE mainly faces users and business. The third difference between QoS and QoE is that QoS relies on the analytic approaches and empirical or simulative measurements, whereas QoE depends on multidisciplinary and multi-methodological approaches.

Despite of these differences, QoE is still dependent on QoS to a certain extent. The relationship between QoS and QoE can be obtained from **Fig. 1**. It can be seen that QoE covers more influence factors than QoS and the conventional timed video quality metrics. Therefore, QoE and QoS are not mutually exclusive; on the contrary, QoE is an extension of QoS, which takes subjective factors (e.g., user and context) into consideration on the basis of QoS. In recent years, researchers have tried to implement QoS to QoE mapping. Aroussi et al. [12] proposed a global correlation model between QoE and QoS based on the multiple linear regression (MLR). Alberti et al. [13] presented a nonlinear psychometric model to evaluate the mean opinion score (MOS) from the QoS parameters for dynamic adaptive streaming over HTTP (DASH) streaming systems. Mansouri et al. [14] developed an integrated QoS and QoE evaluation system in order to evaluate voice over IP (VoIP) service quality in a more comprehensive way. Anchuen et al. [15] estimated the satisfaction of users in terms of QoE using neural network approach, where the input of the proposed model was obtained by five QoS parameters. Ning et al. [16] analyzed the sensitivity of QoE to different QoS parameters and provided the mapping relationship between QoS to QoE. García-Pineda et al. [17] used a statistical technique that employs all kinds of variables related to QoS, to evaluate the subjective QoE.

In summary, the main difference between QoS and QoE is that QoS depends on the network perspective, while QoE focuses on the users' perspective. However, QoS and QoE are not two independent metrics, because QoE adds the effects of context and human to the system factors that are widely studied on



▲ Figure 1. Illustration of the impact of end-to-end system on QoE in video transmission.

QoS. Furthermore, the above studies indicate that it is possible to develop a real-time QoE metric based on QoS factors.

### 3 Classification of QoE Effects

As an overall metric, QoE can be influenced by various factors in the end-to-end video delivery system. In practice, these influence factors include video capture, coding, storage, delivery, decoding, rendering, display and context of use. Furthermore, QoE is also affected by user factors such as user personality and expectations. Apparently, all these factors have direct impact on the design of QoE-aware optimization techniques for video delivery. Here, we present a classification and an enumeration of all relevant influence factors of QoE in video delivery system in this section.

In [18], the contributions of QoE in video delivery were divided into three categories including content preparation, content delivery and the customer environment. In [10], the end-to-end influence factors of QoE in video transmission were summarized into three categories including system influence factors, context influence factors and human influence factors. Following these ideas, we classify the QoE influence factors based on their operational locations in the end-to-end video delivery system. We propose to partition the video delivery system into four major elements including host, channel, terminal, and user, as shown in Fig. 1. Correspondingly, we classify these influence factors into four categories accordingly: host factors, channel factors, terminal factors, and user factors. The major advantage of this classification is that it can support various QoE mappings and cross-layer optimization designs at different taxonomies of the video delivery system.

Base on the partition method mentioned above, the influence factors of QoE are summarized in **Table 1** and are discussed as follows.

(1) Host factors: These factors include video content factors and media factors. At the host, a source video is generally processed and/or coded before being transmitted in order to reduce the storage size and meet the bandwidth budget. In addition, the unimpaired source video is usually available at the host. Therefore, we can utilize it (or the features extracted from it) as a reference to measure the QoE loss during video processing and compression. The temporal and spatial samplings may also have impacts on the user's QoE.

(2) Channel factors: The channel factors are mainly network-related factors. It is known that the packet transmission can be influenced by different network configurations such as bandwidth, throughput, resource requirements, scheduling, and sometimes, zapping time and handoff. However, inappropriate network configurations or poor network conditions may cause packet delay, jitter, loss or error rate, which will degrade the user's QoE. For online video purchasing, the pricing model and the prices also affect the QoE. In addition, the channel factors and relevant host factors can be regarded as QoS param-

▼ **Table 1. QoE influence factors at different taxonomies of a video transmission system**

Taxonomy		QoE influence factors
Host	Content factors	Temporal/spatial requirements, color depth, texture, 2D/3D, content reliability, artifacts, etc.
	Media factors	Encoding, resolution, sampling rate, frame rate, media synchronization, etc.
Channel	Network factors	Delay, jitter, loss, error rate, bandwidth, throughput, path selection, resource requirements, scheduling, zapping time, hand-off, etc.
	Other factors	Pricing, etc.
Terminal	Device factors	Decoding, error concealment, zooming, rendering, display size, screen resolution, color depth, user interface, CPU and memory, battery, etc.
	Other factors	Luminance, viewing distance, movement, interactivity, personalization, security, mobility, etc.
User	Physiological factors	Gender, age, heart rate, electrodermal activity, etc.
	Psychological factors	Attention, interest, personality, mood, preconceptions, user expectation/goal, etc.

CPU: central processing unit    QoE: quality of experience

ters during video delivery process.

(3) Terminal factors: This category includes the device factors (e.g., decoding parameters, reception device settings, and display parameters) and environmental configurations (e.g., luminance, viewing distance, and movement). In addition, the usability and accessibility of graphic user interface (GUI) and video interactivity during playing are also considered as QoE factors drawn into this category. Furthermore, the security and personalization issues also belong to this category in some particular applications [6].

(4) User factors: The user factors are generally composed of physiological factors (e.g., gender, age, and heart rate) and psychological factors (e.g., attention, interest, and mood). In practice, the user factors, especially the psychological factors, are difficult to be directly measured, which leads to a great obstacle to the research of the influence of user factors on QoE. Luckily, in recent years, researchers have found some indirect ways to measure user factors and subsequently made several breakthroughs in the study of the impact of user factors on QoE, which will be discussed in detail in Section 4.

## 4 Further Analysis

In video delivery system, QoE covers the end-to-end factors, which are from host to users, to affect the users' experience on video services. In this section, we further analyze the specific impacts of these factors on QoE.

### 4.1 Impact of Host Factors

It is well known that the distortion is inevitably introduced to reduce the visual quality of videos in the process of acquisition, processing and compressed. Therefore, video quality assessment (VQA) methods have been extensively used to pre-

dict the impact of distortion on video quality. According to the available amount of reference information at the host, VQA can be also divided into three categories: FR, reduced-reference (RR) and no-reference (NR) methods. In an FR method, an unimpaired original video is compared frame by frame with the impaired video to obtain a VQA metric. Typical FR measures are PSNR, SSIM [1], etc. These FR methods achieve higher accuracy because of their reference original videos. However, compared with the other two methods, FR methods have a small scope of applications due to its demand for unimpaired videos, which are generally applicable only to the host. For the RR methods [19]–[21], they only need to extract some features from unimpaired original videos and transmit them along with the impaired videos. Thus they are more applicable than FR measures. How to extract features is the main challenge for them. The NR methods are completely independent of the original video information, which is both its advantage and its difficulty. In recent years, several NR methods have been proposed due to their practicality [22]–[24].

In addition, other influence factors at the host have been studied in the past years. The depth perception assessment metrics for 3D stereoscopic videos were proposed in [25] and [26]. Ou et al. [27] analyzed the impacts of spatial, temporal and amplitude resolution on the bit rate of a compressed video and proposed an analytical rate model. The effect of color depth for QoE was investigated in [28]. A determining method of frame rate and resolution was proposed to improve QoE [29]. In [30], the impact of video resolution was also discussed.

All of the above works study the influences of host factors on QoE, and lay the foundation for the development of VQA based on host factors. In recent years, VQA technology of ordinary video has become more mature, especially the FR methods. However, with the popularity of special videos (e.g., ultra-high-definition (UHD), high dynamic range (HDR), 3D, and 360-degree videos), how to measure the host factors in these video delivery has been an open question. In addition, since original special videos are usually unavailable in many real-world video applications, the RR and NR methods will play an important role in VQA. In summary, the impacts of host factors on QoE will be further investigated in the future.

#### 4.2 Impact of Channel Factors

QoE can be affected by numerous channel factors when the video stream is transmitted over a channel, such as bandwidth and throughput. These can be considered as a part of the QoS parameters. Due to the measurability of QoS, QoE assessment approaches based on channel factors are widely used in video delivery system.

In [31], Frnda et al. discussed the impact of packet loss and delay variation on QoE and designed a prediction model for estimation of triple play services. Maeda et al. [32] investigated the influence of network delay on QoE in a networked haptic drum system. Nunome et al. [33] investigated the effect of two

allocation methods of bandwidth on QoE in multiview video and audio transmission. Begluk et al. [34] proposed a machine-learning model to predict QoE based on network-related factors (e.g., delay, jitter, and loss) as input data. Gutierrez et al. [35] studied the impact of transmission errors in 3DTV and proposed a novel evaluation methodology for QoE.

These researchers extensively investigated the influence of channel factors on QoE, especially network factors. However, other factors (e.g., pricing) are still not well studied at present. They are also highly needed for the study of QoE. Furthermore, the development of heterogenous networks such as 5G network has improved the users' requirements for video transmission. For these new emerging networks, how to measure QoE is a new challenge and research direction. Therefore, the influences of channel factors on QoE will be widely studied on these networks.

#### 4.3 Impact of Terminal Factors

When watching the same video in the same environment, there are some differences in the experience of users using different terminal devices. The reason for the difference of QoE is the influences of the terminal factors, including screen resolution, display size, luminance, viewing distance, etc. With the development of smart devices, more and more researchers have studied the influences of terminal factors on QoE.

Beyer et al. [36] observed a considerable impact of display size on overall quality. In [37], Vucic et al. studied the impacts of smartphone factors (including CPU, screen size and display resolution) on the QoE for multi-party video conferencing. Jegannatan et al. [38] studied the effect of user interfaces on QoE of multiview video and audio over IP networks. Edstrom et al. [39] mainly investigated environmental luminance at different levels and its impact on the user's viewing experience. Triyas-on [40] conducted a subjective experiment to prove screen size has an effect toward the QoE of remote cloud-based virtual desktop.

It should be pointed out that there are not many studies on the impacts of terminal factors on QoE. Most of the current studies are mainly based on device-related and environment-related. They seem to ignore the influences of other terminal factors on QoE, such as interactivity and personalization. However, with the development of smart devices (e.g., smartphones, tablets, and VR devices), users put forward higher requirements for the interactivity, personalization and security of video services. It suggests that the research of these terminal-related factors plays a pivotal role in VQA. Thereby, QoE research based on terminal-related factors will become an important research direction in the future.

#### 4.4 Impact of User Factors

Since the human factors are strongly related to and may affect other factors, they play an increasingly important role in the impact of QoE. They can well reflect each user's personal



experience. However, they are highly complex and not well comprehended because of their subjectivity and relevance [41].

In a study given in [42], Guntuku et al. found that personality and culture play a key role in predicting the intensity of negative affect. Murray et al. [43] evaluated the impact of users' age and gender on user QoE based on user perception of olfaction based multimedia. In addition, Murray et al. [44] proposed a model based on empirical data to estimate user QoE. The result indicates that human factors play an important role in perceptual multimedia quality of olfaction enhanced multimedia. Song et al. [45] developed a user-centric objective QoE evaluation model to predict QoE considering perceptual audiovisual quality and user interest in audiovisual content. In [46], Eynard et al. discussed the impact of verbal communication on the user experience in the context of virtual reality (VR).

All these works demonstrate that the human factors play a key role in QoE assessment. Since these factors differ QoE from QoS, an increasing number of researchers try to build QoE models based on the human factors to achieve more accurate video quality assessment. However, most of these factors are not measured directly. Thence, these studies are currently focused on users' touch, visual and other aspects in the special videos, especially immersive applications. Although these studies have made some progress, how to directly measure the impact of the human factors on QoE is still a challenge due to their complexity.

## 5 Conclusions

In this paper, we discuss three main differences between QoS and QoE and the possibility of QoS and QoE mapping. QoE can be influenced by various factors in the video delivery and we summarize these factors into four categories: host factors, channel factors, terminal factors, and user factors. In addition, we analyze the specific impacts of different types of factors on QoE. We hope our study may promote the development and application of VQA approaches.

## References

- [1] WANG Z, BOVIK A C, SHEIKH H R, et al. Image Quality Assessment: From Error Visibility to Structural Similarity [J]. *IEEE Transactions on Image Processing*, 2004, 13(4): 600–612. DOI: 10.1109/tip.2003.819861
- [2] Definitions of Terms Related to Quality of Service: ITU-T Recommendation E.800 [S]. 2008
- [3] MOLLER S, RAAKE A, Eds. *Quality of Experience: Advanced Concepts, Applications and Methods* [M]. Cham, Switzerland: Springer, 2014
- [4] Definition of Quality of Experience (QoE), ITU TD 109rev2 (PLEN/12) [S]. 2007
- [5] MOLLER S. *Quality Engineering—Qualität kommunikationstechnischer Systeme* [M]. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010
- [6] CALLET P L, MOLLER S, PERKIS A, et al. *Qualinet White Paper on Definitions of Quality of Experience* [R]. Novi Sad: European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003), 2013
- [7] HOBFIELD T, SCHATZ R, BIERACK E, et al. *Internet Video Delivery in YouTube: From Traffic Measurements to Quality of Experience* [M]//HOBFIELD T, SCHATZ R, BIERACK E, et al. eds. *Data Traffic Monitoring and Analysis*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013: 264–301. DOI: 10.1007/978-3-642-36784-7\_11
- [8] REHMAN A, ZENG K, WANG Z. Display Device-Adapted Video Quality-of-Experience Assessment [C]//IS&T/SPIE Annual Symposium on Electronic Imaging. San Francisco, USA, SPIE 9394. DOI: 10.1117/12.2077917
- [9] MAIA O B, YEHIA H C, DE ERICO L. A Concise Review of the Quality of Experience Assessment for Video Streaming [J]. *Computer Communications*, 2015, 57: 1–12. DOI: 10.1016/j.comcom.2014.11.005
- [10] ZHAO T S, LIU Q, CHEN C W. QoE in Video Transmission: A User Experience-Driven Strategy [J]. *IEEE Communications Surveys & Tutorials*, 2017, 19(1): 285–302. DOI: 10.1109/comst.2016.2619982
- [11] LI T T, ZHANG H X, TIAN J, et al. QoE-Driven Centralized Scheduling for HTTP Adaptive Video Streaming Transmission over Wireless Networks [C]//9th International Conference on Wireless Communications and Signal Processing (WCSP). Nanjing, China, 2017: 1–6. DOI: 10.1109/WCSP.2017.8171114
- [12] AROUSSE S, BOUABANA-TEBIBEL T, MELLOUK A. Empirical QoE/QoS Correlation Model Based on Multiple Parameters for VoD Flows [C]//IEEE Global Communications Conference (GLOBECOM). Anaheim, CA, USA, 2012: 1963–1968. DOI: 10.1109/GLOCOM.2012.6503403
- [13] ALBERTI C, RENZI D, TIMMERER C, et al. Automated QoE Evaluation of Dynamic Adaptive Streaming over HTTP [C]//Fifth International Workshop on Quality of Multimedia Experience (QoMEX). Klagenfurt am Wörthersee, Austria, 2013: 58–63. DOI: 10.1109/QoMEX.2013.6603211
- [14] MANSOURI T, NABAVI A, ZARE RAVASAN A, et al. A Practical Model for Ensemble Estimation of QoS and QoE in VoIP Services via Fuzzy Inference Systems and Fuzzy Evidence Theory [J]. *Telecommunication Systems*, 2016, 61(4): 861–873. DOI: 10.1007/s11235-015-0041-6
- [15] ANCHUEN P, UTHANSAKUL P, UTHANSAKUL M. QOE Model in Cellular Networks Based on QoS Measurements Using Neural Network Approach [C]//13th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON). Chiang Mai, Thailand, 2016: 1–5. DOI: 10.1109/ECTICon.2016.7561318
- [16] NING Z L, LIU Y Q, WANG X J, et al. A Novel QoS-Based QoE Evaluation Method for Streaming Video Service [C]//IEEE International Conference on Internet of Things (IThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData). Exeter, UK, 2017: 956–961. DOI: 10.1109/IThings-GreenCom-CPSCom-SmartData.2017.147
- [17] GARCÍA-PINEDA M, SEGURA-GARCÍA J, FELICI-CASTELL S. A Holistic Modeling for QoE Estimation in Live Video Streaming Applications over LTE Advanced Technologies with Full and Non Reference Approaches [J]. *Computer Communications*, 2018, 117: 13–23. DOI: 10.1016/j.comcom.2017.12.010
- [18] VALERDI J, GONZÁLEZ A, GARRIDO F J. Automatic Testing and Measurement of QoE in IPTV Using Image and Video Comparison [C]//Fourth International Conference on Digital Telecommunications. Colmar, France, 2009: 75–81. DOI: 10.1109/ICDT.2009.21
- [19] WANG M M, ZHANG F, AGRAFIOTIS D. A very Low Complexity Reduced Reference Video Quality Metric Based on Spatio-Temporal Information Selection [C]//IEEE International Conference on Image Processing (ICIP). Quebec City, Canada, 2015: 571–575. DOI: 10.1109/ICIP.2015.7350863
- [20] AABED M A, ALREGIB G. Reduced-Reference Perceptual Quality Assessment for Video Streaming [C]//IEEE International Conference on Image Processing (ICIP). Quebec City, Canada, 2015: 2394–2398. DOI: 10.1109/ICIP.2015.7351231
- [21] YU M, ZHENG K H, JIANG G Y, et al. Binocular Perception Based Reduced-Reference Stereo Video Quality Assessment Method [J]. *Journal of Visual Communication and Image Representation*, 2016, 38: 246–255. DOI: 10.1016/j.jvcir.2016.03.010
- [22] CHEN Q G, JIN Y H, YANG T. A Supervised No-Reference QOE Assessment Model on IPTV Services [C]//4th International Conference on Cloud Computing and Intelligence Systems (CCIS). Beijing, China, 2016: 272–277. DOI: 10.1109/CCIS.2016.7790268
- [23] TORRES VEGA M, MOCANU D C, STAVROU S, et al. Predictive No-Reference Assessment of Video Quality [J]. *Signal Processing: Image Communication*, 2017, 52: 20–32. DOI: 10.1016/j.image.2016.12.001
- [24] ZHANG H, LI F, LI N. Compressed-Domain-Based No-Reference Video Quality Assessment Model Considering Fast Motion and Scene Change [J]. *Multimedia Tools and Applications*, 2017, 76(7): 9485–9502. DOI: 10.1007/s11042-016-3558-0



- [25] LEBRETON P, RAAKE A, BARKOWSKY M, et al. Evaluating Depth Perception of 3D Stereoscopic Videos [J]. *IEEE Journal of Selected Topics in Signal Processing*, 2012, 6(6): 710–720. DOI: 10.1109/jstsp.2012.2213236
- [26] NUR YILMAZ G. A no Reference Depth Perception Assessment Metric for 3D Video [J]. *Multimedia Tools and Applications*, 2015, 74(17): 6937–6950. DOI: 10.1007/s11042-014-1945-y
- [27] OU Y F, XUE Y Y, WANG Y. Q-STAR: A Perceptual Video Quality Model Considering Impact of Spatial, Temporal, and Amplitude Resolutions [J]. *IEEE Transactions on Image Processing*, 2014, 23(6): 2473–2486. DOI: 10.1109/tip.2014.2303636
- [28] BOITARD R, POURAZAD M T, NASIOPOULOS P. Evaluation of Chroma Sub-sampling for High Dynamic Range Video Compression [C]//*IEEE International Conference on Electronics, Circuits and Systems (ICECS)*. Monte Carlo, Monaco, 2016: 696–699. DOI: 10.1109/ICECS.2016.7841297
- [29] LI M, SONG J B, HUI L. A Determining Method of Frame Rate and Resolution to Boost the Video Live QoE [C]//*2nd International Conference on Multimedia and Image Processing (ICMIP)*. Wuhan, China, 2017: 206–209. DOI: 10.1109/ICMIP.2017.26
- [30] ASAN A, ROBITZA W, MKWAWA I H, et al. Impact of Video Resolution Changes on QoE for Adaptive Video Streaming [C]//*2017 IEEE International Conference on Multimedia and Expo (ICME)*. Hong Kong, China, 2017: 499–504. DOI: 10.1109/ICME.2017.8019297
- [31] FRNDA J, VOZNAK M, SEVCIK L. Impact of Packet Loss and Delay Variation on the Quality of Real-Time Video Streaming [J]. *Telecommunication Systems*, 2016, 62(2): 265–275. DOI: 10.1007/s11235-015-0037-2
- [32] MAEDA Y, ISHIBASHI Y, FUKUSHIMA N. QoE Assessment of Sense of Presence in Networked Virtual Environment with Haptic and Auditory Senses: Influence of Network Delay [C]//*IEEE 3rd Global Conference on Consumer Electronics (GCCE)*. Tokyo, Japan, 2014: 679–683. DOI: 10.1109/GCCE.2014.7031181
- [33] NUNOME T, FURUKAWA K. The Effect of Bandwidth Allocation Methods on QoE of Multi-View Video and Audio IP Transmission [C]//*IEEE 22nd International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*. Lund, Sweden, 2017: 1–6. DOI: 10.1109/CAMAD.2017.8031625
- [34] BEGLUK T, HUSIC J B, BARAKOVIC S. Machine Learning-Based QoE Prediction for Video Streaming over LTE Network [C]//*17th International Symposium INFOTEH-JAHORINA (INFOTEH)*. East Sarajevo, Bosnia-Herzegovina, 2018: 1–5. DOI: 10.1109/INFOTEH.2018.8345519
- [35] GUTIÉRREZ J, PÉREZ P, JAUREGUIZAR F, et al. Subjective Assessment of the Impact of Transmission Errors in 3DTV Compared to HDTV [C]//*3DTV Conference: the True Vision—Capture, Transmission and Display of 3D Video (3DTV-CON)*. Antalya, Turkey, 2011: 1–4. DOI: 10.1109/3DTV.2011.5877209
- [36] BEYER J, MIRUCHNA V, MÖLLER S. Assessing the Impact of Display Size, Game Type, and Usage Context on Mobile Gaming QoE [C]//*Sixth International Workshop on Quality of Multimedia Experience (QoMEX)*. Singapore, Singapore, 2014: 69–70. DOI: 10.1109/QoMEX.2014.6982297
- [37] VUCIC D, SKORIN-KAPOV L. The Impact of Mobile Device Factors on QoE for Multi-Party Video Conferencing via WebRTC [C]//*13th International Conference on Telecommunications (ConTEL)*. Graz, Austria, 2015: 1–8. DOI: 10.1109/ConTEL.2015.7231206
- [38] JEGANATAN F, FRANCIS W, NUNOME T. QoE Assessment of Multi-View Video and Audio Simultaneous IP Transmission: The Effect of User Interfaces [C]//*International Conference on Information and Communication Technology Convergence (ICTC)*. Busan, South Korea, 2014: 466–471. DOI: 10.1109/ICTC.2014.6983182
- [39] EDSTROM J, CHEN D L, WANG J H, et al. Luminance-Adaptive Smart Video Storage System [C]//*IEEE International Symposium on Circuits and Systems (ISCAS)*. Montreal, Canada, 2016: 734–737. DOI: 10.1109/ISCAS.2016.7527345
- [40] TRIYASON T, KRATHU W. The Impact of Screen Size Toward QoE of Cloud-Based Virtual Desktop [J]. *Procedia Computer Science*, 2017, 111: 203–208. DOI: 10.1016/j.procs.2017.06.054
- [41] BARAKOVIC S, SKORIN-KAPOV L. Survey of Research on Quality of Experience Modelling for Web Browsing [J]. *Quality and User Experience*, 2017, 2: 6. DOI: 10.1007/s41233-017-0009-2
- [42] GUNTUKU S C, LIN W S, SCOTT M J, et al. Modelling the Influence of Personality and Culture on Affect and Enjoyment in Multimedia [C]//*International Conference on Affective Computing and Intelligent Interaction (ACII)*. Xi'an, China, 2015: 236–242. DOI: 10.1109/ACII.2015.7344577
- [43] MURRAY N, LEE B, QIAO Y S, et al. The Influence of Human Factors on Olfaction Based Multimedia Quality of Experience [C]//*Eighth International Conference on Quality of Multimedia Experience (QoMEX)*. Lisbon, Portugal, 2016: 1–6. DOI: 10.1109/QoMEX.2016.7498975
- [44] MURRAY N, MUNTEAN G M, QIAO Y S, et al. Modeling User Quality of Experience of Olfaction-Enhanced Multimedia [J]. *IEEE Transactions on Broadcasting*, 2018, 64(2): 539–551. DOI: 10.1109/tbc.2018.2825297
- [45] SONG J R, YANG F Z, ZHOU Y C, et al. QoE Evaluation of Multimedia Services Based on Audiovisual Quality and User Interest [J]. *IEEE Transactions on Multimedia*, 2016, 18(3): 444–457. DOI: 10.1109/tmm.2016.2520090
- [46] EYNARD R, PALLOT M, CHRISTMANN O, et al. Impact of Verbal Communication on User Experience in 3D Immersive Virtual Environments [C]//*IEEE International Conference on Engineering, Technology and Innovation/International Technology Management Conference (ICE/ITMC)*. Belfast, UK, 2015: 1–8. DOI: 10.1109/ICE.2015.7438679

### Biographies

**CHEN Jinling** received her bachelor of engineering from Fujian Agriculture and Forestry University, China in 2016. She is a master student at the College of Physics and Information Engineering, Fuzhou University, China. Her research interests include image/video processing and visual quality assessment.

**XU Yiwen** (xu.yiwen@fzu.edu.cn) received his Ph.D. degree from Department of Electronic Engineering, Xiamen University, China in 2012. He has been an associate professor with the College of Physics and Information Engineering, Fuzhou University, China, since 2013. His research interests lie in multimedia information processing, video codec and transmission, and video quality assessment.

**LIU Yisang** received her bachelor of engineering from Fuzhou University, China in 2016. She is a master student at the College of Physics and Information Engineering, Fuzhou University. Her research interests lie in multimedia streaming, 3D video quality assessment, and 3D video streaming.

**HUANG Huiwen** received her bachelor of engineering from Fuzhou University, China in 2016. She is a master student at the College of Physics and Information Engineering, Fuzhou University. Her research interests lie in image/video processing and 360 degree video streaming.

**ZHUANG Zhongwen** received his bachelor of engineering from Fuzhou University, China in 2016. He is a master student at the College of Physics and Information Engineering, Fuzhou University. His research interest is video coding.



# Visual Attention Modeling in Compressed Domain: From Image Saliency Detection to Video Saliency Detection

FANG Yuming and ZHANG Xiaoqiang

(Jiangxi University of Finance and Economics, Nanchang, Jiangxi 330032, China)

**Abstract:** Saliency detection models, which are used to extract salient regions in visual scenes, are widely used in various multimedia processing applications. It has attracted much attention in the area of computer vision over the past decades. Since most images or videos over the Internet are stored in compressed domains such as images in JPEG format and videos in MPEG2 format, H.264 format, and MPEG4 Visual format, many saliency detection models have been proposed in the compressed domain recently. We provide a review of our works on saliency detection models in the compressed domain in this paper. Besides, we introduce some commonly used fusion strategies to combine spatial saliency map and temporal saliency map to compute the final video saliency map.

**Keywords:** saliency detection; computer vision; compressed domain; visual attention; fusion strategy

DOI: 10.12142/ZTECOM.201901006

<http://kns.cnki.net/kcms/detail/34.1294.TN.20190314.1717.004.html>, published online March 14, 2019

Manuscript received: 2018-07-19

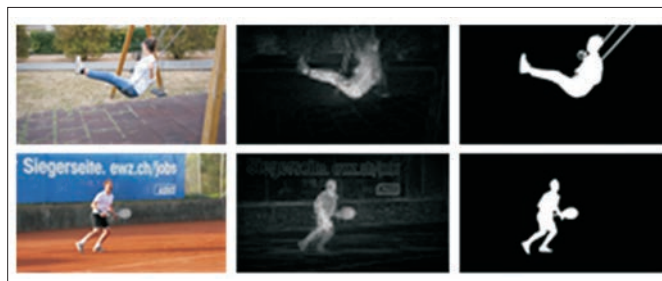
## 1 Introduction

The human visual system (HVS) has limited capacity and cannot process everything that falls onto the retina [1]. Visual attention would selectively bring important information into focus while filtering other parts to reduce the complexity of scene analysis. Saliency detection model, which simulates visual attention mechanism, could identify regions of interest in images or videos. There are two visual attention mechanisms: bottom-up and top-down approaches. The bottom-up attention [2] is determined by characteristics of a visual scene (stimulus-driven), while the top-down attention [3] is determined by cognitive phenomena like expectations, current goals, and knowledge (goal-driven). Saliency estimation from one computational model is shown in Fig. 1, where the brighter the region is, the more salient it is.

Currently, saliency detection has been applied to the important preprocessing step in various multimedia processing applications, such as object tracking [4], [5], image retargeting [6], object detection [7], object recognition [8], person re-identification [9], image compression [10], quality assessment [11]–

[13], abstraction [14], segmentation [15], and so on.

Saliency detection models can be divided into pixel-domain models and compressed-domain models. Early research on saliency detection mostly focuses on feature extraction in the pixel domain [16]–[34]. However, most images or videos over the Internet are basically stored in the compressed domain. For example, images over the Internet are stored in Joint Photographic Experts Group (JPEG) format, while videos are stored in H.264 and Moving Picture Experts Group (MPEG2) format.



▲ Figure 1. Saliency estimation results [16] on the public database Densely Annotated Video Segmentation (DAVIS) [17]. From the first column to the last column: original images, saliency maps, and ground truth maps.

Compressed images or videos are widely used in various multimedia applications over the Internet, because they can reduce storage space and improve transmission efficiency. The current saliency detection models have to decompress the compressed images or videos into the pixel domain for feature extraction, which is time consuming. To avoid the process, some saliency detection models are proposed in the compressed domain [6], [35]–[43].

As a pioneer, Itti et al. [18] proposed a conceptually computational model for saliency detection based on multiscale image features including intensity, color, and orientation. Harel et al. [19] introduced a bottom-up visual saliency model (GBVS) with the new definition of dissimilarity to extract saliency information. After that, Yang et al. [20] computed visual saliency by ranking the similarity of the image elements (pixels or regions) with foreground cues or background cues via graph-based manifold ranking. However, many graph-based models [19], [20] heavily depend to the performance of the superpixel segmentation preprocessing. Therefore, Li et al. [21] introduced a saliency detection approach that considers the advantages of both region-based features and image details by the regularized random walk ranking. Tu et al. [24] proposed a method for measuring the image boundary efficiently based on the minimum spanning tree. The method established by Qin et al. [22] calculates saliency by Cellular Automata—a dynamic evolution model, which can obtain the relevance of similar regions. Tong et al. [23] exploited both weak and strong models for saliency detection by developing a bootstrap learning algorithm. Recently, deep learning based methods become more and more popular in saliency detection. Wang et al. [25] estimated saliency by integrating local features and global features extracted by two deep neural networks, respectively. Based on auto-encoder neural network, Zhang [26] presented a saliency detection model by learning uncertain convolutional features.

Recently, some video saliency detection models were also explored [28]–[31] in the pixel domain. Kim et al. [28] introduced the approach of random walk with restart to detect spatially and temporally salient regions. They calculated spatiotemporal saliency by finding the steady-state distribution of the walker. In [29], temporal background priors are combined with spatial background priors to generate spatiotemporal background priors. Then, saliency estimation is conducted by a dual-graph based structure using spatiotemporal background priors. A spectral foreground extraction algorithm, Quantum Cuts (QCUT), is applied to estimate the saliency probability of regions [30]. Chen et al. [31] designed a video saliency detection model based on the spatiotemporal saliency fusion and low-rank coherency guided saliency diffusion. In [44], Li et al. proposed an unsupervised approach for video saliency object detection by using stacked auto-encoder neural network. In that approach, three saliency cues including pixel, superpixel, and object levels are extracted based on the algorithms of [45], [46], and [47]. Then the three saliency cues are fed into

stacked auto-encoders to infer a saliency score for each pixel.

The models mentioned above are all saliency detection models in the pixel domain. Recently, there have been some works explored on saliency detection in the compressed domain. Muthuswamy et al. [35] used discrete cosine transform (DCT) coefficients [6] and motion vectors [48] as features for MPEG2 video saliency detection. Khatoonabadi et al. [36] proposed a new feature, operational block description length (OBDL), as a measure of saliency. The OBDL represents the minimum number of bits required to encode a given video block under a certain distortion criterion [36]. In [37], Khatoonabadi et al. introduced two video features called Motion Vector Entropy and Smoothed Residual Norm extracted from the compressed video bitstream. Using the statistics of these two features in videos, they proposed a visual saliency detection model for compressed video. Two compressed domain features called residual DCT coefficient norms and operational block description are extracted from video bitstream [38], [39]. Then Li et al. [38], [39] used a fusion algorithm whose fusion coefficients vary with quantization parameters to fuse the two feature maps for saliency estimation. Xu et al. [40] first extracted High Efficiency Video Coding (HEVC) features in the HEVC domain and then those features are integrated by the learned support vector machine for video saliency detection. Jian et al. [41] introduced a saliency detection model by extracting three features including Quaternionic Distance Based Weber Descriptor (QDWD), pattern distinctness, and local contrast. By exploiting MPEG4 AVC compression principles, Ammar et al. [42] calculated the intensity, color, orientation, and motion feature maps by extracting the energy of luma coefficients, energy of chroma coefficients, gradient of the prediction modes, and amplitude of motion vectors. Finally, spatiotemporal saliency map is obtained by an average fusion algorithm. We proposed a saliency detection model in the compressed domain for images [6] and video [43].

The remaining of this paper is organized as follows. Section 2 describes our works in the compressed domain. Section 3 compares our fusion strategies of spatial and temporal saliency with those from other existing fusion strategies. The final Section 4 concludes the paper.

## 2 Compressed-Domain Visual Saliency Models

There are two works in computational modeling of visual attention in the compressed domain [6], [43]. In [6], the saliency detection model in compressed domain is built for 2D images, while the model is established for visual saliency modeling of video sequences in [43].

### 2.1 Saliency Detection Model for Compressed-Domain Image

We proposed a saliency detection model for images in com-

pressed domain [6]. The general framework of the proposed model is shown in **Fig. 2**. Three kinds of features (including intensity, color, and texture features) are firstly extracted from DCT coefficients. Then four visual saliency maps (one intensity saliency map, two color saliency maps, and one texture saliency map) are estimated by calculating feature contrast based on small DCT blocks weighted by a Gaussian model. Finally, by using coherent normalization-based fusion method to fuse these saliency maps, the final saliency map is obtained. Some saliency detection results of the proposed model [6] are shown in **Fig. 3**.

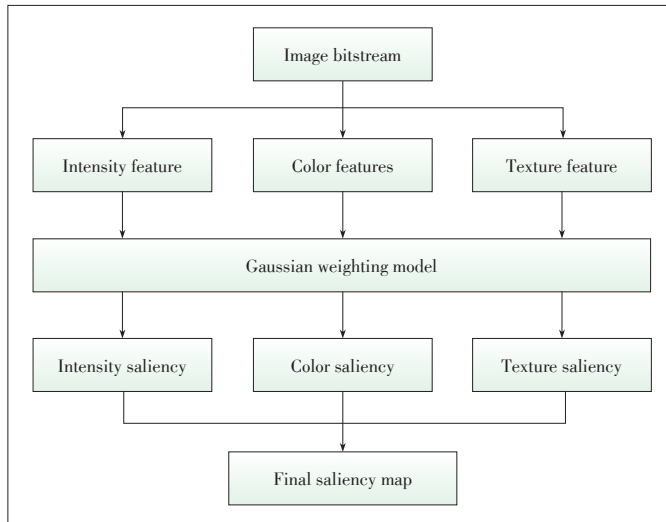
This work [6] in saliency detection mainly includes two contributions: the first one is how to extract features (intensity, color, and texture features) directly from the JPEG bitstream; the second is to design a new computational model of visual attention based on DCT blocks in the compressed domain. The details of the model are described as follows.

#### (1) Feature Extraction from the JPEG Bitstream

The color space of the input JPEG images is converted from RGB color space to YCbCr color space. YCbCr color space could be used to extract the three kinds of features mentioned above. Specifically, L channel contains the intensity and texture information while Cb and Cr channels contain color information. Each channel is divided into  $8 \times 8$  blocks, and the DCT is carried out for each small block. DCT coefficients in each block include the DC coefficient and AC coefficients. Please note that DC coefficient is a measure of the average energy of this block, while the remaining 63 AC coefficients represent high frequency information of this block. Therefore, we could use DC coefficient in L channel to extract intensity feature  $L$ . DC coefficients in Cb and Cr channels are used to extract two color features ( $C_1$  and  $C_2$ ). AC coefficients in L channel are used to extract texture feature  $T$ .

#### (2) Saliency Estimation in the Compressed Domain

Four saliency maps (one intensity saliency map, two color sa-



▲ **Figure 2.** The framework of the model proposed in [6].



▲ **Figure 3.** Saliency estimation results [6] on the public database in [49]. The first row is original images, while the second row represents saliency maps.

liency map, and one texture saliency map) are computed by feature contrast based on DCT blocks. The saliency value of each DCT block in each feature map is determined by two factors, including the block differences and weights between this block and all other blocks of the input image. Intensity and color feature differences of each DCT block are calculated by L1-norm distance, while texture difference of each DCT block is estimated by Hausdorff distance. The saliency value for each block is proportional to the block difference. The human eye is more sensitive about the differences between the current block and nearer blocks compared with the relatively distant area. A Gaussian model is thus used to weight the block differences for saliency detection. The saliency map for the  $n$ th feature can be calculated as follows:

$$S_i^n = \sum_{j \neq i} \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{d_{ij}^2}{2\sigma^2}} D_{ij}^n, \quad (1)$$

where  $d_{ij}$  represents the Euclidean distance between DCT blocks  $i$  and  $j$ ;  $n \in \{L, C_1, C_2, T\}$  and  $D_{ij}$  is DCT block difference;  $\sigma$  is a parameter of the Gaussian model. In the study [6],  $\sigma$  is set to 5.

According to Eq. (1), different saliency maps are calculated based on different features. These saliency maps include one intensity saliency map, two color saliency maps, and one texture saliency map. The final saliency map  $S$  for a given JPEG image can be calculated by fusing these four saliency maps. In [6], the coherent normalization-based fusion method is used to combine these four saliency maps as follows:

$$S = \beta \sum N(k) + \gamma \prod N(k), \quad (2)$$

where  $\beta$  and  $\gamma$  are parameters determining the weights for each components. In [6], the two parameters are both set to  $1/5$ .  $N$  is the normalization operation;  $k \in \{S^n\}$ .

## 2.2 Saliency Detection Model for Compressed-Domain Video

Similar with images, videos over the Internet are almost stored in the compressed domain such as H.264 and MPEG2.



In the study [43], a video saliency detection model is proposed based on feature contrast in the compressed domain.

The framework of the model [43] is shown in **Fig. 4**. That video saliency detection model could be roughly divided into two stages, including spatial saliency and temporal saliency estimation and fusion of these two kinds of visual saliency. Specifically, spatial saliency is calculated based on the three features (including luminance, color, and texture) extracted from the video bitstream. Temporal saliency is calculated based on the motion features extracted from the motion vectors in video bitstream. In the second stage, based on a new fusion method of parameterized normalization, sum and product (PNSP), the final saliency map for the video frame is calculated. Some saliency detection results of the proposed model [43] are shown in **Fig. 5**. The details of the model will be described as follows.

#### (1) Feature Extraction in the Video Bitstream

In MPEG4 advanced simple profile (ASP) video, there are two kinds of predicted frames: P frames use motion compensated prediction from a past reference frame, while B frames are bidirectionally predictive-coded by using motion compensated prediction from a past and/or a future reference frame. As there are two kinds of frames, there are two kinds of ways to calculate the motion feature. The motion vector  $MV$  is used to represent the motion feature for P frames. The motion feature for B frames can be calculated by both motion prediction from the past and future reference frames. Assume the motion com-

pensated prediction from the past reference and the future reference frames are  $MV_1$  and  $MV_2$ . The motion feature  $V$  of B frames is computed as follows:

$$V = MV_1 - MV_2. \quad (3)$$

Please note that no matter what kinds of frames are used, motion features are computed based on DCT blocks. And the motion feature of P/B frames can be obtained as  $V$ . For spatial features include intensity, color, and texture, they can be extracted as [6].

#### (2) Saliency Estimation in the Compressed Domain

Based on the motion feature  $V$ , the feature map of each video frame is computed as follows:

$$S_i^v = \sum_{j \neq i} w_{ij} D_{ij}^v, \quad (4)$$

$$w_{ij} = \frac{1}{\sigma_v \sqrt{2\pi}} e^{-\frac{d_{ij}^2}{2\sigma_v^2}}, \quad (5)$$

where  $S_i^v$  represents temporal saliency value of the  $i$ th DCT block in the motion feature map;  $D_{ij}^v$  is the motion feature difference between DCT blocks  $i$  and  $j$ ;  $\sigma_v$  is a parameter of the Gaussian model. The spatial saliency map  $S^s$  is calculated by linearly combining the four spatial feature maps from intensity, color, and texture features ( $L, C_1, C_2, T$ ).

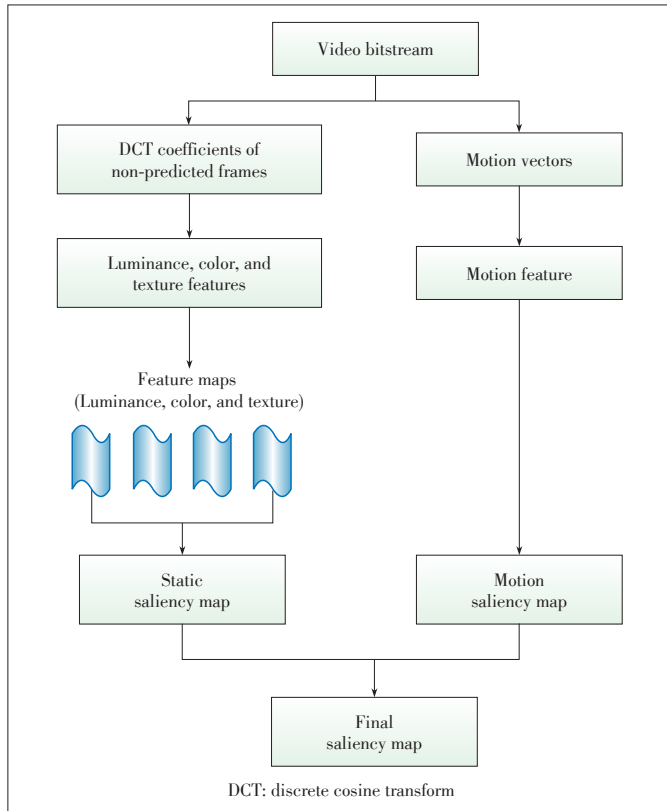
In [43], based on the characteristics of the spatial saliency map and temporal saliency map, a new fusion method called PNSP is proposed. The final saliency map for video frame is calculated as follows:

$$S^f = \beta_1 S^s + \beta_2 S^v + \beta_3 S^s S^v, \quad (6)$$

where  $S^f$  denotes the final saliency map for the video frame;  $\beta_1$ ,  $\beta_2$ , and  $\beta_3$  are the parameters determining the weights of each component;  $S^s$  is the spatial saliency map and  $S^v$  is the temporal saliency map.

### 3 Spatiotemporal Weighting Strategy

Here, we introduce our works on spatiotemporal weighting strategy [16], [43]. In [16], based on Gestalt theory, the spatial



▲ Figure 4. The framework of the model proposed by [43].



▲ Figure 5. Saliency estimation results [43] on the public database Densely Annotated Video Segmentation (DAVIS) [17]. The first row is original images, while the second row represents saliency maps.



PNSP is designed to combine the spatial and temporal saliency to obtain the final saliency map. We also introduce several common fusion algorithms in [50] for integrating spatial saliency and temporal saliency map.

### 3.1 Common Fusion Approaches

#### (1) Normalization and Sum (NS)

The most simple and direct method for fusing spatial saliency and temporal saliency is to normalize these two salient maps to the same dynamic range (between 0 and 1) and then sum the two maps to get the final saliency map as follows:

$$S = \sum_n N(S_n), \quad (7)$$

where  $S$  is the final saliency map;  $n \in \{1, 2\}$  and  $S_n$  is the spatial saliency map or temporal saliency map.

#### (2) Normalization and Maximum (NM)

The fusion algorithm tries to normalize spatial saliency map and temporal saliency map to the same dynamic range and then uses the maximum value as the final saliency value at each location.

$$S = \max_n N(S_n), \quad (8)$$

where  $\max$  is the maximum operator.

#### (3) Normalization and Product (NP)

Compared to NS and NM methods, the summation and maximum are replaced by the product operator in NP.

$$S = \prod_n N(S_n). \quad (9)$$

### 3.2 The Fusion Approach Based on Uncertainty Weighting

Compared with image saliency detection, video saliency detection is a more challenging problem due to its complex background and utilization of motion information. So far, only a few video saliency detection models have been proposed [51]–[53]. In [16], a novel method is proposed to estimate video saliency by using Gestalt theory and uncertainty weighting.

The algorithm [16] could be divided into two main stages including the spatial and temporal saliency estimation stage and the fusion stage of the two saliency maps. Spatial saliency is calculated by extracted spatial features including luminance, color, and texture features from a given video frame using DCT coefficients [6]. Temporal saliency can be measured based on a psychological study of human visual speed perception [54]. Based on uncertainty weighting strategy, we can fuse the spatial saliency map and temporal saliency map for obtaining the final saliency map. Spatial uncertainty estimation is conceptually rooted in the Gestalt theory including the law of proximity and the law of continuity [55], [56]. Temporal uncertainty estimation is calculated based on the psychovisual studies in [54]. Some saliency detection results of the proposed model [16] are shown in Fig. 1.

The proximity law of Gestalt theory states that elements which are close to each other tend to be perceived as a group, while the continuity law of Gestalt theory indicates that elements which are connected to each other tend to be perceived as a group. These two laws can be applied to saliency detection as follows: first, the spatial location which is closer to the saliency center in an image is more likely to be a salient location; second, a spatial location which is more connected to other saliency regions is more likely to be a salient location. Then the spatial uncertainty for each pixel in the spatial saliency map is calculated as follows:

$$U^s = U^d + U^c, \quad (10)$$

where  $U^s$  is the spatial uncertainty map;  $U^d$  is the probability of a pixel being salient given its distance from saliency center;  $U^c$  represents the probability of a pixel being salient given its connectedness to other salient pixels.

When the background motion is very large in the video, or the local contrast increases, the system cannot detect motion of the object accurately. This temporal uncertainty evaluation  $U^t$  is conducted based on the psychovisual studies in [54]. Therefore, the spatial and temporal saliency map can be integrated into spatiotemporal saliency map of the given video sequence by using these two uncertainty weighting map as follows:

$$S_{sp} = \frac{U^t S_s + U^s S_t}{U^t + U^s}, \quad (11)$$

where  $S_{sp}$  is the spatiotemporal saliency map;  $U^t$  is the temporal uncertainty map;  $U^s$  represents the spatial uncertainty map;  $S_s$  is the spatial saliency map calculated by DCT coefficients;  $S_t$  is the temporal saliency map calculated by optical flow algorithm.

### 3.3 The Fusion approach Based on PNSP Weighting

Another fusion method [43] has already been described in Section 2.2. And we can find that this fusion method is the combination of the NS method and NP method mentioned above. If the salient regions have low spatial saliency value with high temporal saliency value, the NS method can highlight the saliency by summation operation. If the non-salient regions have low spatial saliency value with high temporal saliency value, the NP method can suppress the saliency by product operation. Therefore, this algorithm can combine the advantages of these two algorithms.

## 4 Conclusions

In this paper, we review some works in the pixel-domain and compressed-domain. We first attempt to provide a comprehensive description of two compressed-domain saliency detection models. These two models are designed to handle with different tasks including compressed-domain 2D images and compressed-domain 2D video saliency detection. The difficulty of

the tasks continues to increase since increasing complexity of scene. Then we exhaustively review our two fusion strategies of spatial saliency map and temporal saliency map. The PNSP fusion algorithm considers the advantages of NS algorithm and NP algorithm. Another fusion algorithm is designed based on proximity law and proximity law of Gestalt theory.

Specifically, According to image saliency detection or video saliency detection, feature contrast for each block is calculated by the differences between the features of this block and other blocks in the whole image. In the future, we hope that we could propose more effective methods to handle the computational modeling for visual attention.

## References

- [1] JAMES W. The Principles of Psychology [M]. England: Read Books Ltd, 2013
- [2] NOTHDURFT H C. Saliency of Feature Contrast [M]//NOTHDURFT H C. eds. Neurobiology of Attention. Amsterdam, Netherlands: Elsevier, 2005: 233–239. DOI: 10.1016/b978-012375731-9/50042-2
- [3] ITTI L, KOCH C. Computational Modelling of Visual Attention [J]. *Nature Reviews Neuroscience*, 2001, 2(3): 194–203. DOI: 10.1038/35058500
- [4] MAHADEVAN V, VASCONCELOS N. Saliency-Based Discriminant Tracking [C]//IEEE Conference on Computer Vision and Pattern Recognition, Miami, USA, 2009: 1007–1013. DOI: 10.1109/CVPR.2009.5206573
- [5] MA C, MIAO Z J, ZHANG X P, et al. A Saliency Prior Context Model for Real-Time Object Tracking [J]. *IEEE Transactions on Multimedia*, 2017, 19(11): 2415–2424. DOI: 10.1109/tmm.2017.2694219
- [6] FANG Y M, CHEN Z Z, LIN W S, et al. Saliency Detection in the Compressed Domain for Adaptive Image Retargeting [J]. *IEEE Transactions on Image Processing*, 2012, 21(9): 3888–3901. DOI: 10.1109/tip.2012.2199126
- [7] GUO M W, ZHAO Y Z, ZHANG C B, et al. Fast Object Detection Based on Selective Visual Attention [J]. *Neurocomputing*, 2014, 144: 184–197. DOI: 10.1016/j.neucom.2014.04.054
- [8] REN Z X, GAO S H, CHIA L T, et al. Region-Based Saliency Detection and Its Application in Object Recognition [J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2014, 24(5): 769–779. DOI: 10.1109/tcsvt.2013.2280096
- [9] ZHAO R, WANLI O Y, WANG X G. Unsupervised Saliency Learning for Person Re-Identification [C]//IEEE Conference on Computer Vision and Pattern Recognition, Portland, USA, 2013: 3586–3593. DOI: 10.1109/CVPR.2013.460
- [10] HOU Y H, WANG P C, XIANG W, et al. A Novel Rate Control Algorithm for Video Coding Based on fuzzy-PID Controller [J]. *Signal, Image and Video Processing*, 2015, 9(4): 875–884. DOI: 10.1007/s11760-013-0518-2
- [11] CULIBRK D, MIRKOVIC M, ZLOKOLICA V, et al. Salient Motion Features for Video Quality Assessment [J]. *IEEE Transactions on Image Processing*, 2011, 20(4): 948–958. DOI: 10.1109/tip.2010.2080279
- [12] LIU H T, HEYNDERICKX I. Visual Attention in Objective Image Quality Assessment: Based on Eye-Tracking Data [J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2011, 21(7): 971–982. DOI: 10.1109/tcsvt.2011.2133770
- [13] FENG X, LIU T, YANG D, et al. Saliency Inspired Full-Reference Quality Metrics for Packet-Loss-Impaired Video [J]. *IEEE Transactions on Broadcasting*, 2011, 57(1): 81–88. DOI: 10.1109/tbc.2010.2092150
- [14] JI Q G, FANG Z D, XIE Z H, et al. Video Abstraction Based on the Visual Attention Model and Online Clustering [J]. *Signal Processing: Image Communication*, 2013, 28(3): 241–253. DOI: 10.1016/j.image.2012.11.008
- [15] MISHRA A K, ALOIMONOS Y, CHEONG L F, et al. Active Visual Segmentation [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012, 34(4): 639–653. DOI: 10.1109/tpami.2011.171
- [16] FANG Y M, WANG Z, LIN W S, et al. Video Saliency Incorporating Spatiotemporal Cues and Uncertainty Weighting [J]. *IEEE Transactions on Image Processing*, 2014, 23(9): 3910–3921. DOI: 10.1109/tip.2014.2336549
- [17] PERAZZI F, PONT-TUSET J, MCWILLIAMS B, et al. A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation [C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, USA, 2016: 724–732. DOI: 10.1109/CVPR.2016.85
- [18] ITTI L, KOCH C, NIEBUR E. A Model of Saliency-Based Visual Attention for Rapid Scene Analysis [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1998, 20(11): 1254–1259. DOI: 10.1109/34.730558
- [19] HAREL J, KOCH C, PERONA P. Graph-Based Visual Saliency [M]//Advances in Neural Information Processing Systems 19. Cambridge, USA: The MIT Press, 2007: 545–552. DOI: 10.7551/mitpress/7503.003.0073
- [20] YANG C, ZHANG L H, LU H C, et al. Saliency Detection Via Graph-Based Manifold Ranking [C]//IEEE Conference on Computer Vision and Pattern Recognition, Portland, USA, 2013: 3166–3173. DOI: 10.1109/CVPR.2013.407
- [21] LI C Y, YUAN Y C, CAI W D, et al. Robust Saliency Detection Via Regularized Random Walks Ranking [C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, USA, 2015: 2710–2717. DOI: 10.1109/CVPR.2015.7298887
- [22] QIN Y, LU H C, XU Y Q, et al. Saliency Detection Via Cellular Automata [C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, USA, 2015: 110–119. DOI: 10.1109/CVPR.2015.7298606
- [23] TONG N, LU H C, RUAN X, et al. Salient Object Detection Via Bootstrap Learning [C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, USA, 2015: 1884–1892. DOI: 10.1109/CVPR.2015.7298798
- [24] TU W C, HE S F, YANG Q X, et al. Real-Time Salient Object Detection with a Minimum Spanning Tree [C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, USA, 2016: 2334–2342. DOI: 10.1109/CVPR.2016.256
- [25] WANG L J, LU H C, RUAN X, et al. Deep Networks for Saliency Detection Via Local Estimation and Global Search [C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, USA, 2015: 3183–3192. DOI: 10.1109/CVPR.2015.7298938
- [26] ZHANG P P, WANG D, LU H C, et al. Learning Uncertain Convolutional Features for Accurate Saliency Detection [C]//IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017: 212–221. DOI: 10.1109/ICCV.2017.32
- [27] YUAN Y C, LI C Y, KIM J, et al. Reversion Correction and Regularized Random Walk Ranking for Saliency Detection [J]. *IEEE Transactions on Image Processing*, 2018, 27(3): 1311–1322. DOI: 10.1109/tip.2017.2762422
- [28] KIM H, KIM Y, SIM J Y, et al. Spatiotemporal Saliency Detection for Video Sequences Based on Random Walk with Restart [J]. *IEEE Transactions on Image Processing*, 2015, 24(8): 2552–2564. DOI: 10.1109/tip.2015.2425544
- [29] XI T, ZHAO W, WANG H, et al. Salient Object Detection with Spatiotemporal Background Priors for Video [J]. *IEEE Transactions on Image Processing*, 2017, 26(7): 3425–3436. DOI: 10.1109/tip.2016.2631900
- [30] AYTEKIN C, POSSEGGGER H, MAUTHNER T, et al. Spatiotemporal Saliency Estimation by Spectral Foreground Detection [J]. *IEEE Transactions on Multimedia*, 2018, 20(1): 82–95. DOI: 10.1109/tmm.2017.2713982
- [31] CHEN C, LI S, WANG Y G, et al. Video Saliency Detection Via Spatial-Temporal Fusion and Low-Rank Coherency Diffusion [J]. *IEEE Transactions on Image Processing*, 2017, 26(7): 3156–3170. DOI: 10.1109/tip.2017.2670143
- [32] FANG Y M, LIN W S, LEE B S, et al. Bottom-Up Saliency Detection Model Based on Human Visual Sensitivity and Amplitude Spectrum [J]. *IEEE Transactions on Multimedia*, 2012, 14(1): 187–198. DOI: 10.1109/tmm.2011.2169775
- [33] FANG Y M, WANG J L, NARWARIA M, et al. Saliency Detection for Stereoscopic Images [J]. *IEEE Transactions on Image Processing*, 2014, 23(6): 2625–2636. DOI: 10.1109/tip.2014.2305100
- [34] FANG Y M, ZHANG C, LI J, et al. Visual Attention Modeling for Stereoscopic Video: A Benchmark and Computational Model [J]. *IEEE Transactions on Image Processing*, 2017, 26(10): 4684–4696. DOI: 10.1109/tip.2017.2721112
- [35] MUTHUSWAMY K, RAJAN D. Salient Motion Detection in Compressed Domain [J]. *IEEE Signal Processing Letters*, 2013, 20(10): 996–999. DOI: 10.1109/lsp.2013.2277884
- [36] KHATOONABADI S H, VASCONCELOS N, BAJICI V, et al. How Many Bits does it Take for a Stimulus to be Salient? [C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, USA, 2015: 5501–5510. DOI: 10.1109/CVPR.2015.7299189
- [37] KHATOONABADI S H, BAJICI V, SHAN Y F. Compressed-Domain Correlates of Human Fixations in Dynamic Scenes [J]. *Multimedia Tools and Applications*, 2015, 74(22): 10057–10075. DOI: 10.1007/s11042-015-2802-3
- [38] LI Y J, LI Y S. A Fast and Efficient Saliency Detection Model in Video Compressed-Domain for Human Fixations Prediction [J]. *Multimedia Tools and Applications*, 2017, 76(24): 26273–26295. DOI: 10.1007/s11042-016-4118-3
- [39] LI Y J, LI Y S, LIU W J, et al. Human Fixation Detection Model in Video Compressed Domain Based on Markov Random Field [J]. *Journal of Electronic Imaging*, 2017, 26(1): 013008. DOI: 10.1117/1.jei.26.1.013008

- [40] XU M, JIANG L, SUN X Y, et al. Learning to Detect Video Saliency with HEVC Features [J]. *IEEE Transactions on Image Processing*, 2017, 26(1): 369–385. DOI: 10.1109/tip.2016.2628583
- [41] JIAN M W, QI Q, DONG J Y, et al. Integrating QDWD with Pattern Distinctness and Local Contrast for Underwater Saliency Detection [J]. *Journal of Visual Communication and Image Representation*, 2018, 53: 31–41. DOI: 10.1016/j.jvcir.2018.03.008
- [42] AMMAR M, MITREA M, HASNAOUI M, et al. MPEG-4 AVC Stream-Based Saliency Detection. Application to Robust Watermarking [J]. *Signal Processing: Image Communication*, 2018, 60: 116–130. DOI: 10.1016/j.image.2017.09.007
- [43] FANG Y M, LIN W S, CHEN Z Z, et al. A Video Saliency Detection Model in Compressed Domain [J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2014, 24(1): 27–38. DOI: 10.1109/tcsvt.2013.2273613
- [44] LI J, XIA C Q, CHEN X W. A Benchmark Dataset and Saliency-Guided Stacked Autoencoders for Video-Based Salient Object Detection [J]. *IEEE Transactions on Image Processing*, 2018, 27(1): 349–364. DOI: 10.1109/tip.2017.2762594
- [45] ZHANG J M, SCLAROFF S, LIN Z, et al. Minimum Barrier Salient Object Detection at 80 FPS [C]//*IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, 2015: 1404–1412. DOI: 10.1109/ICCV.2015.165
- [46] PENG H W, LI B, LING H B, et al. Salient Object Detection via Structured Matrix Decomposition [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, 39(4): 818–832. DOI: 10.1109/tpami.2016.2562626
- [47] LI J, LEVINE M D, AN X J, et al. Visual Saliency Based on Scale-Space Analysis in the Frequency Domain [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013, 35(4): 996–1010. DOI: 10.1109/tpami.2012.147
- [48] AGARWAL G, ANBU A, SINHA A. A Fast Algorithm to Find the Region-of-Interest in the Compressed MPEG Domain [C]//*International Conference on Multimedia and Expo. (ICME'03)*, Baltimore, USA, 2003: 133–136. DOI: 10.1109/ICME.2003.1221571
- [49] ACHANTA R, HEMAMI S, ESTRADA F, et al. Frequency-Tuned Salient Region Detection [C]//*IEEE Conference on Computer Vision and Pattern Recognition*, Miami, USA, 2009: 1597–1604. DOI: 10.1109/CVPR.2009.5206596
- [50] CHAMARET C, CHEVET J C, LE MEUR O. Spatio-Temporal Combination of Saliency Maps and Eye-Tracking Assessment of Different Strategies [C]//*IEEE International Conference on Image Processing*, Hong Kong, China, 2010: 1077–1080. DOI: 10.1109/ICIP.2010.5651381
- [51] GUO C L, ZHANG L M. A Novel Multiresolution Spatiotemporal Saliency Detection Model and Its Applications in Image and Video Compression [J]. *IEEE Transactions on Image Processing*, 2010, 19(1): 185–198. DOI: 10.1109/tip.2009.2030969
- [52] LE MEUR O, LE CALLET P, BARBA D. Predicting Visual Fixations on Video Based on Low-Level Visual Features [J]. *Vision Research*, 2007, 47(19): 2483–2498. DOI: 10.1016/j.visres.2007.06.015
- [53] MAHADEVAN V, VASCONCELOS N. Spatiotemporal Saliency in Dynamic Scenes [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010, 32(1): 171–177. DOI: 10.1109/tpami.2009.112
- [54] STOCKER A A, SIMONCELLI E P. Noise Characteristics and Prior Expectations in Human Visual Speed Perception [J]. *Nature Neuroscience*, 2006, 9(4): 578–585. DOI: 10.1038/nn1669
- [55] BANERJEE J C. Gestalt Theory of Perception (M)//*Encyclopaedic Dictionary of Psychological Terms*. New Delhi, India: MD Publications Pvt. Ltd., 1994: 107–109
- [56] STEVENSON H. Emergence: The Gestalt Approach to Change [EB/OL]. (2012). <http://www.clevelandconsultinggroup.com/articles/emergence-gestalt-approach-to-change.php>

### Biographies

**FANG Yuming** (fa0001ng@e.ntu.edu.sg) received his Ph.D. degree from Nanyang Technological University, Singapore, M.S. degree from Beijing University of Technology, China, and B.E. degree from Sichuan University, China. Currently, he is a professor in the School of Information Management, Jiangxi University of Finance and Economics, China. He serves as an associate editor of *IEEE Access* and is on the editorial board of *Signal Processing: Image Communication*. His research interests include visual attention modeling, visual quality assessment, image retargeting, computer vision, 3D image/video processing, etc.

**ZHANG Xiaoqiang** is currently pursuing the master's degree with the School of Information Technology, Jiangxi University of Finance and Economics, China. His research interests include saliency detection, computer vision, machine learning, and deep learning.

# Perceptual Quality Assessment of Omnidirectional Images: Subjective Experiment and Objective Model Evaluation



DUAN Huiyu, ZHAI Guangtao, MIN Xiongkuo, ZHU Yucheng, FANG Yi, and YANG Xiaokang

(Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, Shanghai 200240, China)

**Abstract:** Virtual reality (VR) environment can provide immersive experience to viewers. Under the VR environment, providing a good quality of experience is extremely important. Therefore, in this paper, we present an image quality assessment (IQA) study on omnidirectional images. We first build an omnidirectional IQA (OIQA) database, including 16 source images with their corresponding 320 distorted images. We add four commonly encountered distortions. These distortions are JPEG compression, JPEG2000 compression, Gaussian blur, and Gaussian noise. Then we conduct a subjective quality evaluation study in the VR environment based on the OIQA database. Considering that visual attention is more important in VR environment, head and eye movement data are also tracked and collected during the quality rating experiments. The 16 raw and their corresponding distorted images, subjective quality assessment scores, and the head-orientation data and eye-gaze data together constitute the OIQA database. Based on the OIQA database, we test some state-of-the-art full-reference IQA (FR-IQA) measures on equirectangular format or cubic format omnidirectional images. The results show that applying FR-IQA metrics on cubic format omnidirectional images could improve their performance. The performance of some FR-IQA metrics combining the saliency weight of three different types are also tested based on our database. Some new phenomena different from traditional IQA are observed.

**Keywords:** perceptual quality assessment; omnidirectional images; subjective experiment; objective model evaluation; visual saliency

DOI: 10.12142/ZTECOM.201901007

<http://kns.cnki.net/kcms/detail/34.1294>.

TN.20190313.0855.002.html, published online March 13, 2019

Manuscript received: 2019-01-16

## 1 Introduction

Omnidirectional content could provide observers with immersive perception with the help of Head-Mounted Displays (HMDs). As an important component of virtual reality (VR), natural immersive videos provide the viewers with real-world scenes. The omnidirectional visual experience makes the user experience more immersive compared to traditional VR content generated by computer-aided 3D modeling. Therefore, we mainly consider natural immersive content, i.e., omnidirectional images, in this paper.

Because of the immersive experience providing by VRHMD, it is exciting to experience omnidirectional contents. However, due to the limitation of the photographic apparatus, transmis-

sion bandwidth, and display devices, etc, the content viewed by observers usually cannot live up to the expectation. As a consequence, it is significant to study the quality of experience (QoE) in VR environments. Many traditional image quality assessment (IQA) databases have been constructed by researchers, such as Live Image Quality Assessment Database (LIVE) [1], TID2008—a database for evaluation of full-reference visual quality assessment metrics [2], categorical image quality (CSIQ) database [3], and quality assessment considering viewing distance and image resolution (VDID) [4], and some works related to assessing the quality of omnidirectional visual contents have also been done, such as [5]–[9]. However as far as we know, the databases relevant to omnidirectional image quality assessment are very few. And there is no database includ-



ing both subjective evaluation scores and eye movement data. So we have constructed one omnidirectional IQA (OIQA) database [10]. For traditional videos or images quality assessment, many efforts have been made on designing human visual system (HVS) based IQA metrics [11]–[15]. Visualizing immersive videos or omnidirectional (360-degree, equirectangular, VR) images [5] is different from traditional 2D videos or images. Observers are supposed to be in the central position of a sphere when visualizing immersive contents. The results in [5] and [16] illustrate that the view-port visualized by observers usually only occupies a portion of the whole omnidirectional images or videos. Because of the immersive experience, the visual attention of observers in the view-port of omnidirectional videos or images is different from the visual attention in plane 2D videos or images. Therefore, it is significant to study the new method to evaluate the quality of images and videos combining visual saliency in VR environment.

In this paper, we also explore the method of using human visual preferences to assess the quality of omnidirectional images. We test our ideas based on our OIQA database, which includes 16 raw reference omnidirectional images and their corresponding 320 degradation images under four kinds of distortion types. These distortions are JPEG [17] compression, JPEG2000 [18] compression, Gaussian blur, and Gaussian noise. The head and eye movement data are also collected in OIQA database. We first discuss the influence of intrinsic distortion of equirectangular projection. For comparison, the IQA metrics are tested on cubic images and we think cubic images have almost no such distortion. The performance of some FR-IQA metrics combining the saliency weight of three different types is also tested based on our database. Three different kinds of saliency maps include the global head movement characteristic map, global eye viewing preference map, and ground-truth visual saliency map.

The remainder of this paper is arranged as follows. We introduce the subjective omnidirectional IQA in Section 2. In Section 3, we evaluate several state-of-the-art FR-IQA models on the OIQA database and combine human visual preference in some IQA models. Some inspiring observations are proposed. We summarize and conclude the whole paper in Section 4.

## 2 Subjective Quality Assessment of Omnidirectional Images

The image collection and quality degradation processes are first introduced in this section. Next, we introduce our experimental methodology to conduct subjective quality rating and capture head or eye movement data. Finally, we process and analyze the collected visual attention data and subjective quality ratings and present some conclusions we have observed.

### 2.1 Original and Distorted Equirectangular Images

We collect 16 raw images which are captured by profession-

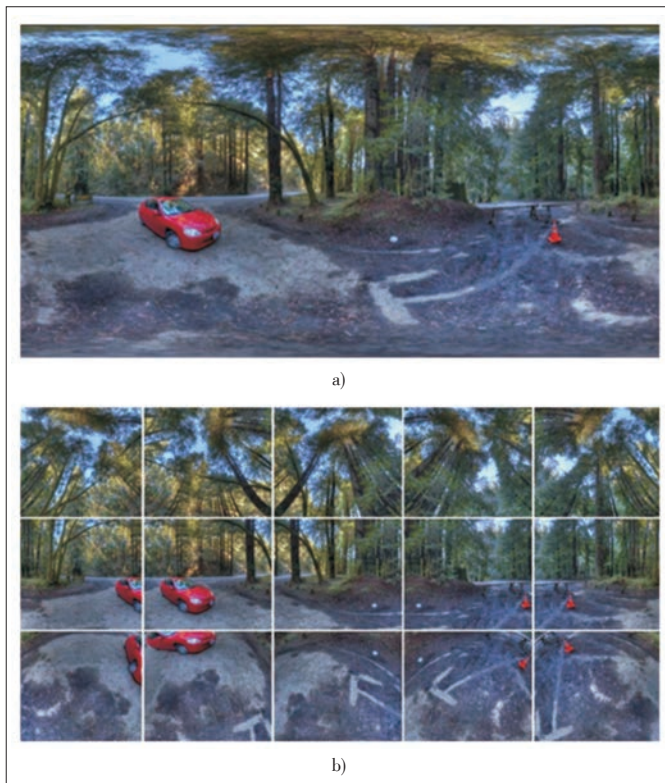
al photographers and available under Creative Commons (CC) copyright. The collected images are representative and have diversified textures. We show several sample raw images in **Fig. 1**. We zoomed in and carefully checked all raw images to avoid easily observed artifacts. This procedure can avoid the “intrinsic artifacts”. All of raw images have close resolutions which range from 11 332×5 666 to 13 320×6 660, and close perceptual quality. This procedure can reduce the influence of the original content’s quality on subjective ratings. We introduce four types of distortions to raw images, with five distortion levels for each type. The four types of distortions we introduced are JPEG compression, JPEG2000 compression, Gaussian blur, and white Gaussian noise (WGN), respectively, which are four commonly encountered distortions.

JPEG and JPEG2000 are the two commonly used compression methods to simulate the artifacts introduced during compression. In this paper, we introduced these two methods with five compression levels each to the raw images. The five levels are manually set to cover a wide perceptual quality range. Since omnidirectional contents are generally created, stored, compressed and transmitted in equirectangular format, we compress all images in equirectangular format directly and get degraded images. We also introduced another two commonly encountered distortions, which are Gaussian blur and WGN. In this paper, we mainly consider the blur and noise introduced during capturing. Omnidirectional images are usually captured by multiple cameras (e.g. camera array) and then stitched. To simulate the blur and noise introduced during capturing, the raw images are split into 15 small blocks. **Fig. 2a** represents the raw image. **Fig. 2b** represents the 15 split images of the raw image in Fig. 2a. These split images represent the scenes captured by each camera of the camera array. To simulate the distortions introduced at the sensor of each camera, we add Gaussian blur and WGN to 15 split images respectively. Then these split images with distortions added are projected back to one equirectangular image. Following these procedures, we



▲ **Figure 1.** Some source images in the omnidirectional image quality assessment database.





▲ Figure 2. One source image and 15 split images: a) the source image; b) 15 split images.

add the Gaussian blur and Gaussian noise to images more uniformly compared with adding distortions to equirectangular images directly. We also introduce five levels of blur and noise distortions to generate images with varying distortions. Thus we have 336 images in total in our OIQA database, including 16 raw images and their corresponding 320 distorted images.

## 2.2 Equipment and Software

The equipment we used to show the omnidirectional images to the subjects is HTC VIVE. This HMD has high-precision tracking ability and excellent graphics display technology. Specifically, the resolution of this display is 1 080×1 200 per eye and 2 160×1 200 combined with a field of view (FOV) of about 100 horizontal degrees and about 110 vertical degrees. The refresh rate of the HTC VIVE is 90 Hz. Additionally, the sensor of this HMD could provide head-orientation data at the same rate as the frame rate. In order to obtain eye movement data, a small eye-tracker named aGlass [19] is installed into the HMD. aGlass is an excellent VR eye-tracker with an error less than 0.5°. We also develop an interactive software based on Unity to display omnidirectional images and collect rating scores, head-orientation data and eye movement data.

## 2.3 Subjects

The total number of the subjects participated in our experiments was 20, including 5 females and 15 males. The age of

subjects ranged from 18 years to 30 years with an average of 24 years. All of the subjects reported normal or corrected-to-normal vision. As illustrated in [20], visually induced motion sickness (VIMS) in virtual reality environment could make the quality of experience (QoE) worse. And all of the subjects in our experiments reported that they did not have travel sickness.

## 2.4 Subjective Experiment Methodology

With the help of HTC VIVE, aGlass and the software, the experiments were conducted to obtain subjective rating score, head-orientation data and eye-movement data at the same time. Subjects were asked to seat in a rolling chair and be free to rotate the chair. This procedure is to ensure that the whole omnidirectional image could be visualized by observers. At the start of each experiment, the subjects were asked to calibrate the eye tracker which is installed in the HMD. Then, the visual attention data of the 16 raw omnidirectional images were collected. Each image was displayed for 20 seconds with a five-second gray screen displayed before showing the following omnidirectional image. In order to collect natural viewing visual attention data, all the subjects were asked to look around at least one circle in this step. Next, in order to make the subjects familiar with the distortions types and levels of the database, we conducted a training procedure. Finally, we conducted the formal quality rating experiment and collected rating scores. All images were displayed in a random order in this step. To avoid VIMS and fatigue, the subjects had enough rest time every 10 minutes during the experiment.

## 2.5 Data Processing and Analysis

Three types of data are collected, including raw subjective quality scores given by subjects, head movement data and eye gaze data, through the subjective experiment. In this section, we discuss the processing and analyzing procedure of these three kinds of data.

### 2.5.1 Subjective Quality Score Processing and Analysis

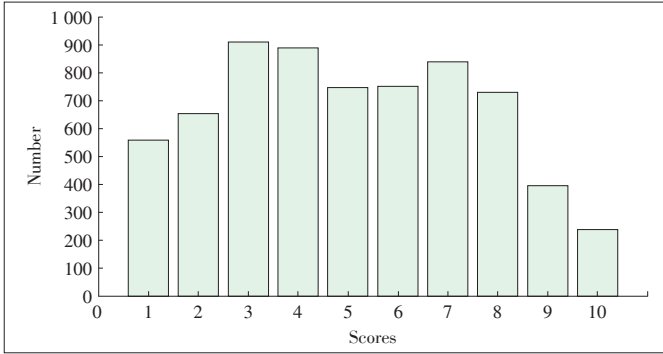
We first process the subjective rating scores of images. The mean opinion scores (MOS) are computed by the following formula:

$$MOS_j = \frac{\sum_{i=1}^N m_{ij}}{N}, \quad (1)$$

where  $N$  is the number of subjects and  $m_{ij}$  is the score assigned by subject  $i$  to image  $j$ . We also use the  $3\sigma$  principle to remove the outliers, which are scores far away from the average value. Fig. 3 illustrates the histogram of the distribution of subjective quality scores. Obviously, the subjective rating scores are distributed across all perceptual quality range.

### 2.5.2 Visual Attention Data Processing and Analysis

Head-orientation and eye-movement data within 20 seconds of 16 raw images were collected. To make the data more intuit-



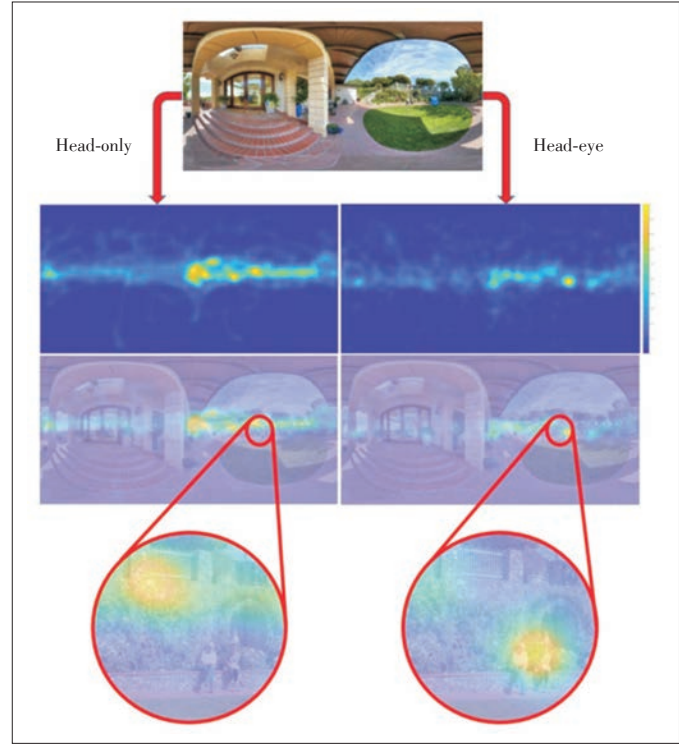
▲ Figure 3. Histogram of the subjective quality scores.

tive, we projected the view direction data and the eye-movement data from the 3D sphere space to the 2D equirectangular image. Head-only (view direction centered) saliency maps and head-eye saliency maps were created using the view direction information and the eye movement information respectively. We followed the method in [21] to get the saliency map. We applied a Gaussian filter of  $3.34^\circ$  of visual angle [22], [23] to fixation maps of the view-port image. Then the viewport images with spread fixations were back-projected into the sphere-map and then to the final equirectangular visual attention map. The OIQA database that includes the view-direction information, eye-fixation maps head saliency maps, and head-eye saliency maps was then released.

As shown in **Fig. 4**, for an image in OIQA database, we display its corresponding head-only saliency map and head-eye saliency map. Comparing the two saliency maps, we can see that the salient regions mainly centralize in the middle part of the equirectangular image nearby the equator. When viewing omnidirectional images in HMD, the top and bottom regions of an equirectangular image, i.e., the north and the south pole regions of the sphere, are less observed by the human subjects. Moreover, only a small part of the whole scene can be observed by users when viewing omnidirectional images in HMD. From an overall perspective, the head-only saliency map is similar with the head-eye saliency map. However, in details, there are many differences. Therefore, no matter in two-dimensional space or in three-dimensional space, it is reasonable and significant to assess the quality of images combining the visual saliency information.

### 2.5.3 Global Viewing Direction Bias

From **Fig. 4**, we can see that whether in head-only saliency map or in head-eye saliency map, the salient regions are all located around the equator of map. On the one hand, observers are more comfort when viewing the horizontal direction in HMD. On the other hand, when shooting panoramic images, salient scenes or objects are usually near the equator. Therefore we believe that it is one of the bottom layer features when viewing omnidirectional images using HMD. **Fig. 5** shows the scatter diagrams of global viewing direction (head or eye) weight



▲ Figure 4. The head-only saliency map and head-eye saliency map.

proportion along with latitude, clustering over all subjects and all omnidirectional images. One-, two- and three-term gaussian fitting curves are also plotted in this figure. From the figure, we can see that three-term gaussian fitting curves can get relative good fitting performance. Three-term gaussian fitting curves can be plotted by:

$$f(x) = \alpha_1 e^{-\left(\frac{x-\beta_1}{\gamma_1}\right)^2} + \alpha_2 e^{-\left(\frac{x-\beta_2}{\gamma_2}\right)^2} + \alpha_3 e^{-\left(\frac{x-\beta_3}{\gamma_3}\right)^2}. \quad (2)$$

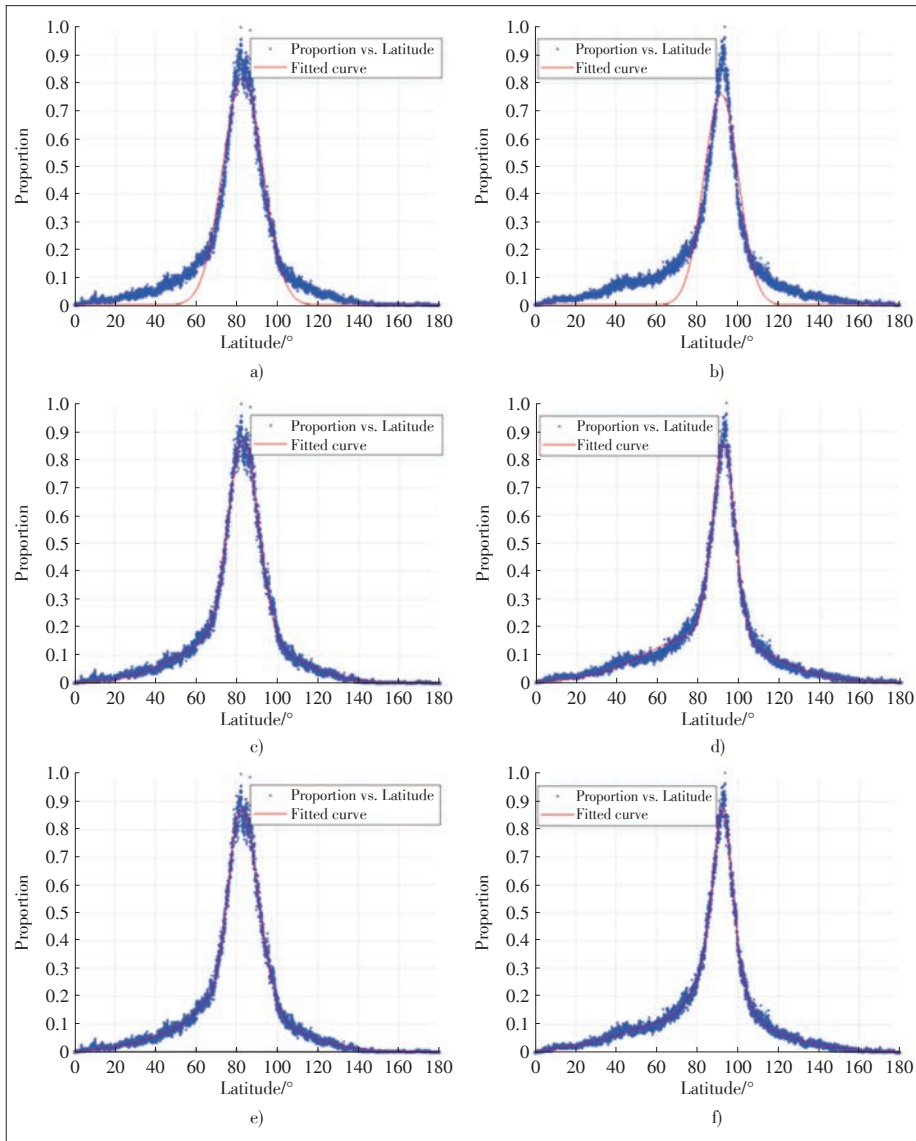
**Table 1** shows the coefficient of fitting curves. The first row lists the nine parameters in Equ. (2). The second and third rows list nine coefficient values of global head movement direction and global eye viewing direction fitting curves, respectively. From the fitting curves, we could get the global viewing direction bias when viewing omnidirectional images. This global viewing direction bias can be used to generate global saliency weight, which is shown in **Fig. 6**. The global viewing bias of omnidirectional images can be used in their perceptual quality assessment. We will discuss this method in the following.

## 3 Comparison of Objective Quality Assessment on the OIQA database

### 3.1 Experimental Protocol

#### 3.1.1 FR-IQA Measures

After the experiment, we compared the performance of 11



▲ Figure 5. The scatter diagrams of head movement direction or eye viewing direction weight proportion along with latitude, clustering over all the subjects and all the omnidirectional images; a) One-term fitting curves of head movement direction; b) one-term fitting curves of eye viewing direction; c) two-term fitting curves of head movement direction; d) two-term fitting curves of eye viewing direction; e) three-term fitting curves of head movement direction; f) three-term fitting curves of eye viewing direction.

▼ Table 1. The coefficient values of head movement direction fitting curves and eye viewing direction fitting curves

Parameter	$\alpha_1$	$\beta_1$	$\gamma_1$	$\alpha_2$	$\beta_2$	$\gamma_2$	$\alpha_3$	$\beta_3$	$\gamma_3$
Fitting value (head)	-0.2404	72.75	7.53	0.7957	81.19	12.53	0.1353	77.78	40.94
Fitting value (eye)	0.4943	92.75	6.05	0.2622	91.03	13.53	0.1288	81.26	48.26

state-of-the-art objective FR-IQA measures, which include 1) feature similarity (FSIM) [24], 2) gradient magnitude similarity deviation (GMSD) [25], 3) GSM [25] 4) gradient similarity (GSI) [26], 5) information content weighted structural similarity

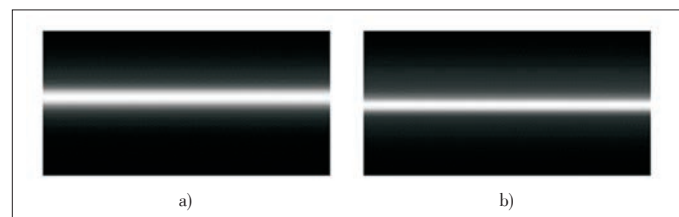
(IW-SSIM) [27], 6) mean squared error (MSE), 7) Multiscale structural similarity (MS-SSIM) [28], 8) peak signal-to-noise ratio (PSNR), 9) structural similarity (SSIM) [29], 10) visual information fidelity (VIF) [30], 11) visual saliency-induced index (VSI) [31]. When calculating the performance, we firstly mapped the predictions of the IQA models to subjective quality ratings through a five-parameter logistic function [32]–[34]:

$$f(x) = \beta_1 \left( \frac{1}{2} - \frac{1}{1 + e^{\beta_2(x - \beta_3)}} \right) + \beta_4 x + \beta_5, \quad (3)$$

in which  $x$  denotes the predicted scores;  $f(x)$  represents the corresponding mapped score;  $\beta_i$  ( $i=1, 2, 3, 4, 5$ ) are the parameters to be fitted. Then the mapped scores are compared with the subjective scores to measure the performance of the IQA models. In this paper, we use Pearson's Linear Correlation Coefficient (PLCC), root mean square error (RMSE) and spearman rank correlation coefficient (SRCC) as criteria to evaluate the performance of algorithms. Table 2 lists the performance of aforementioned IQA models under these three criteria.

### 3.1.2 Combining Human Visual Preference

The FR-IQA metrics are not only calculated on equirectangular images, but also calculated on cubic images. Fig. 7 shows the omnidirectional images in equirectangular format or cubic format. We supposed that the cubic format omnidirectional images could simulate the view-port which the subjects really saw in HMD. It is obvious that cubic images have less distortion than equirectangular images in this figure. The omnidirectional images in cubic format are similar with traditional 2D images so we think the tra-



▲ Figure 6. The global saliency weight map generated from global viewing direction bias: a) Global head movement characteristic map; b) global eye viewing preference map.



ditional IQA metrics could perform better on images of this kind of format. We also combined the saliency weight information aforementioned on part of FR-IQA models to compare the influence of saliency map with different accuracy on the evaluation results. Three kinds of saliency maps, including global head movement characteristic map, global eye viewing preference map and ground-truth visual saliency map are discussed in this paper. The results are shown in Table 2.

### 3.2 Performance Comparison

As show in Table 2a, FSIM, GSI and VSI perform better than other metrics. Although their performance is pretty well, we still believe that the performance could be improved better, e.g., using saliency or other human visual preference to OIQA. In this paper, we first compare the SSIM metrics with or without the pre-processing method, which includes a low-pass filter and downsampling process. SSIM2 is the metric with pre-processing procedure while SSIM1 without in Table 2. It is obvious that this pre-processing method contribute a lot to the performance promotion of SSIM in the OIQA database. Except for these models, other state-of-the-art IQA models perform not well, and they undergo some performance drop when transferring from traditional images to omnidirectional images.



▲ Figure 7. Omnidirectional images of equirectangular format or cubic format: a) Equirectangular image; b)–g) cubic images, which are Front, Right, Back, Left, top and bottom in sequence.

▼ Table 2. Performance of FR-IQA models in terms of PLCC, SRCC and RMSE. The best three performing metrics are highlighted with bold font

a) Assessing the perceptual quality of omnidirectional images on equirectangular format images

Metrics	PLCC	SRCC	RMSE
<b>FSIMc</b>	<b>0.9188</b>	<b>0.9140</b>	<b>5.6800</b>
GMSD	0.7412	0.7378	9.6574
GMSM	0.6768	0.6642	10.590
<b>GSI</b>	<b>0.9008</b>	<b>0.8924</b>	<b>6.2473</b>
IW-MSE	0.6207	0.7328	11.280
IW-PSNR	0.7371	0.7328	9.7223
IW-SSIM	0.7805	0.7766	8.9934
MSE	0.3279	0.4971	13.590
MS-SSIM	0.6745	0.6653	10.621
PSNR	0.5060	0.4971	12.408
SSIM1	0.5271	0.3479	12.225
SSIM2	0.8888	0.8800	6.5917
VIF	0.7878	0.7867	8.8614
VIFp	0.7555	0.7501	9.4246
<b>VSI</b>	<b>0.9087</b>	<b>0.9055</b>	<b>6.0059</b>

b) Assessing the perceptual quality of omnidirectional images on cubic format images

Metrics	PLCC	SRCC	RMSE
<b>FSIMc</b>	<b>0.9316</b>	<b>0.9278</b>	<b>5.2273</b>
GMSD	0.7120	0.7042	10.101
GMSM	0.6448	0.6393	10.996
<b>GSI</b>	<b>0.9215</b>	<b>0.9162</b>	<b>5.5878</b>
IW-MSE	0.6165	0.7110	11.327
IW-PSNR	0.7179	0.7054	10.014
IW-SSIM	0.7799	0.7755	9.0045
MSE	0.3919	0.5693	13.235
MS-SSIM	0.6699	0.6651	10.681
PSNR	0.5621	0.5603	11.898
SSIM1	0.4462	0.3870	12.875
SSIM2	0.8843	0.8740	6.7175
VIF	0.7725	0.7716	9.1351
VIFp	0.7761	0.7699	9.0723
<b>VSI</b>	<b>0.9236</b>	<b>0.9192</b>	<b>5.5158</b>

c) Assessing the perceptual quality of omnidirectional images combining head movement direction information (Fig. 6a)

Metrics	PLCC	SRCC	RMSE
<b>FSIMc</b>	<b>0.9118</b>	<b>0.9049</b>	<b>5.9061</b>
GMSM	0.6530	0.7035	10.895
MSE	0.3420	0.5026	13.518
PSNR	0.4123	0.3958	13.106
SSIM1	0.4481	0.3663	12.861
<b>SSIM2</b>	<b>0.8967</b>	<b>0.8844</b>	<b>6.3664</b>
<b>VSI</b>	<b>0.9009</b>	<b>0.8946</b>	<b>6.2451</b>

d) Assessing the perceptual quality of omnidirectional images combining eye viewing direction information (Fig. 6b)

Metrics	PLCC	SRCC	RMSE
<b>FSIMc</b>	<b>0.9148</b>	<b>0.9078</b>	<b>5.8090</b>
GMSM	0.7350	0.7283	9.7549
MSE	0.3407	0.5175	13.525
PSNR	0.4364	0.4154	12.943
SSIM1	0.4665	0.3849	12.725
<b>SSIM2</b>	<b>0.8988</b>	<b>0.8849</b>	<b>6.3075</b>
<b>VSI</b>	<b>0.9064</b>	<b>0.9005</b>	<b>6.0783</b>

e) Assessing the perceptual quality of omnidirectional images combining original saliency map from subjects

Metrics	PLCC	SRCC	RMSE
<b>FSIMc</b>	<b>0.9113</b>	<b>0.9015</b>	<b>5.9245</b>
GMSM	0.7508	0.7449	9.5026
MSE	0.3475	0.5317	13.489
PSNR	0.4858	0.4534	12.574
SSIM1	0.5083	0.4077	12.388
<b>SSIM2</b>	<b>0.8927</b>	<b>0.8779</b>	<b>6.4817</b>
<b>VSI</b>	<b>0.9086</b>	<b>0.9027</b>	<b>6.0071</b>

FR-IQA: full reference image quality assessment

FSIM: feature similarity

GMSD: gradient magnitude similarity deviation

GMSM: gradient magnitude similarity mean

GSI: gradient similarity

IW-MSE: information content weighted mean squared error

IW-PSNR: information content weighted peak signal-to-noise ratio

IW-SSIM: information content weighted structural similarity

MSE: mean squared error

MS-SSIM: multiscale structural similarity

PLCC: Pearson's Linear Correlation Coefficient

PSNR: peak signal-to-noise ratio

RMSE: root mean square error

SRCC: spearman rank correlation coefficient

SSIM1: structural similarity

SSIM2: structural similarity with pre-processing procedure

VIF: visual information fidelity

VSI: visual saliency-induced index

There is much room to improve these models.

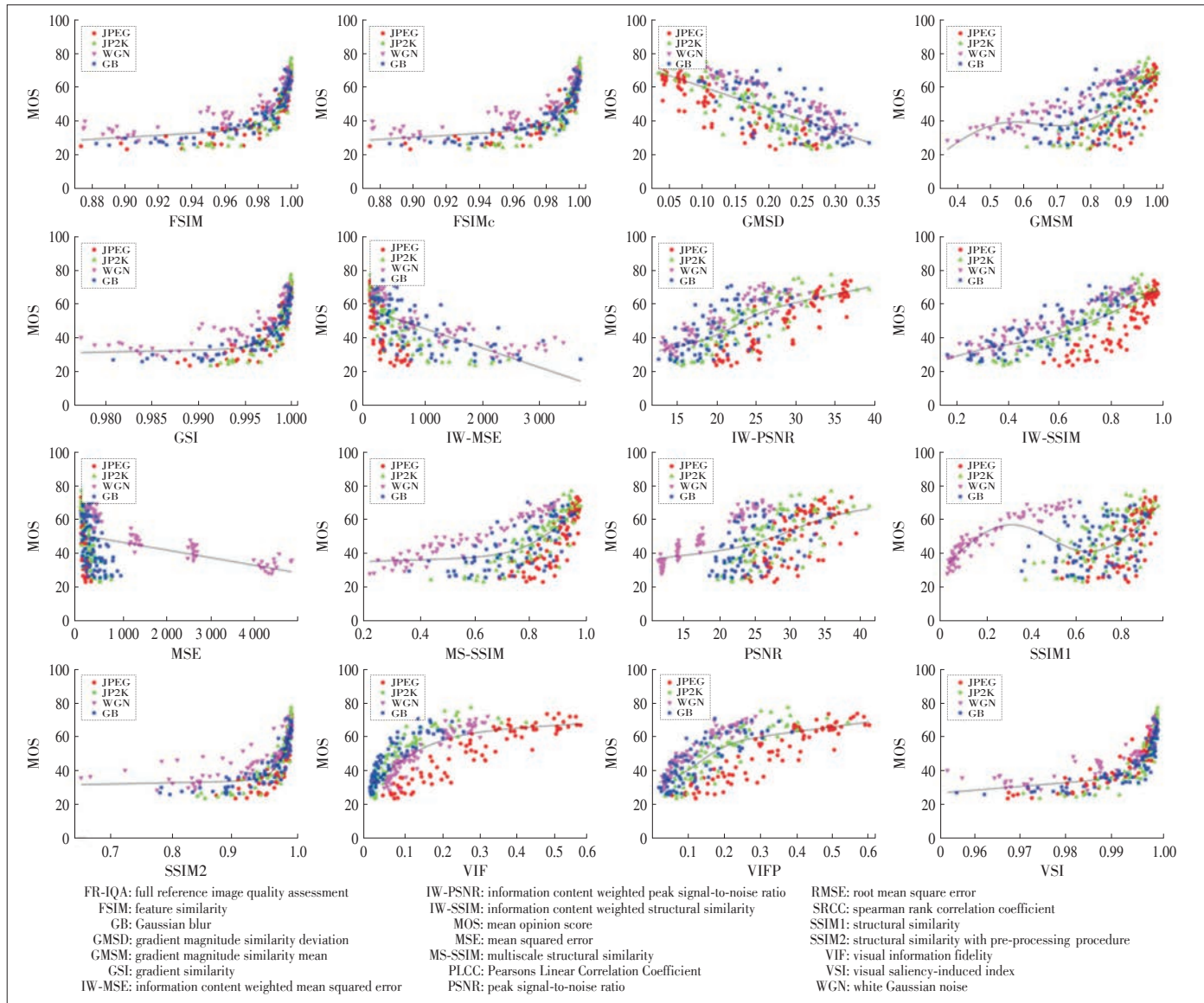
#### 1) Performance Comparison of FR-IQA Metrics on Equirectangular Images and Cubic Images

The performance of these metrics are also calculated on omnidirectional images in cubic format, as shown in Table 2b. The best three performing models in Table 2b are also FSIM, GSI, and VSI. Compared with Table 2a, they have significant performance improvement. The FSIMc calculated on cubic images get the best performance in Table 2. The significant performance improvement also appears in MSE and PSNR metrics. However, for some other FR-IQA metrics, the performance improvement is not obvious. For some FR-IQA metrics, the performance even decreases when calculating on cubic images. It illustrates that this method cannot improve the perfor-

mance of all FR-IQA metrics. We compared the scatter diagrams of FR-IQA metrics on equirectangular images (**Fig. 8**) and on cubic images (**Fig. 9**), respectively. Detailed illustration will be discussed in Subsection 3.3.

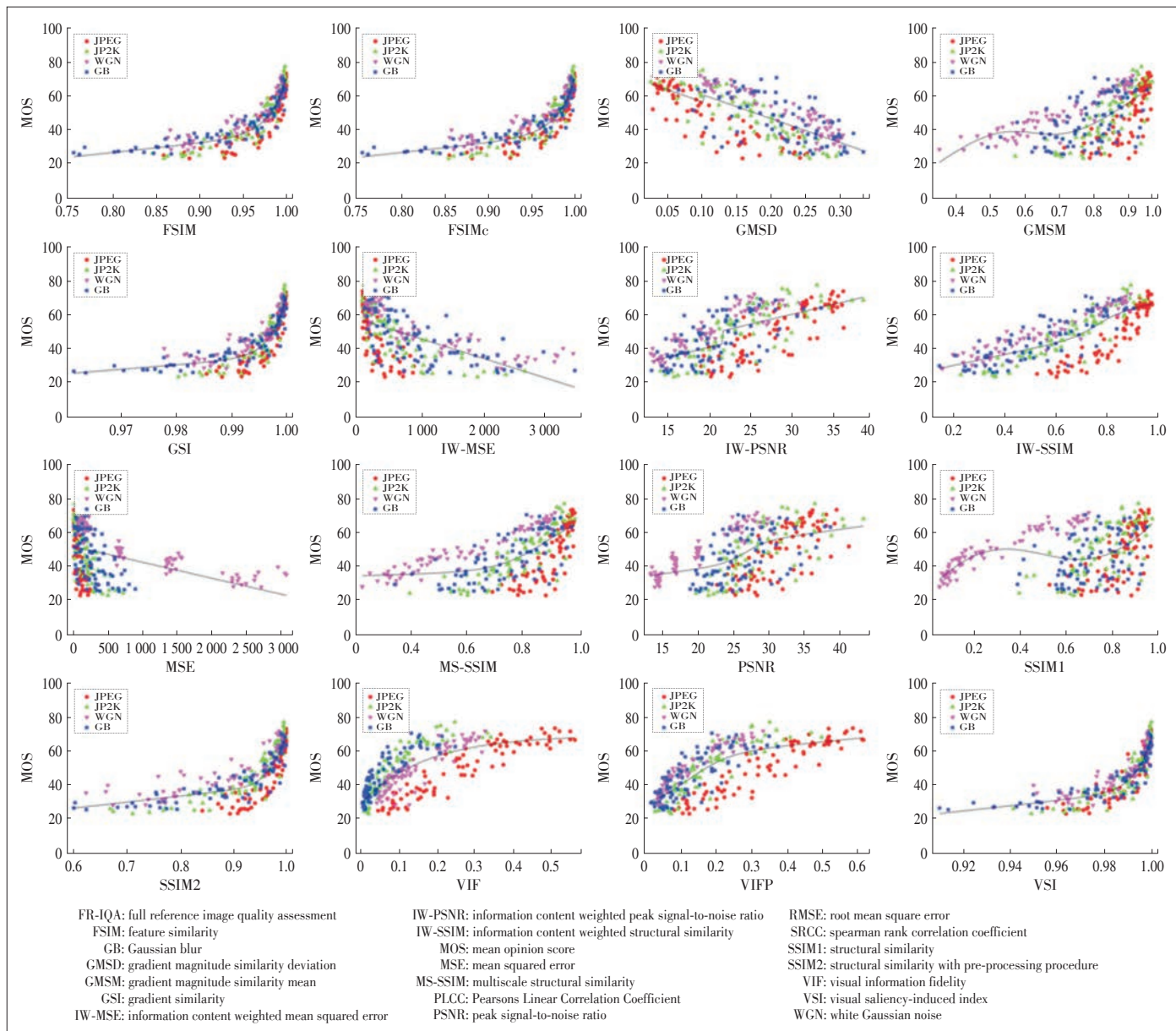
#### 2) Performance Comparison Combining Saliency Weight of Three Different types

In Tables 2c, 2d and 2e, we display the performance of seven FR-IQA metrics combining saliency weight with three different types, including the global head movement characteristic map, global eye viewing preference map and ground-truth visual saliency map. From these three tables, we find that as the accuracy of saliency map increases, the performance of these FR-IQA metrics increase in general, but the improvement is not impressive. Thus we propose that if the accuracy of



▲ **Figure 8.** Scatter plots of subjective MOS versus FR-IQA model prediction, including FSIM, FSIMc, GMSD, GMSM, GSI, IW-MSE, IW-PSNR, IW-SSIM, MSE, MS-SSIM, PSNR, SSIM1, SSIM2, VIF, VIFp, and VSI, based on equirectangular format images. The distortion types are JPEG compression (red points), JPEG2000 compression (green points), WGN (magenta points), and GB (blue points).





▲ Fig. 9. Scatter plots of subjective MOS versus FR-IQA model prediction, including FSIM, FSIMc, GMSD, GMSM, GSI, IW-MSE, IW-PSNR, IW-SSIM, MSE, MS-SSIM, PSNR, SSIM1, SSIM2, VIF, VIFp, and VSI, based on cubic format images. The distortion types are JPEG compression (red points), JPEG2000 compression (green points), WGN (magenta points), GB (blue points).

saliency map of omnidirectional images is not assured, researchers could combining the global saliency weight (Fig. 6) in their IQA algorithms. We find from the performance in Table 2 that the performance of some metrics decreases after combining the saliency weight. We think it is because that the intrinsic distortion of equirectangular projection is inconsistent with the saliency map. Relative issues need further research.

### 3.3 Differences Between Omnidirectional IQA and Traditional IQA

We select 16 FR - IQA models (FSIM, FSIMc, GMSD, GMSM, GSI, IW-MSE, IW-PSNR, IW-SSIM, MSE, MS-SSIM,

PSNR, SSIM1, SSIM2, VIF, VIFp, and VSI) and illustrate their scatter plots on equirectangular images (Fig. 8) and on cubic images (Fig. 9). The models we selected contain high performance metrics (such as FSIMc and VSI) and classical IQA metrics (such as SSIM and PSNR). As shown in Fig. 8, the scatter points whose color are magenta represent the distortion type of WGN. Compared with the scatter points of the other three distortion types, it is obvious that these scatter points are always far from the fitted curve. The scatter points of WGN are almost always higher than the scatter points of the other three distortion types. It means that these IQA models have predicted lower quality scores than the ideal values for distortion type WGN.

We believe this phenomenon is partly caused by the subjective ratings and partly caused by the objective IQA models. Another phenomenon we observed from the scatter plots in Figs. 8 and 9 is that the scatter points of JPEG compression do not fit well in some metrics. We think it is mainly caused by the subjective ratings.

#### 1) Subjective Rating Differences

For the exception to the WGN distortion, we see that this phenomenon is observed in various IQA models in Fig. 8. In traditional images, for all kinds of distortions, most IQA models show quite consistent predictions. However, in omnidirectional images, exception to the WGN distortion is observed. Therefore we believe it is partly caused by the subjective ratings. We believe that people prefer high-frequency content when viewing VR stimuli. And this preference leads to the exceptional subjective ratings. Observers will have more comfortable visual experiences when viewing high frequency content. Compared with traditional displays, subjects can only see the view-port image. This limited displaying effects of current VR-HMD also leads to the exceptional subjective ratings. Because the contents are not completed in the view-port and subjects will be annoyed with losing image details. We introduce four types of distortions in this paper, including JPEG compression, JPEG2000 compression, Gaussian noise, and Gaussian blur. Gaussian noise adds high frequency information to the image. The rest three distortions reduce the high frequency information and image details. For the exception to the JPEG compression distortion, we think humans' perceptual assessment of color distortion in VR environment is also different with that in traditional 2D displays. Some following work can be done regarding this phenomenon.

#### 2) Objective Measure Differences

For the exception to the WGN distortion, we also believe that this phenomenon is caused by the intrinsic distortion of equirectangular projection. Fig. 9 shows the scatter diagrams of MOS versus FR-IQA metrics on cubic images and we think cubic images have little intrinsic distortion. We excitingly find from the figure that for some models, such as FSIM, GSI, and VSI, the scatter points of the distortion type WGN are closer to the fitted curves, compared with Fig. 8, although they also a little far away from the fitted curve. Thus the performance of these metrics in Table 2b are better than those in Table 2a. Thus the phenomenon aforementioned is partly caused by the intrinsic distortion of equirectangular projection.

## 4 Conclusion and Future Work

In this paper, we investigate the methodology of assessing the quality of omnidirectional images. We first construct an omnidirectional IQA database. The database includes 16 source images with their corresponding 320 degraded images. We add four most commonly encountered distortions, including JPEG compression, JPEG2000 compression, Gaussian noise, and

Gaussian blur. We collect the subjective quality scores, view-orientation information, and eye-movement data during the experiment. By comparing objective FR-IQA models on the OIQA database, we propose that humans prefer high frequency content and image details in VR HMDs, and the losing of image details can do a lot of harm to the visual experience in the VR case. By comparing the performance of state-of-the-art objective FR-IQA models tested on the equirectangular images and cubic images, respectively, we find that calculating the IQA metrics on cubic images could improve some metrics' performance. Visual saliency information should also be combined in the IQA metrics, and more accurate saliency information will make the performance better.

## References

- [1] SHEIKH H R, WANG Z, CORMACK L, et al. Live Image Quality Assessment Database Release 2 [EB/OL]. (2005)[2018]. <http://live.ece.utexas.edu/research/quality>
- [2] PONOMARENKO N, LUKIN V, ZELENSKY A. Tid2008-A Database for Evaluation of Full-Reference Visual Quality Assessment Metrics [J]. *Advances of Modern Radioelectronics*, 2009, 10(4): 30–45
- [3] LARSON E C, CHANDLER D. Categorical Image Quality (CSIQ) Database [EB/OL]. (2010)[2018]. <http://vision.okstate.edu/csiq>
- [4] GU K, LIU M, ZHAI G T, et al. Quality Assessment Considering Viewing Distance and Image Resolution [J]. *IEEE Transactions on Broadcasting*, 2015, 61(3): 520–531. DOI: 10.1109/tbc.2015.2459851
- [5] RAI Y, LE CALLET P, GUILLOT P. Which Saliency Weighting for Omnidirectional Image Quality Assessment? [C]//Ninth International Conference on Quality of Multimedia Experience (QoMEX), Erfurt, Germany, 2017: 1–6. DOI: 10.1109/QoMEX.2017.7965659
- [6] UPENIK E, RERÁBEK M, EBRAHIMI T. Testbed for Subjective Evaluation of Omnidirectional Visual Content [C]//2016 Picture Coding Symposium (PCS), Nuremberg, Germany, 2016: 1–5. DOI: 10.1109/PCS.2016.7906378
- [7] YU M, LAKSHMAN H, GIROD B. A Framework to Evaluate Omnidirectional Video Coding Schemes [C]//IEEE International Symposium on Mixed and Augmented Reality, Fukuoka, Japan, 2015: 31–36. DOI: 10.1109/ISMAR.2015.12
- [8] SUN W, GU K, ZHAI G T, et al. CVIQD: Subjective Quality Evaluation of Compressed Virtual Reality Images [C]//IEEE International Conference on Image Processing (ICIP), Beijing, China, 2017: 3450–3454. DOI: 10.1109/ICIP.2017.8296923
- [9] DUAN H Y, ZHAI G T, YANG X K, et al. IVQAD 2017: An Immersive Video Quality Assessment Database [C]//International Conference on Systems, Signals and Image Processing (IWSSIP), Poznan, Poland, 2017: 1–5. DOI: 10.1109/IWSSIP.2017.7965610
- [10] DUAN H Y, ZHAI G T, MIN X K, et al. Perceptual Quality Assessment of Omnidirectional Images [C]//IEEE International Symposium on Circuits and Systems (ISCAS), Florence, Italy, 2018: 1–5. DOI: 10.1109/ISCAS.2018.8351786
- [11] ZHAI G T, CAI J F, LIN W S, et al. Cross-Dimensional Perceptual Quality Assessment for Low Bit-Rate Videos [J]. *IEEE Transactions on Multimedia*, 2008, 10(7): 1316–1324. DOI: 10.1109/tmm.2008.2004910
- [12] ZHAI G T, WU X L, YANG X K, et al. A Psychovisual Quality Metric in Free-Energy Principle [J]. *IEEE Transactions on Image Processing*, 2012, 21(1): 41–52. DOI: 10.1109/tip.2011.2161092
- [13] GU K, ZHAI G T, YANG X K, et al. Using Free Energy Principle for Blind Image Quality Assessment [J]. *IEEE Transactions on Multimedia*, 2015, 17(1): 50–63. DOI: 10.1109/tmm.2014.2373812
- [14] GU K, ZHAI G T, LIN W S, et al. No-Reference Image Sharpness Assessment

- in Autoregressive Parameter Space [J]. *IEEE Transactions on Image Processing*, 2015, 24(10): 3218–3231. DOI: 10.1109/tip.2015.2439035
- [15] ZHU W, ZHAI G, SUN W, et al. On the Impact of Environmental Sound on Perceived Visual Quality [C]//*Pacific Rim Conference on Multimedia (PCM)*, Harbin, China, 2017: 723–734
- [16] XU M, LI C, LIU Y F, et al. A Subjective Visual Quality Assessment Method of Panoramic Videos[C]//*IEEE International Conference on Multimedia and Expo (ICME)*, Hong Kong, China, 2017: 517–522. DOI: 10.1109/ICME.2017.8019351
- [17] WALLACE G K. The JPEG still Picture Compression Standard [J]. *IEEE Transactions on Consumer Electronics*, 1992, 38(1): xviii–xxxiv. DOI: 10.1109/30.125072
- [18] SKODRAS A, CHRISTOPOULOS C, EBRAHIMI T. The JPEG 2000 still Image Compression Standard [J]. *IEEE Signal Processing Magazine*, 2001, 18(5): 36–58. DOI: 10.1109/79.952804
- [19] 7invensun. Aglass [EB/OL]. (2017)[2018]. <http://www.aglass.com>
- [20] DUAN H Y, ZHAI G T, MIN X K, et al. Assessment of Visually Induced Motion Sickness in Immersive Videos [C]//*18th Pacific-Rim Conference on Multimedia (PCM)*, Harbin, China, 2017, pp. 662–672. DOI: 10.1007/978-3-319-77380-3\_63
- [21] RAI Y, CALLET P L. A dataset of head and eye movements for 360 degree images [C]//*ACM on Multimedia Systems Conference*, Taipei, China, 2017: 205–210. DOI: 10.1145/3083187.3083218
- [22] CURCIO C A, SLOAN K R, KALINA R E, et al. Human Photoreceptor Topography [J]. *The Journal of Comparative Neurology*, 1990, 292(4): 497–523. DOI: 10.1002/cne.902920402
- [23] ENGELKE U, ZHANG W, CALLET P L. Perceived Interest Versus Overt Visual Attention in Image Quality Assessment [J]. *Proceedings of SPIE*, vol. 9394, 2015. DOI: 10.1117/12.2086371
- [24] ZHANG L, ZHANG L, MOU X Q, et al. FSIM: A Feature Similarity Index for Image Quality Assessment [J]. *IEEE Transactions on Image Processing*, 2011, 20(8): 2378–2386. DOI: 10.1109/tip.2011.2109730
- [25] XUE W F, ZHANG L, MOU X Q, et al. Gradient Magnitude Similarity Deviation: A Highly Efficient Perceptual Image Quality Index [J]. *IEEE Transactions on Image Processing*, 2014, 23(2): 684–695. DOI: 10.1109/tip.2013.2293423
- [26] LIU A M, LIN W S, NARWARIA M. Image Quality Assessment Based on Gradient Similarity [J]. *IEEE Transactions on Image Processing*, 2012, 21(4): 1500–1512. DOI: 10.1109/tip.2011.2175935
- [27] WANG Z, LI Q. Information Content Weighting for Perceptual Image Quality Assessment [J]. *IEEE Transactions on Image Processing*, 2011, 20(5): 1185–1198. DOI: 10.1109/tip.2010.2092435
- [28] WANG Z, SIMONCELLI E P, BOVIK A C. Multiscale Structural Similarity for Image Quality Assessment [C]//*Thirty-Seventh Asilomar Conference on Signals, Systems & Computers*, Pacific Grove, USA, 2003: 1398–1402. DOI: 10.1109/ACSSC.2003.1292216
- [29] WANG Z, BOVIK A C, SHEIKH H R, et al. Image Quality Assessment: From Error Visibility to Structural Similarity [J]. *IEEE Transactions on Image Processing*, 2004, 13(4): 600–612. DOI: 10.1109/tip.2003.819861
- [30] SHEIKH H R, BOVIK A C. Image Information and Visual Quality [J]. *IEEE Transactions on Image Processing*, 2006, 15(2): 430–444. DOI: 10.1109/tip.2005.859378
- [31] ZHANG L, SHEN Y, LI H Y. VSI: A Visual Saliency-Induced Index for Perceptual Image Quality Assessment [J]. *IEEE Transactions on Image Processing*, 2014, 23(10): 4270–4281. DOI: 10.1109/tip.2014.2346028
- [32] MIN X K, GU K, ZHAI G T, et al. Blind Quality Assessment Based on Pseudo-Reference Image [J]. *IEEE Transactions on Multimedia*, 2018, 20(8): 2049–2062. DOI: 10.1109/tmm.2017.2788206
- [33] MIN X K, GU K, ZHAI G T, et al. Saliency-Induced Reduced-Reference Quality Index for Natural Scene and Screen Content Images [J]. *Signal Processing*, 2018, 145: 127–136. DOI: 10.1016/j.sigpro.2017.10.025
- [34] MIN X K, MA K D, GU K, et al. Unified Blind Quality Assessment of Compressed Natural, Graphic, and Screen Content Images [J]. *IEEE Transactions on Image Processing*, 2017, 26(11): 5462–5474. DOI: 10.1109/tip.2017.2735192

## Biographies

**DUAN Huiyu** (huiyuduan@sjtu.edu.cn) received the B.E. degree from the University of Electronic Science and Technology of China in 2017. He is currently pursuing the Ph.D. degree at the Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, China. His research interests include image quality assessment, visual attention modeling, and perceptual signal processing.

**ZHAI Guangtao** received the B.E. and M.E. degrees from Shandong University, China in 2001 and 2004, respectively, and the Ph.D. degree from Shanghai Jiao Tong University, China in 2009. From 2008 to 2009, he was a visiting student with the Department of Electrical and Computer Engineering, McMaster University, Hamilton, Canada, where he was a post-doctoral fellow from 2010 to 2012. From 2012 to 2013, he was a Humboldt Research Fellow with the Institute of Multimedia Communication and Signal Processing, Friedrich Alexander University of Erlangen - Nuremberg, Germany. He is currently a Research Professor with the Institute of Image Communication and Information Processing, Shanghai Jiao Tong University. His research interests include multimedia signal processing and perceptual signal processing. He received the National Excellent Ph.D. Thesis Award from the Ministry of Education of China in 2012.

**MIN Xiongkuo** received the B.E. degree from Wuhan University, China in 2013, and the Ph.D. degree from Shanghai Jiao Tong University, China in 2018. From 2016 to 2017, he was a visiting student with the Department of Electrical and Computer Engineering, University of Waterloo, Canada. He is currently a post-doctoral fellow with Shanghai Jiao Tong University. His research interests include image quality assessment, visual attention modeling, and perceptual signal processing. He received the Best Student Paper Award from the IEEE ICME 2016.

**ZHU Yucheng** received the B.E. degree from the Shanghai Jiao Tong University, China in 2015. He is currently pursuing the Ph.D. degree at the Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University. His research interests include image quality assessment, visual attention modeling, and perceptual signal processing. He received Grand Challenge Best Performance Awards in ICME 2017 and 2018.

**FANG Yi** is an undergraduate student at Shanghai Jiao Tong University, China and will receive the B.E. degree in 2019. She will pursue the M.E. degree at the Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University. Her research interests include image quality assessment, visual attention modeling, and perceptual signal processing.

**YANG Xiaokang** received the B.S. degree from Xiamen University, China in 1994, the M.S. degree from the Chinese Academy of Sciences, China in 1997, and the Ph.D. degree from Shanghai Jiao Tong University, China in 2000. From 2000 to 2002, he was a Research Fellow with the Centre for Signal Processing, Nanyang Technological University, Singapore. From 2002 to 2004, he was a Research Scientist with the Institute for Infocomm Research, Singapore. From 2007 to 2008, he visited the Institute for Computer Science, University of Freiburg, Freiburg im Breisgau, Germany, as an Alexander von Humboldt Research Fellow. He is currently a Distinguished Professor with the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, where he is also the Deputy Director of the Institute of Image Communication and Information Processing. His current research interests include image processing and communication, computer vision, and machine learning.

# Quality-of-Experience in Human-in-the-Loop Haptic Communications



LIU Qian<sup>1</sup> and ZHAO Tiesong<sup>2</sup>

(1. Dept. of Computer Science and Technology, Dalian University of Technology, Dalian, Liaoning 116024, China;  
2. College of Physics and Information Engineering, Fuzhou University, Fuzhou, Fujian 350108, China)

**Abstract:** With the worldwide rapid development of 5G networks, haptic communications, a key use case of the 5G, has attracted increasing attentions nowadays. Its human-in-the-loop nature makes quality of experience (QoE) the leading performance indicator of the system design. A vast number of high quality works were published on user-level, application-level and network-level QoE-oriented designs in haptic communications. In this paper, we present an overview of the recent research activities in this progressive research area. We start from the QoE modeling of human haptic perceptions, followed by the application-level QoE management mechanisms based on these QoE models. High fidelity haptic communications require an orchestra of QoE designs in the application level and the quality of service (QoS) support in the network level. Hence, we also review the state-of-the-art QoS-related QoE management strategies in haptic communications, especially the QoS-related QoE modeling which guides the resource allocation design of the communication network. In addition to a thorough survey of the literature, we also present the open challenges in this research area. We believe that our review and findings in this paper not only provide a timely summary of prevailing research in this area, but also help to inspire new QoE-related research opportunities in haptic communications.

**Keywords:** QoE; human-in-the-loop; haptic communications; kinesthetic signals; tactile signals; haptic

DOI: 10.12142/ZTECOM.201901008

<http://kns.cnki.net/kcms/detail/34.1294.TN.20190305.1718.004.html>, published online March 5, 2019

Manuscript received: 2019-01-11

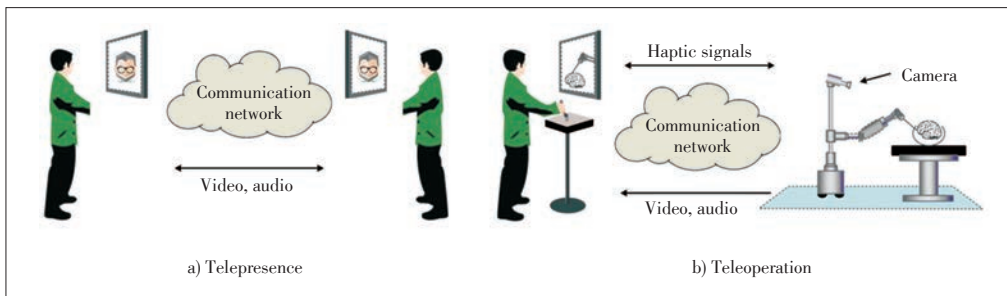
## 1 Introduction

The advent of mobile phones in the late 1980s broke the geographic barrier of landline telephones, so that we can have auditory conversation and interaction from anywhere. The past decade has witnessed the global blooming of mobile internet where audio-visual communications shape the way humans interact with technical systems or each other. Most recently, the world embraces the rise of the fifth generation (5G) of mobile networks, an excellent enabler of haptic communications, which will promote the human-to-human and human-to-machine interaction from the current audio-visual experience to the next-generation audio-visual-haptic perception.

**Fig. 1a** illustrates the conventional audio-visual communica-

tions (also known as telepresence), where one user remotely interacts with another user by exchanging audio/visual signals through the communication network. **Fig. 1b** demonstrates a typical haptic communication scenario for bilateral teleoperation. The global loop of haptic interaction consists of a human operator, a remote robot and the communication network. The audio/visual signals are transmitted from the remote robot to the operator, while haptic information exchanged bidirectionally between the human operator and the remote robot. Different from the conventional audio-visual communications, the “haptic” modality enables humans to actually alter the physical world remotely. It is obvious that in this human-in-the-loop haptic communication system, the quality of experience (QoE) of the human operator plays an important role in the operating/interacting performance of the entire system.





◀Figure 1.  
A comparison of conventional audio-visual communications and audio-visual-haptic communications.

A comprehensive definition of QoE was presented in [1] for the conventional audio-visual communication as “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.” In the context of haptic communications, the QoE inherits all genes from that of the conventional audio/visual communication, while extending the audio-visual perception to a third dimension, the haptic perception, referring to the sense of touch.

In this paper, we summarize the prevailing studies on QoE strategies in haptic communications. We should point out that this survey stands at the communication perspective, and reviews the QoE-related techniques, algorithms and mechanisms operating in the haptic communication chain. For the perspective of automotive control, a comprehensive review of network designs for the Quality of Control (QoC) can be found in [2]. A survey of various control systems stabilizing the global haptic communications can be found in [3]. The mechanical design of haptic interface devices is out of the scope of this paper.

The roadmap of this survey is shown in **Fig. 2**. We first introduce the user-level QoE models derived from the psychophysical factors of human haptic sensations in Section 2. Based on these models, various haptic applications are developed to enhance the performance of user perception, such as haptic-enabled virtual reality (VR) gaming and haptic-enabled cinema. In Section 3, we stand at the application level, and illustrate QoE-oriented designs in latest haptic applications including haptic data reduction schemes, multiplexing schemes for the multi-user transparency and the perceivable synchrony of multi-modal data delivery. In Section 4, we present the network-level QoS-related QoE management, focusing on the QoS-related QoE modeling which provide guidance to the resource allocation

of haptic signals. Finally, Section 5 concludes the paper with the future work and a summary.

## 2 User-Level QoE Management

The haptic perception of human beings generally refers to the sense of touch, composed of kinesthetic sense and tactile sense. In haptic communications, the human haptic perception is tightly connected to physical stimuli. As a result, we will first review the psychophysical impact factors of haptic interactions, based on which the user-level QoE models are developed in the literature.

### 2.1 Psychophysical Impact Factors

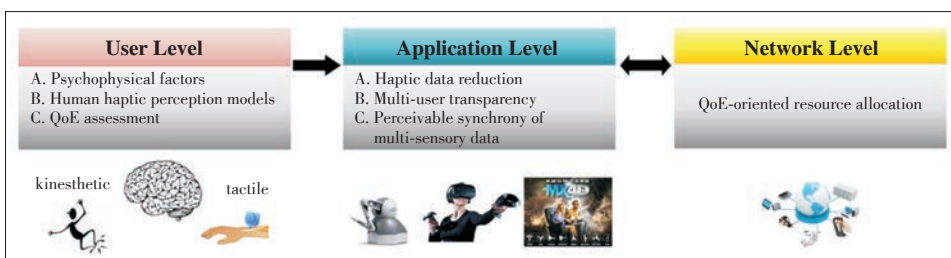
The haptic sensation is directly relevant to the human psychophysical perception mechanism. The kinesthetic sense allows for the perception of the position and orientation of our body parts and joints and external forces and torques applied to them. Hence, position, velocity, angular velocity, force, and torque all fall into the category of kinesthetic information. On the other hand, the tactile perception is sensed by different types of mechanoreceptors in the skin and allows humans to feel the surface texture (see [4] for details about the five psychophysical dimensions of tactile perception of textures), friction, temperature, etc. [5].

Haptic interactions generally involve both kinesthetic and tactile perceptions. For example, when a human operator controls a robot to grasp a rubber ball, the force and torque (kinesthetic) feedback presents the mass of the ball while the friction (tactile) feedback tells the texture of the surface. Since the haptic perception is an integration of multi-dimensional influence factors, the sense of touch is considered as the most complex sense to study [6]. A summary of the psychophysical factors

and corresponding human perception mechanism and exemplar signal generation approaches [7]–[14] are listed in **Table 1**.

### 2.2 Human Haptic Perception Models

The first human haptic perception study was performed by Weber [15], who examined the precision of the



▲Figure 2. Roadmap of this survey.

▼Table 1. Psychophysical factors, corresponding human perception mechanism and exemplar signal generation approaches

	Kinesthetic sense	Tactile sense
Signal type	Position, velocity, angular velocity, force, and torque	Surface texture and friction
Human perception mechanism	Sensed by the muscles, joints, and tendons of the body	Sensed by different types of mechanoreceptors in the skin
Exemplar signal generation solutions	Using high torque motors to generate kinesthetic force feedback, such as Geomagic Touch (used to be called as Phantom Omni [7]) and Omega 3 [8]	Multi-pin display attached to the human skin [9]–[11], or using vibrators to display vibrotactile stimuli [12], such as TPad [13] and TeslaTouch [14]

touch sense and developed the famous Weber's law [16]. In this perceptual law, the perceivable difference between two stimuli, the Just Noticeable Difference (JND), is proportional to the initial stimulus itself, which can be expressed as

$$\Delta I = k \cdot I, \quad (1)$$

where  $k$  is a constant;  $I$  and  $\Delta I$  denote the initial stimulus and the JND, respectively. The constant  $k$ , also called the Weber fraction, depends on the investigated stimulus, e.g. force, stiffness or velocity, and is generally obtained via experiments [16]. From Eq. (1), we can conclude that the human haptic perception system has different sensitivity with respect to the magnitude of the initial stimulus, e.g. the JND of large initial force is larger than that of a small initial force. Larger JND results in smaller sensitivity. In addition, the change of kinesthetic stimulus is linearly proportional to its intensity. Later, Fechner developed a logarithm model to reveal the relationship between the intensity of the stimulus and the change of kinesthetic perception in the brain [17]. Weber's linear model and Fechner's logarithmic model were proved essentially equivalent by Da-haene [18]. Since the linear model is simpler than the logarithmic counterpart, Weber's law is widely accepted as the QoE model of kinesthetic perception, which is then widely adopted in kinesthetic data reduction (see Section 3.1.1 for details).

The kinesthetic signal involves large amplitude/low frequency force feedback, and has been shown to lack realism due to

the absence of high-frequency transients (e.g., tapping on hard surfaces) and small-scale surface details (e.g., palpation of textured surfaces). Fortunately, the tactile signal provides an enhanced fidelity compared with the kinesthetic signal. The state-of-the-art tactile perception models concentrated on the modeling of vibrotactile texture signals [19]–[21]. Surprisingly, there is a strong similarity between texture signals and speech signals. This characteristic is then utilized in tactile data reduction technology (see Section 3.1.2 for details). As the rapid advances of machine learning algorithms, data-driven modeling and rendering approaches have been proposed for sophisticated tactile primitives, e.g., surface textures [22], viscoelasticity reactions [23], and thermal properties [24].

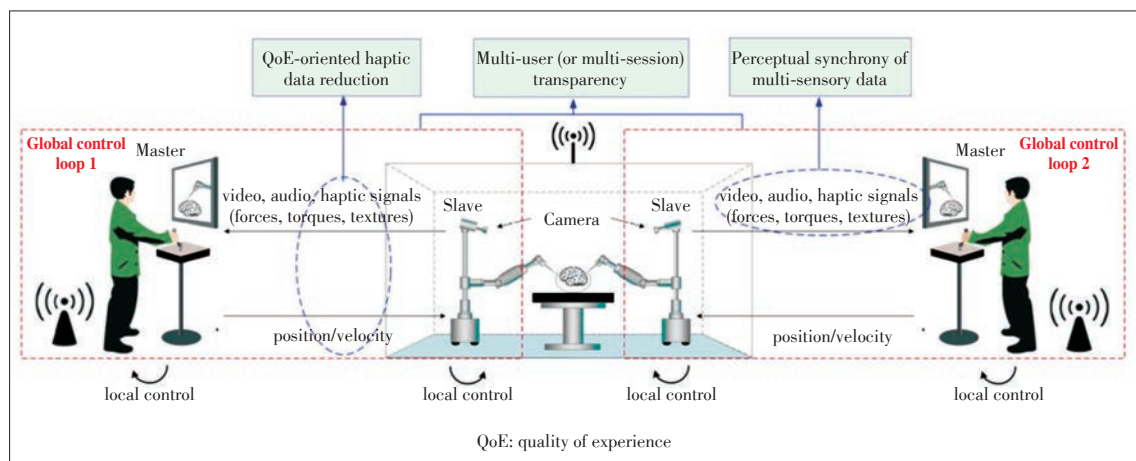
### 3 Application-Level QoE Management

The key objective of haptic applications is to satisfy the user's demand on haptic perceptual experience. In this section, we will summarize the key enablers of application-level QoE management, including QoE-oriented haptic data reduction, multi-user transparency, and perceptual synchrony of multi-sensory data, whose operational regions are illustrated in Fig. 3 by using an exemplar haptic teleoperation application with two teleoperation sessions, respectively.

#### 3.1 QoE-Oriented Haptic Data Reduction

It is known from Section 2.1 that two types of haptic signals (i.e. kinesthetic and tactile signals) are sensed by different human perception mechanisms and therefore, possess different properties. In addition, they have different tolerance on the communication delay. In particular, kinesthetic interactions are delay sensitive and will experience performance degradation in the presence of communication delay. The delay requirement of tactile interactions is quite relaxed compared with kinesthetic counterparts. On the other hand, the extreme high packet rate of kinesthetic feedback introduces a heavy burden to the network. The tactile feedback also contains multi-modal signals. Both kinesthetic and tactile interactions prefer

Figure 3.► Application-level QoE management in an exemplar haptic teleoperation application with two teleoperation sessions.



a data reduction module to improve the efficiency of the system. Therefore, in this section, we will review the data reduction solutions of kinesthetic and tactile signals, which are of significant importance to haptic communications.

### 3.1.1 Kinesthetic Codecs

In order to guarantee system stability, a high sampling rate of 1 kHz (or even higher) for the kinesthetic signals is required in the implementation of teleoperation systems over the network. The haptic sensor readings are typically packetized and transmitted once available in order to keep the communication delay as small as possible. As a result, 1 000 or more haptic data packets need to be transmitted every second between the master and the slave devices in addition to the audio and video streams. This phenomenon introduces a severe burden to the communication networks. In order to address this problem, perceptual data reduction schemes [25]–[32] were developed based on the Weber's law.

This principle dynamically selects the to-be-transmitted samples according to human perception thresholds (as shown in **Fig. 4** for a 1-DoF example). Samples with black dots represent the output of the perceptual deadband (PD) data reduction scheme. The perception thresholds are represented by deadbands, illustrated as gray zones in Fig. 4. Grey samples falling within the current deadband can be dropped, indicating that the signal change is too small to be perceived by human beings. This way, the PD data reduction strategy can reduce the average packet rate by approximately 80%–90%.

This single degree-of-freedom (DoF) approach has been extended to 3-DoF [29], [30], and has been refined with velocity-dependent force thresholds [32]. Furthermore, an error-resilient PD data reduction scheme were proposed in [33] to reduce the impact of packet losses.

In the presence of communication delays, the aforementioned haptic data reduction schemes have to be combined with stability-ensuring control schemes, e.g. wave variable (WV) scheme, time-domain passivity approach (TDPA), and model mediated teleoperation (MMT). The haptic packet rate reduction scheme has been combined with the WV control scheme in [34] and [35]. The resulting approach operates on haptic signals in the time domain (i.e., directly on the force and velocity signals). This scheme, however, is suited only for constant communication delay. Xu et al. [36] combined the

haptic packet rate reduction approach with the TDPA control scheme to reduce the packet rate over the communication network while preserving system stability in the presence of time-varying and unknown delays. This scheme is named TDPA + PD in the following. Compared to the existing WV-based haptic data reduction approaches, this scheme robustly deals with time-varying delays. Similarly, Xu et al. [37] incorporated a perception-based model update scheme into a point cloud-based MMT control architecture. This scheme is called MMT + PD in the following. The stability of the MMT architecture requires a stable and precise parameter estimation method to model the environment on the slave side. To address this issue, online environment modeling approaches were proposed for static objects [37], deformable objects [38], and movable objects [39]. Simple object models such as a rigid plane/sphere, a deformable thin membrane, or a freely movable cube are employed to approximate the remote environments. In [40], a passivity-based model update scheme was proposed to guarantee system stability during model update. In summary, the goal of our previous works in the area of MMT is to achieve stability while improving the transparency for networked interaction with simple or complex environments. MMT, however, can become computationally expensive and also requires a large amount of data to be transmitted between the slave and the master. Furthermore, its applicability is reduced as the environment dynamics increase.

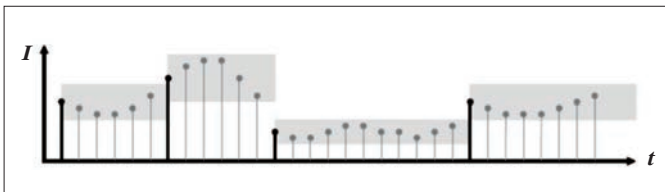
### 3.1.2 Tactile Codecs

Towards the compression of vibrotactile signals, offline algorithms were proposed in [41] and [42] with known prior knowledge of the surface texture (e.g., pre-scanning procedure). The first online compression of vibrotactile signals can be found in [43] for bilateral teleoperation. The compression algorithm is inspired by the similarities observed between texture signals and speech signals. Thus, a well-developed speech coding technique, the Algebraic Code-Excited Linear Prediction coding (ACE-LPC) [44], is adapted for developing a perceptually transparent texture codec. The authors of [43] reported a compression rate of 8:1 with a very low bitrate (4 kbits/s) on data transmission. An extended version of this compression algorithm was proposed in [45], in which the masking phenomenon in the perception of wide-band vibrotactile signals was applied to further improve the efficiency of the texture codec.

**Table 2** summarizes the human perception models and corresponding data reduction solutions for both kinesthetic and tactile signals, denoting as kinesthetic codecs and tactile codecs, respectively.

## 3.2 Multi-User Transparency

For the exemplar haptic teleoperation application shown in Fig. 3, the ideal system transparency is defined in this context as a perfect match between the master and slave positions and force signals, or alternatively a match between the environment



▲ **Figure 4.** Perceptual deadband principle. The perception thresholds (boundaries of gray zones) are a function of the stimulus intensity  $I$ . Samples that fall within the deadbands can be dropped (adapted from [25]).

▼ **Table 2. Overview of the human perception models and corresponding data reduction solutions (reproduced from [3])**

Haptic type	Human perception model	Data reduction solutions			
Kinesthetic	Weber's law (linear) [16]; Fechner's law (logarithmic) [17]	Without considering network conditions  Perceptual deadband schemes: Single DoF [2]; 3-DoF [29], [30]	considering network conditions		
			Solutions	Known const. delay	Unknown const. delay
			WV+ PD	[34]	[35]
			TDPA+ PD	[36]	[36]
Tactile	Data-driven ML models	[41], [42]			-
	Similar to speech signal	[43], [45]			-

DoF: degree of freedom  
ML: maximum likelihood

MMT: model mediated teleoperation  
PD: perceptual deadband

TDPA: time-domain passivity approach  
WV: wave variable

impedance and the impedance displayed to human operators [46]. When multiple users are remotely operating at the same environment (e.g. the patient's organ in Fig. 3), the maintenance of multi-user/multi-session transparency becomes a crucial but evitable problem. The state-of-the-art solutions basically focus on two aspects: plug-and-play (PnP) mechanisms to increase the interchangeability of multiple haptic interfaces, and the multi-user/multi-session synchronization strategy.

### 3.2.1 PnP Mechanisms

For the multi-user scenario of a given haptic application, it is essential to enable flexible and dynamic connections among multiple haptic interfaces and devices in order to support a rapid session setup and to ensure the interoperability of the employed components, e.g., detailed knowledge exchange of system parameters, functional capabilities, and the requirements of the deployed hardware (e.g. the description of DoF, workspace, and maximum forces/torques).

In [47], an Extensible Markup Language (XML)-based description language was proposed for virtual environment (VE)-based teleoperation systems. Based on this language, a web interface was developed in [48] to facilitate the PnP of haptic devices for a server-client-based teleoperation infrastructure. Another trend of the PnP mechanism [49] is to leverage the existing internet session and presence protocols, e.g. session initiation protocol (SIP).

### 3.2.2 Multi-User/Multi-Session Synchronization Strategy

In order to maintain the performance of the multi-user scenario in haptic communications, Schuwerk et al. [50]–[52] proposed to integrate the data compression, communication and control aiming at providing stable and perceptually transparent visual-haptic collaboration between two or more users. A VE-based teleoperation system [50] was developed based on the client-server architecture where the server manages the state consistency of the distributed VE, while the haptic feedback is

computed locally at each client. The PD data reduction principle was adopted to reduce the update rate of network traffic from the server to the client. However, this work neglected the communication delay which may lead to unavoidable inconsistencies in the VE states. A delay compensation strategy was proposed in [52] to solve this problem. Then, the work of [51] was further extended in [52] for deformable objects.

## 3.3 Perceivable Synchrony of Multi-Sensory Data

From Fig. 3, we can observe that a third modality, the haptic signal, is

transmitted from the remote environment to the human operator in haptic communication, in addition to the audio and visual modalities in conventional multimedia communications. It is well known that video data are bandwidth hungry, while the haptic signal has relatively higher delay requirement than the video and audio signals. Therefore, a perceivable synchrony of multi-sensory data streams should be achieved in order to provide a satisfactory QoE performance.

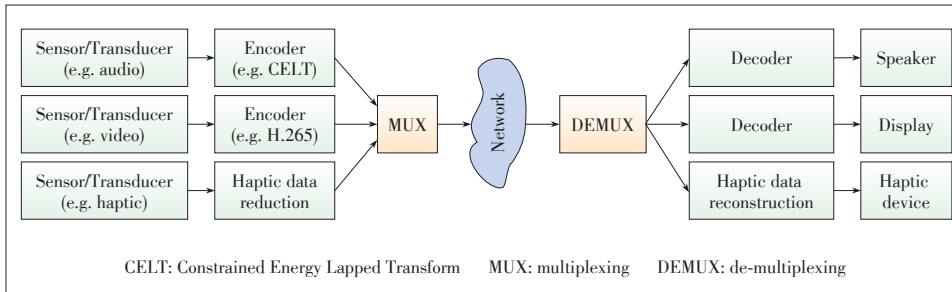
### 3.3.1 QoE Factor of Perceivable Delay for Multi-Sensory Data Synchronization

The first investigation on the effects of latency in human-machine interactions was conducted by Robert Miller in 1968 [53] through experiments of simple interaction events as keyboard typing and audio conversations. It was reported in [53] that a time delay of 100 ms is perceived as instantaneous. Yuan et al. [54] studied the perceived delays for multi-sensory data delivery in mulsemmedia services via extensive subjective tests. It was reported that haptic media could be presented up to 1 s behind video contents, air-flow media could be released either 5 s before or 3 s behind the video contents, while achieving an unperceivable asynchrony.

### 3.3.2 Application-Level Multiplexing Scheme

The state-of-the-art application-level multiplexing schemes for multi-sensory data delivery were presented in [55]–[58] with a typical structure shown in Fig. 5. This kind of scheme is able to multiplex different media modalities, with minimum computation cost and response time. They can interact with different network layers (such as the application layer for handling haptic data representation and the network layer for handling QoS requirements), and can also interact with the haptic-audio-video application to optimize network resources utilization to fit specific application needs. A synchronization approach is implemented in the multiplexing scheme for timely de-multiplexing of the communicated data in order to recover





▲ Figure 5. An application-level multiplexing scheme for synchronized multi-sensory data delivery.

individual media streams.

## 4 Network-Level QoE Management

In the literature review process of QoE-oriented designs in haptic communications, we discovered that most research work in this area was conducted at the application layer, leaving the network-level less visited. Pioneer work in this research area lies in the QoE-oriented resource allocation (RA) based on QoS-related QoE models. In this section, we will focus on the development of QoS-related QoE metric which provides a vital guidance to RA mechanisms.

The RA approach developed in [59] takes full advantage of the QoE-delay model developed in [60]. It is known that different control and communication approaches lead to different types of artifacts in haptic communications. Based on the characteristics of control schemes, XU et al. [60] proposed a hypothesis between the QoE and the end-to-end delay for different control schemes as shown in **Fig. 6a**, then obtained a QoE-delay model based on the subjective test results of a VE-based spring-damper teleoperation system, as shown in **Fig. 6b**. This model was later utilized in [59] to guide the network resource allocation of multi-session haptic communications aiming at achieving the maximal QoE performance.

Leveraging the bidirectional information exchange characteristic of haptic signals, Aijaz [61] developed a symmetric downlink and uplink RA strategy for haptic communications. Condoluci et al. [62] first assumed that the QoE performance is in-

versely proportional to the communication delay. Under this assumption, they [62] performed a data-driven study on the delay characteristic of haptic information, and then developed a soft resource reservation strategy aiming at minimize the communication delay of haptic data.

## 5 Conclusions and Future Work

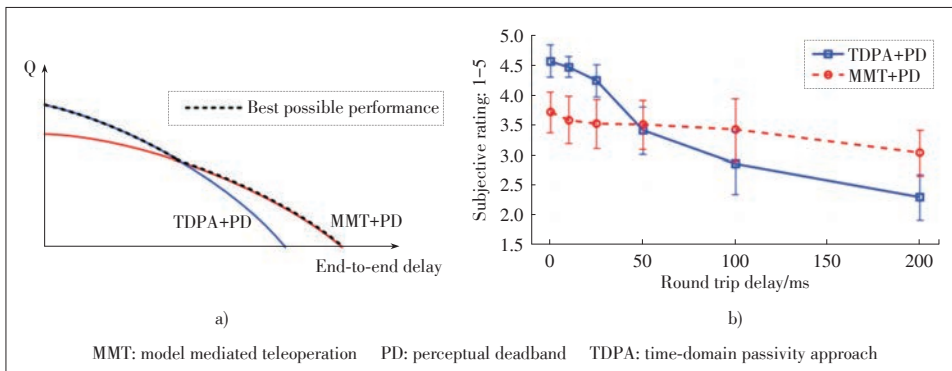
In this paper, we performed an extensive review on the research activities of QoE-oriented designs in haptic communications, starting from the user-level QoE (psychophysical) impact factors and models, the application-level QoE management (including perceptual haptic data reduction, transparency maintenance in multi-user scenario, and the perceptual synchrony of multi-sensory data), to the network-level QoS-related QoE management.

Compared with the QoE-related technologies in conventional audio-visual communications, QoE in haptic communications is a relatively young research direction, with new open challenges and exciting new research opportunities. Potential future research directions may related but not limited to the following aspects:

(1) Exploring the relationship among QoE, quality-of-task (QoT), quality-of-control (QoC), and QoS: The user experience of haptic communications is influenced by the network conditions (e.g. delay, jitter, and packet loss), the adopted control scheme, and the complexity of assigned tasks (e.g. free space versus contact and soft objects versus rigid surface). A thorough consideration of all impact factors will absolutely enhance the accuracy of QoE performance models, and provide a better guidance to various applications, e.g. the preferred control scheme of a given task under a given QoS setup.

(2) Low-latency video coding: It is well known that video streams are bandwidth hungry. The user perceivable delay constraints violation problem will become even more severe when

high resolution video cameras are adopted in haptic communications. Therefore, low-latency video coding should be included in the application-level multiplexing scheme in order to assure the perceptual synchrony of multi-sensory data. The state-of-the-art video coding standard, H.265/HEVC, provides special support for low-delay video applications. The development of parallel processing tools, the new “dependent slice segments” concept and the new “hypothetical reference decoder process-



▲ Figure 6. a) Hypothesis between quality of experience and delay for different control schemes; b) subjective tests results of a virtual environment-based spring-damper teleoperation system (adopted from [60]).

ing” concept, makes the low-delay video encoding and decoding a reality. In addition, intra-block refreshing [63], insignificant frames dropping [64], and key frame selection mechanisms will also help realize low-delay video coding technology.

We expect that this paper can promote the research activities in the abovementioned developed research directions for QoE in haptic communications, and also trigger the development of complex multi-sensory media systems, including senses of audio, visual, haptic, olfaction, and gustation, as well as the associated QoE studies.

## References

- [1] ZHAO T S, LIU Q, CHEN C W. QoE in Video Transmission: A User Experience-Driven Strategy [J]. *IEEE Communications Surveys & Tutorials*, 2017, 19(1): 285–302. DOI: 10.1109/comst.2016.2619982
- [2] PARK P, COLERI ERGEN S, FISCHIONE C, et al. Wireless Network Design for Control Systems: A Survey [J]. *IEEE Communications Surveys & Tutorials*, 2018, 20(2): 978–1013. DOI: 10.1109/comst.2017.2780114
- [3] ANTONAKOGLU K, XU X, STEINBACH E, et al. Toward Haptic Communications over the 5G Tactile Internet [J]. *IEEE Communications Surveys & Tutorials*, 2018, 20(4): 3034–3059. DOI: 10.1109/comst.2018.2851452
- [4] OKAMOTO S, NAGANO H, YAMADA Y. Psychophysical Dimensions of Tactile Perception of Textures [J]. *IEEE Transactions on Haptics*, 2013, 6(1): 81–93. DOI: 10.1109/toh.2012.32
- [5] CHA J, HO Y-S, KIM Y, et al. A Framework for Haptic Broadcasting [J]. *IEEE Multimedia*, 2019, 16(3): pp. 16–27.
- [6] STEINBACH E, HIRCHE S, ERNST M, et al. Haptic Communications [J]. *Proceedings of the IEEE*, 2012, 100: 937–956. DOI: 10.1109/JPROC.2011.2182100
- [7] SILVA A, RAMIREZ O, et al. PHANTOM OMNI Haptic Device: Kinematic and Manipulability [C]//*Electronics, Robotics and Automotive Mechanics Conference (CERMA)*, Cuernavaca, Morelos, 2009: 193–198
- [8] FORCE DIMENSION. Force Dimension User Manual of Omega.x Haptic Device, Version 1.7 [Z]. Nyon, Switzerland, 2013.
- [9] BREWSTER S, CHOCHAN F, BROWN L. Tactile Feedback for Mobile Interactions [C]//*SIGCHI Conference on Human Factors in Computing Systems*, Gaitheburg, USA, 2007: 159–162
- [10] DREWING K, FRITSCHI M, ZOPF R, et al. First Evaluation of a Novel Tactile Display Exerting Shear Force Via Lateral Displacement [J]. *ACM Transactions on Applied Perception*, 2005, 2(2): 118–131. DOI: 10.1145/1060581.1060586
- [11] BENALI-KHOUDJA M, HAFEZ M, ALEXANDRE J M, et al. Tactile Interfaces: A State-of-the-Art Survey [C]//*International Symposium of Robotics*, Paris, France, 2004: 23–26
- [12] CHOUVARDAS V G, MILIOU A N, HATALIS M K. Tactile Display Applications: A State-of-the-Art Survey [C]//*2nd Balkan Conference in Informatics*, Ohrid, Macedonia, 2005: 290–303
- [13] WINFIELD L, GLASSMIRE J, COLGATE J E, et al. T-PaD: Tactile Pattern Display Through Variable Friction Reduction [C]//*Second Joint EuroHaptics Conference and Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems (WHC'07)*, Tsukuba, Japan, 2007: 421–426. DOI: 10.1109/WHC.2007.105
- [14] BAU O, POUPYREV I, ISRAR A et al. TeslaTouch: Electro-vibration for Touch Surfaces [C]//*User Interface Software and Technology (UIST)*, New York, USA: 283–292, 2010.
- [15] WEBER E. *The Sense of Touch* [M]. Cambridge, USA: Academic Press, 1978
- [16] GESCHIEDER G A. *Psychophysics: The Fundamentals* [M]. 3rd ed. Mahwah, USA: Lawrence Erlbaum, 1997
- [17] DEHAENE S. The Neural Basis of the Weber-Fechner Law: A Logarithmic Mental Number Line [J]. *Trends in Cognitive Sciences*, 2003, 7(4): 145–147. DOI: 10.1016/S1364-6613(03)00055-X
- [18] DEHAENE S. Subtracting Pigeons: Logarithmic or Linear? [J] *Psychological Science*, 2001, 12(3): 244–246. DOI: 10.1111/1467-9280.00343
- [19] OKAMURA A M, DENNERLEIN J T, HOWE R D. Vibration Feedback Models for Virtual Environments [C]//*IEEE International Conference on Robotics and Automation (Cat. no.98CH36146)*, Leuven, Belgium, 1998: 674–679. DOI: 10.1109/ROBOT.1998.677050
- [20] KUCHENBECKER K J, ROMANO J, MCMAHAN W. *Haptography: Capturing and Recreating the Rich Feel of Real Surfaces* [M]//KUCHENBECKER K J, ROMANO J, MCMAHAN W. eds. *Springer Tracts in Advanced Robotics*. Berlin, Heidelberg, Germany: Springer Berlin Heidelberg, 2011: 245–260. DOI: 10.1007/978-3-642-19457-3\_15
- [21] GURUSWAMY V L, LANG J, LEE W S. IIR Filter Models of Haptic Vibration Textures [J]. *IEEE Transactions on Instrumentation and Measurement*, 2011, 60(1): 93–103. DOI: 10.1109/tim.2010.2065751
- [22] SHIN S, OSGOUEI R H, KIM K D, et al. Data-Driven Modeling of Isotropic Haptic Textures Using Frequency-Decomposed Neural Networks [C]//*IEEE World Haptics Conference (WHC)*, Evanston, USA, 2015: 131–138. DOI: 10.1109/WHC.2015.7177703
- [23] YIM S, JEON S, CHOI S. Data-Driven Haptic Modeling and Rendering of Viscoelastic and Frictional Responses of Deformable Objects [J]. *IEEE Transactions on Haptics*, 2016, 9(4): 548–559. DOI: 10.1109/toh.2016.2571690
- [24] CHOI H, CHO S, SHIN S, et al. Data-Driven Thermal Rendering: An Initial Study [C]//*IEEE Haptics Symposium (HAPTICS)*, San Francisco, USA, 2018: 344–350. DOI: 10.1109/HAPTICS.2018.8357199
- [25] STEINBACH E, HIRCHE S, KAMMERL J, et al. Haptic Data Compression and Communication [J]. *IEEE Signal Processing Magazine*, 2011, 28(1): 87–96. DOI: 10.1109/MSP.2010.938753
- [26] HINTERSEER P, STEINBACH E, HIRCHE S, et al. A Novel, Psychophysically Motivated Transmission Approach for Haptic Data Streams in Telepresence and Teleaction Systems [C]//*ICASSP'05. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Philadelphia, USA, 2005: ii/1097–ii/1100. DOI: 10.1109/ICASSP.2005.1415600
- [27] HINTERSEER P, STEINBACH E, CHAUDHURI S. Perception-Based Compression of Haptic Data Streams Using Kalman Filters [C]//*IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, Toulouse, France, 2006. DOI: 10.1109/ICASSP.2006.1661315
- [28] HINTERSEER P, STEIBACH E, CHAUDHURI S. Model Based Data Compression for 3D Virtual Haptic Teleinteraction [C]//*International Conference on Consumer Electronics*, Las Vegas, USA, 2006: 23–24. DOI: 10.1109/ICCE.2006.1598291
- [29] HINTERSEER P, HIRCHE S, CHAUDHURI S, et al. Perception-Based Data Reduction and Transmission of Haptic Data in Telepresence and Teleaction Systems [J]. *IEEE Transactions on Signal Processing*, 2008, 56(2): 588–597. DOI: 10.1109/tsp.2007.906746
- [30] KAMMERL J, CHAUDHARI R, STEINBACH E. Combining Contact Models with Perceptual Data Reduction for Efficient Haptic Data Communication in Networked VEs [J]. *IEEE Transactions on Instrumentation and Measurement*, 2011, 60(1): 57–68. DOI: 10.1109/tim.2010.2065670
- [31] KAMMERL J, VITTORIAS I, NITSCH V, et al. Perception-Based Data Reduction for Haptic Force-Feedback Signals Using Velocity-Adaptive Deadbands [J]. *Presence: Teleoperators and Virtual Environments*, 2010, 19(5): 450–462. DOI: 10.1162/pres\_a\_00008
- [32] KAMMERL J, CHAUDHARI R, STEINBACH E. Exploiting Directional Dependencies of Force Perception for Lossy Haptic Data Reduction [C]//*IEEE International Symposium on Haptic Audio Visual Environments and Games*, Phoenix, USA, 2010: 1–6. DOI: 10.1109/HAVE.2010.5623975
- [33] BRANDI F, KAMMERL J, STEINBACH E. Error-Resilient Perceptual Coding for Networked Haptic Interaction [C]//*ACM International Conference on Multimedia*, At Firenze, Italy. 2010: pp. 351–360. DOI: 10.1145/1873951.1874000
- [34] HIRCHE S, BUSS M. Transparent Data Reduction in Networked Telepresence and Teleaction Systems. Part II: Time-Delayed Communication [J]. *Presence: Teleoperators and Virtual Environments*, 2007, 16(5): 532–542. DOI: 10.1162/pres.16.5.532
- [35] VITTORIAS I, KAMMERL J, HIRCHE S, et al. Perceptual Coding of Haptic Data in Time-Delayed Teleoperation [C]//*World Haptics 2009—Third Joint EuroHaptics Conference and Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems*, Salt Lake City, USA, 2009: 208–213. DOI: 10.1109/WHC.2009.4810811
- [36] XU X, SCHUWERK C, CIZMECI B, et al. Energy Prediction for Teleoperation Systems that Combine the Time Domain Passivity Approach with Perceptual Deadband-Based Haptic Data Reduction [J]. *IEEE Transactions on Haptics*, 2016, 9(4): 560–573. DOI: 10.1109/toh.2016.2558157
- [37] XU X, CIZMECI B, AL-NUAIMI A, et al. Point Cloud-Based Model-Mediated Teleoperation with Dynamic and Perception-Based Model Updating [J]. *IEEE Transactions on Instrumentation and Measurement*, 2014, 63(11): 2558–2569. DOI: 10.1109/tim.2014.2323139
- [38] XU X, KAMMERL J, CHAUDHARI R, et al. Hybrid Signal-Based and Geometry-Based Prediction for Haptic Data Reduction [C]//*IEEE International Work-*

- shop on Haptic Audio Visual Environments and Games, Qinghuangdao, China, 2011: 68–73. DOI: 10.1109/HAVE.2011.6088394
- [39] XU X, STEINBACH E. Towards Real-Time Modeling and Haptic Rendering of Deformable Objects for Point Cloud-Based Model-Mediated Teleoperation [C]//IEEE International Conference on Multimedia and Expo Workshops (ICMEW), Chengdu, China, 2014: 1–6. DOI: 10.1109/ICMEW.2014.6890578
- [40] XU X, CHEN S L, STEINBACH E. Model-Mediated Teleoperation for Movable Objects: Dynamics Modeling and Packet Rate Reduction [C]//IEEE International Symposium on Haptic, Audio and Visual Environments and Games (HAVE), Ottawa, Canada, 2015: 1–6. DOI: 10.1109/HAVE.2015.7359474
- [41] OKAMOTO S, YAMADA Y. Perceptual Properties of Vibrotactile Material Texture: Effects of Amplitude Changes and Stimuli beneath Detection Thresholds [C]//IEEE/SICE International Symposium on System Integration, Sendai, Japan, 2010: 384–389. DOI: 10.1109/SII.2010.5708356
- [42] CRAIG J C. Difference Threshold for Intensity of Tactile Stimuli [J]. *Perception & Psychophysics*, 1972, 11(2): 150–152. DOI: 10.3758/bf03210362
- [43] CHAUDHARI R, CIZMECI B, KUCHENBECKER K J, et al. Low Bitrate Source-Filter Model Based Compression of Vibrotactile Texture Signals in Haptic Teleoperation [C]//20th ACM International Conference on Multimedia. New York, USA, 2012: 409–418
- [44] ITU-T. Coding Of Speech At 8 Kbit/S Using Conjugate-Structure Algebraic-Code-Excited Linear-Prediction (CS-ACELP): Reduced Complexity 8 kbit/s CSACELP Speech Codec [S]. 1996
- [45] CHAUDHARI R, SCHUWERK C, DANAEI M, et al. Perceptual and Bitrate-Scalable Coding of Haptic Surface Texture Signals [J]. *IEEE Journal of Selected Topics in Signal Processing*, 2015, 9(3): 462–473. DOI: 10.1109/jstsp.2014.2374574
- [46] LAWRENCE D A. Stability and Transparency in Bilateral Teleoperation [J]. *IEEE Transactions on Robotics and Automation*, 1993, 9(5): 624–637. DOI: 10.1109/70.258054
- [47] EID M, ALAMRI A, SADDIK A E. MPEG-7 Description of Haptic Applications Using HAML [C]//IEEE International Workshop on Haptic Audio Visual Environments and their Applications (HAVE 2006), Ottawa, Canada, 2006: 134–139. DOI: 10.1109/HAVE.2006.283780
- [48] GAO Y, OSMAN H A, EL SADDIK A. MPEG-V Based Web Haptic Authoring Tool [C]//IEEE International Symposium on Haptic Audio Visual Environments and Games (HAVE), Istanbul, Turkey, 2013: 87–91. DOI: 10.1109/HAVE.2013.6679616
- [49] KING H H, HANNAFORD B, KAMMERLY J, et al. Establishing Multimodal Telepresence Sessions Using the Session Initiation Protocol (SIP) and Advanced Haptic Codecs [C]//IEEE Haptics Symposium, Waltham, USA, 2010: 321–325. DOI: 10.1109/HAPTIC.2010.5444637
- [50] SCHUWERK C, XU X, CHAUDHARI R, et al. Compensating the Effect of Communication Delay in Client-Server: Based Shared Haptic Virtual Environments [J]. *ACM Transactions on Applied Perception*, 2015, 13(1): 1–22. DOI: 10.1145/2835176
- [51] SCHUWERK C, PAGGETTI G, CHAUDHARI R, et al. Perception-Based Traffic Control for Shared Haptic Virtual Environments [J]. *Presence: Teleoperators and Virtual Environments*, 2014, 23(3): 320–338. DOI: 10.1162/pres\_a\_00196
- [52] SCHUWERK C, XU X, STEINBACH E. On the Transparency of Client/Server-Based Haptic Interaction with Deformable Objects [J]. *IEEE Transactions on Haptics*, 2017, 10(2): 240–253. DOI: 10.1109/toh.2016.2612635
- [53] MILLER R B. Response Time in Man-Computer Conversational Transactions [C]//AFIPS'68 (Fall, part I), San Francisco, USA, 1968: 267–277
- [54] YUAN Z H, BI T, MUNTEAN G M, et al. Perceived Synchronization of Multimedia Services [J]. *IEEE Transactions on Multimedia*, 2015, 17(7): 957–966. DOI: 10.1109/tmm.2015.2431915
- [55] EID M, CHA J, EL SADDIK A. Admux: An Adaptive Multiplexer for Haptic-Audio-Visual Data Communication [J]. *IEEE Transactions on Instrumentation and Measurement*, 2011, 60(1): 21–31. DOI: 10.1109/tim.2010.2065530
- [56] CIZMECI B, XU X, CHAUDHARI R, et al. A Multiplexing Scheme for Multimodal Teleoperation [J]. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2017, 13(2): 1–28. DOI: 10.1145/3063594
- [57] CIZMECI B, CHAUDHARI R, XU X, et al. A Visual-Haptic Multiplexing Scheme for Teleoperation Over Constant-Bitrate Communication Links [M]//CIZMECI B, CHAUDHARI R, XU X, et al. eds. *Haptics: Neuroscience, Devices, Modeling, and Applications*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014: 131–138. DOI: 10.1007/978-3-662-44196-1\_17
- [58] EID M, CHA J, EL SADDIK A. An Adaptive Multiplexer for Multi-Modal Data Communication [C]//IEEE International Workshop on Haptic Audio Visual Environments and Games, Lecco, Italy, 2009: 111–116. DOI: 10.1109/HAVE.2009.5356121
- [59] LIU S W, LI M L, XU X, et al. QoE-Driven Uplink Scheduling for Haptic Communications over 5G Enabled Tactile Internet [C]//IEEE International Symposium on Haptic, Audio and Visual Environments and Games (HAVE), Dalian, China, 2018: 1–5. DOI: 10.1109/HAVE.2018.8547503
- [60] XU X, LIU Q, STEINBACH E. Toward QoE-Driven Dynamic Control Scheme Switching for Time-Delayed Teleoperation Systems: A Dedicated Case Study [C]//IEEE International Symposium on Haptic, Audio and Visual Environments and Games (HAVE), Abu Dhabi, United Arab Emirates, 2017: 1–6. DOI: 10.1109/HAVE.2017.8240352
- [61] ALJAZ A. Toward Human-In-The-Loop Mobile Networks: A Radio Resource Allocation Perspective on Haptic Communications [J]. *IEEE Transactions on Wireless Communications*, 2018, 17(7): 4493–4508. DOI: 10.1109/twc.2018.2825985
- [62] CONDOLUCI M, MAHMOODI T, STEINBACH E, et al. Soft Resource Reservation for Low-Delayed Teleoperation over Mobile Networks [J]. *IEEE Access*, 2017, 5: 10445–10455. DOI: 10.1109/access.2017.2707319
- [63] SONG R, WANG Y L, HAN Y, et al. Statistically Uniform Intra-Block Refresh Algorithm for very Low Delay Video Communication [J]. *Journal of Zhejiang University SCIENCE C*, 2013, 14(5): 374–382. DOI: 10.1631/jzus.c1200333
- [64] LIU T, CHOUDARY C. Real-Time Content Analysis and Adaptive Transmission of Lecture Videos for Mobile Applications [C]//12th Annual ACM International Conference on Multimedia, New York, USA, 2005: 400–403

### Biographies

**LIU Qian** (qianliu@dlut.edu.cn) currently works as an associate professor at the Dept. of Computer Science and Technology, Dalian University of Technology, China. She received her B.S. and M.S. degrees from Dalian University of Technology in 2006 and 2009, respectively, and the Ph.D. degree from The State University of New York at Buffalo, USA, in 2013. She served as a post-doctoral fellow with the Ubiquitous Multimedia Laboratory, State University of New York at Buffalo from 2013 to 2015. She received Alexander von Humboldt Fellowship and worked with the Chair of Media Technology and the Chair of Communication Networks, Technical University of Munich from 2016 to 2017. Her research interests include haptic-audio-visual multimodal signal processing and communications, wireless multimedia communications, and human-machine interactions.

**ZHAO Tiesong** received the B.S. degree in electrical engineering from the University of Science and Technology of China in 2006, and the Ph.D. degree in computer science from the City University of Hong Kong, China in 2011. He served as a research associate with the Department of Computer Science, City University of Hong Kong, from 2011 to 2012, a post-doctoral fellow with the Department of Electrical and Computer Engineering, University of Waterloo, Canada from 2012 to 2013, and a research scientist with the Ubiquitous Multimedia Laboratory, State University of New York at Buffalo until 2015. He is currently a professor with the College of Physics and Information Engineering, Fuzhou University, China. His research interests include image/video signal processing, visual quality assessment, and video coding and transmission.



# Comparison Analysis on Feasible Solutions for LTE Based Next-Generation Railway Mobile Communication System

SUN Bin<sup>1</sup>, DING Jianwen<sup>1</sup>, LIN Siyu<sup>1</sup>,  
WANG Wei<sup>2</sup>, CHEN Qiang<sup>2</sup>, and  
ZHONG Zhangdui<sup>1</sup>

(1. Beijing Jiaotong University, Beijing 100044, China;  
2. ZTE Corporation, Shenzhen, Guangdong 518057, China)

DOI: 10.12142/ZTECOM.201901009

<http://kns.cnki.net/kcms/detail/34.1294.TN.20190314.1717.004.html>, published online March 14, 2019

Manuscript received: 2018-01-01

**Abstract:** Wireless communication technologies play an essential role in supporting railway operation and control. The current Global System for Mobile Communications-Railway (GSM-R) system offers a rich set of voice services and data services related with train control, but it has very limited multimedia service bearer capability. With the development of commercial wireless industry, Long-Term Evolution (LTE) mobile broadband technology is becoming the prevalent technology in most of commercial mobile networks. LTE is also a promising technology of future railway mobile communication systems. The 3rd Generation Partner Project (3GPP) and China Communications Standards Association (CCSA) have proposed two feasible LTE based broadband trunking communication solutions: the 3GPP Mission Critical Push to Talk (MCPTT) solution and B-TrunC solution. In this paper, we first introduce the development of railway mobile communications and LTE technology. The user requirements of future railway mobile communication system (FRMCS) are then discussed. We also analyze the suitability of the two LTE-based solutions for LTE based Next-Generation Railway Mobile Communication System (LTE-R) from different aspects.

**Keywords:** 3GPP MCPTT; B-TrunC; FRMCS; LTE-R

## 1 Introduction

Railway mobile communications originated from the incompatibility of track cables and analogue railway radio networks. Due to their limited bearing capability, less anti-interference capability, and lack of encryption, the analogue railway radio networks were replaced by the Global System for Mobile Communications-Railways (GSM-R) systems. GSM-R bears the railway trunking dispatching voice services and the train control data services as well, which fulfills the mobile communication service requirements of railway systems. However, the limited capacity, higher cost and development cycle inherited in GSM has led to the situation that the market of GSM-R is moving inexorably towards its end.

In 2012, International Union of Railways (UIC) started to evaluate the actual situation of global railway market and to study the needs of railway operators and passengers. The second version of "User Requirements Specification" for Future Railway Mobile Communication System (FRMCS) was released

in March 2016, which is an indispensable step towards the introduction of a successor of GSM-R. Besides the existing GSM-R services, several potential new application requirements would also be added into the FRMCS, such as multimedia dispatching communications and real time video monitoring [1]. GSM-R cannot satisfy the users' requirements of FRMCS and GSM industry will be terminated in 2025, which are motivating the design of GSM-R successor.

Long-Term Evolution (LTE) mobile broadband technology is currently the most successful commercial broadband mobile communication system. LTE provides higher data capacity than GSM and presents a flat architecture to reduce deployment and maintenance costs as long as standard entities are used. In this sense, two feasible LTE based broadband trunking communication solutions, 3GPP Mission Critical Push to Talk (MCPTT) [2] and LTE-based broadband trunking communication (B-TrunC) [3], have been proposed by the 3rd Generation Partner Project (3GPP) and China Communications Standards Association (CCSA). MCPTT is a new global standard for replacing the conventional group communication systems, such as Trans European Trunked Radio (TETRA), P25 and others. MCPTT is based on the LTE architecture and will be available for governments/public safety or railway operators. 3GPP is focusing on the network infrastructure design, while the client-side is left up to manufacturers to design and implement ac-

This work was partly supported by ZTE Industry-Academia-Research Cooperation Funds, Fundamental Research Funds for the Central Universities (No. 2016JBM076), the National Natural Science Foundation of China (No. 61501023, No. U1334202, and No. U1534201), and the Project of China Railway Corporation (No. 2016X009-E).



according to the requirements of users. Developed by CCSA, B-TrunC is designed based on LTE Release 9 [4]. The B-TrunC standard has been developed to address the evolving needs driven by the emergence of new trunking communication requirements, such as multimedia dispatch applications.

However, neither MCPTT nor B-TrunC is specially designed for railway mobile communications. In this context, the suitability of the two LTE-based solutions for LTE based FRMCS (LTE-R) will be analyzed in this paper from the following six aspects: architecture, system functionality and performance, standardization level, interoperability, high mobility support capability, and backward compatibility with GSM-R.

The content of this paper is organized as follows. First, a brief introduction of the railway mobile communications and LTE systems are presented in Section 2. Then, the user requirements of FRMCS are discussed in Section 3, where the MCPTT and B-TrunC are introduced. In Section 4, we analyze the suitability of the two feasible solutions for LTE-R. Finally, conclusions are drawn in Section 5.

## 2 Development of Railway Mobile Communications

The high-speed railway (HSR) is becoming the major mode of transportation for a journey and has been developed rapidly worldwide, benefitting from its advantages of safety, reliability, convenience, comfortableness, and low energy consumption. To guarantee the reliability, availability and safety of the train-ground data transmission, railway mobile communication systems play a key role in the HSR operation [1].

In the early stage of railway mobile communications, several analogue radio networks supported mobile communication applications for drivers and trackside workers. For example, British Rail developed the National Radio Network (NRN) specifically used for the operational railways, which provides radio coverage for 98% of the rail network through base stations and radio exchanges. NRN could provide dedicated open channels on talk-through mode for incident management and an override priority facility, in order to ensure that all emergency calls could be immediately connected to the railway's train control offices and electrical control rooms [5].

The analog radio network has reached to its limits due to the rapid growth of communications traffic volume and the increasing demands for security, economy, efficiency and safety of railway traffic.

In 1994, the GSM standard of European Telecommunication Standards Institute (ETSI) was selected by UIC as the first Digital Railways Radio Communication System standard, because it is the sole system in commercial operation and it is already proven with off-the-shelf products available, requiring the minimum modifications. However, GSM could not fulfill all the necessary requirements of the efficient railway services. Therefore it was necessary that the advanced voice call features should

be identified and added to the standard GSM. UIC, together with the railway operators, launched the European Integrated Radio Enhanced Network (EIRENE) to specify the requirements for mobile networks to fulfill the needs of railways including the features of additional group and broadcast calls [6]. To validate whether these EIRENE functional specifications could be transferred into technical implementations, three prototypes were developed by manufacturers and three pilot-lines were planned and realized in France, Italy and Germany. These three pilot-line systems were built to test different aspects of operation, such as railway station environment, complex radio coverage topology with tunnels and bends, and high speed lines at speeds up to 300 km/h [7].

In 1997, 32 railway operators all over Europe signed a Memorandum of Understanding (MoU) to terminate the investment in the analogue radio systems and start to invest the implementation of GSM-R. As of today, the number of signatories has increased to 38, including railway operators outside Europe. Today, over 100 000 km of railway lines are operated by GSM-R systems and this amount is still growing. In addition, the industry has given commitment to support the GSM-R technology until at least 2030 [1].

## 3 User Requirements of Next-Generation Railway Mobile Communication System

As telecommunication standards are evolving and it usually takes a long time to realize the application of any technology specifications, it is urgent to start the research work on the successor of GSM-R. Thus UIC decided to set up the FRMCS project to prepare the necessary steps towards the introduction of a successor of GSM-R in 2012. The project started with the situation evaluation of actual railway systems and investigation on needs of users, and ended with the delivery of the first specification for user requirements of the next-generation railway mobile communication system.

In 2016, a new version of FRMCS user requirement specification was published [1], in which the traffic requirements are classified into two categories including communication applications and supporting applications. Besides, the users are classified into three groups: critical users, performance users, and business users. The critical users refer to the applications that can enhance the reliability, availability, maintainability and safety (RAMS) of railway systems. The performance users refer to the applications that can improve the performance of railway operation, such as train departure and telemetry. The business users refer to the applications that can support the railway business operation in general.

Comparing with the functional requirement specification of GSM-R system, the new version has some new broadband mobile services that are high demanded, such as real time video and wireless Internet on - train for passengers. Furthermore, more train-ground data services are included, such as monitor-

ing and control of non-critical infrastructure and trackside maintenance communications.

All the above mentioned traffic requirements cannot be met by narrow-band mobile communication systems, and the industry support of GSM-R will be ended in 2030. Thus, it is urgent to develop a dedicated broadband mobile communication system as the successor of GSM-R.

Besides the traffic requirements, the fundamental design principles of FRMCS are proposed for user requirements, which include application decoupled with system architecture, interoperability, reuse of the existing infrastructure, high mobility support, backward compatibility with GSM-R, etc. [1]. These principles will guide the design of LTE-R systems.

## 4 Two Feasible Solutions for LTE-R

To provide a wide range of data-centric services, such as video sharing, multimedia dispatching, and ubiquitous Internet and intranet access for governments and organizations involved in public safety and security, two LTE based broadband trunking communication systems, B-TrunC system and 3GPP MCPTT system, are designed. These two systems are considered as the candidates for LTE-R because the user requirements of railway systems are similar to public safety communication services and LTE is the currently most popular 4G mobile broadband system there.

### 4.1 LTE Based Broadband Trunking Communication system (B-TrunC)

To accomplish broadband multimedia trunking service requirements, i.e., group communication demand for voice, data, and video, CCSA initially started to develop a Broadband Trunking Communication (B-TrunC) system in 2012, which has been admitted by ITU-R as the Public Protection and Disaster Relief (PPDR) Recommendation Standard.

The B-TrunC system is designed based on the TD-LTE system with 3GPP Release 9. The single-cell point-to-multipoint (SC-PTM) is the feature of air interface of B-TrunC system, which is designed to realize the group communications. The SC-PTM is a new type of radio access method dedicated to multicast through the Physical Downlink Shared Channel (PDSCH) in a single cell. For SC-PTM transmission, user equipment in a group receives the group data through a common radio resource region in the PDSCH. This concept naturally allows the group data to be multiplexed with the normal unicast data within a PDSCH subframe and thus does not cause the problem of radio resource granularity.

Besides the SC-PTM technology, two trunking communication entities are involved in System Architecture Evolution (SAE) of LTE system, which are Trunking Control Function (TCF) and Trunking Media Function (TMF). TCF is responsible for trunking service scheduling, call setup and release, session management, authentication, registration, and cancella-

tion. TMF is responsible for trunking user plane management, routing, data forwarding, encoding, etc. The interface between terminals and the system, Uu-T, and the interface between the core network and the dispatcher, D, are designed. To enhance the latency performance of B-TrunC, the Multimedia Broadcast/Multicast Service (MBMS) and Push to Talk over Cellular (PoC) technologies of LTE are carried out. Now CCSA has completed the general technical requirements and air interface standards for B-TrunC.

### 4.2 LTE and MCPTT Standardization Roundup

In this section, the 3GPP LTE standardization process related with MCPTT is introduced.

The fully commercial operation of LTE systems started from the 3GPP Release 8 specification, which was finalized in 2008. The latest version of LTE, Release 14 had been frozen by mid-2017.

LTE Release 8 specified one primary broadband technology based on Orthogonal Frequency Division Multiple Access (OFDM). LTE Release 8 is mainly deployed in a macro/micro-cell layout and can provide improved system capacity and coverage, high peak data rates, low latency, reduced operating costs, multi-antenna support, flexible bandwidth operation and seamless integration with existing systems [8]. LTE Release 9 provides some minor enhancements to LTE Release 8 with respect to the air interface. These features include dual-layer beamforming and time difference of arrival based location techniques. To support the video on demand, video conference and other multimedia services, MBMS architecture is included in the Evolved Packet Core (EPC) of LTE.

LTE Release 10 realizes the following features: bandwidth extension via carrier aggregation to support deployment bandwidth up to 100 MHz, downlink spatial multiplexing including single-cell multi-user MIMO transmission and uplink spatial multiplexing, and heterogeneous networks with emphasis on Type 1 and Type 2 relays [9]. For this release, LTE technology refers to LTE-A as a formal 4G system. As the capacity and performance of the LTE traffic channels are progressively improved, the downlink control channels may become a bottleneck. To address this issue, an enhanced physical downlink control channel (EPDCCH) has been introduced in 3GPP LTE Release 11 [10].

The core specification item of Release 12 in terms of group communication is Group Communication System Enabler (GCSE). The conventional physical channels in LTE can be good media for providing group communications in some basic scenarios [11]. GCSE defines the requirements for group communications and proposes system architecture on top of the existing physical channels. The LTE system of Release 12 has two fundamental physical channels for transferring the data: the PDSCH, which is commonly used for normal unicast data, and the physical multicast channel (PMCH), which is designed for evolved MBMS (eMBMS). Furthermore, direct device-to-device

(D2D) communication is the other feature of Release 12.

Releases 13 and 14 focus on the air interface aspects. They are a part of the continued evolution of LTE - Advanced and play a role as a bridge from 4G to 5G. To meet the user requirements of PPDR, the MCPTT services and system architecture are defined, whose technology enhancement and realization has been completed in Release14 [12], [13]. Other techniques in this release include enhancement of D2D proximity services, indoor positioning enhancements, and the single-cell point-to-multipoint (SC-PTM) [14].

## 5 Analysis of Feasible Architecture Solutions of LTE for Railway

Neither 3GPP MCPTT nor B-TrunC is specially designed for Railway communications. They cannot fully fulfill the user requirements of LTE - R. Next, the suitability of the two LTE - based solutions for FRMCS are analyzed from six aspects, including architecture, system functionality and performance, standardization, interoperability, high mobility support capability, and backward compatibility with GSM-R.

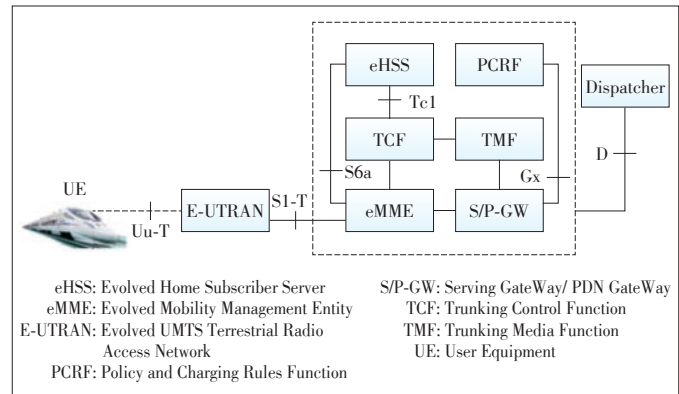
### 5.1 System Architecture

The functionality and commoditization of future railway system determine the design principles of future railway mobile communications system. The FRMCS architecture design should satisfy the services and architecture decoupling principle. As GSM-R is a modified off-the-shelf technology system based on GSM offering, the enhancement to deliver specific “R” (railway) functionality has proven expensive for the railways. For future railway communication system, the specific “R” (railway) functionality provisioning should be decoupled with communication system.

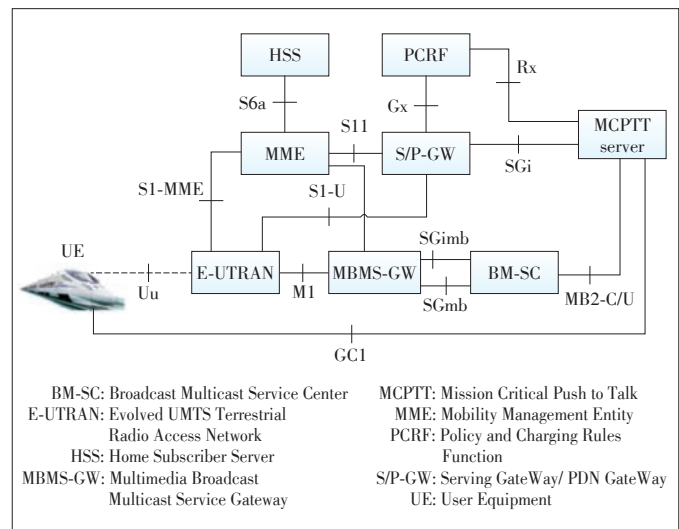
B-TrunC provides the fundamental solution for group communication. To provide the group communication services, several logical channels are designed including Trunking Control Channel (TCCH), Trunking Traffic Channel (TTCH), Trunking Paging Channel (TPCH), and Trunking Paging Control Channel (TPCCH). These channels are mapped to the PDSCH. To implement the trunking service management, two logical function entities, i.e. TCF and TMF, are included in EPC, and the non - access stratum (NAS) signaling messages related with Trunking Service Management (TSM) are designed. The B-TrunC system architecture is shown in **Fig. 1**.

The 3GPP designs the GCSE architecture to realize group communications through the PMCH in MCPTT system based on the eMBMS architecture. The 3GPP MCPTT system architecture is shown in **Fig. 2**. The PMCH allocation is fixed in different uplink-downlink configurations and the trunking applications are provided by the MCPTT server in the application layer.

In summary, B-TrunC redesigns the LTE bearer layer and service layer of Uu and NAS interfaces in LTE system to sup-



▲ Figure 1. B-TrunC system architecture. The grey entities are the special entity designed for trunking services [4].



▲ Figure 2. 3GPP LTE MCPTT system architecture. The grey entities are the special entity designed for trunking services.

port a variety of trunking services. 3GPP MCPTT enhances the bearing capability for trunking services based on the mature commercial LTE product; all the special applications are realized in the application layer, which is independent with the system architecture. Therefore, the 3GPP MCPTT architecture is similar to the decoupled design version of LTE-R.

Because the 3GPP scheme can support the trunking services on the application layer of LTE system, the implementation of railway functional is flexible. GCSE realizes the group communication via fix allocated PMCH, while the PDSCH is used to implement the group communication, so the resource utilization of B-TrunC system is better.

### 5.2 System Functionality and Performance

Besides the narrow band trunking communication services borne in the TETRA system, B-TrunC is designed to bear more multimedia trunking services, such as video group call, video individual call, video forwarding, and delivery to group.

3GPP MCPTT system collects the trunking service require-

ments of PPDR in MCPTT over LTE stage 1, which includes group call, broadband, late call entry, dynamic group call, call prioritization, etc.

For railway trunking communications, functional addressing and location dependent addressing are two typical services. The functionality and technical realization of the two services in the B-TrunC and 3GPP are being standardized. Due to the decoupling design of MCPTT system, more extra undefined new service requirements can be developed in the application layer of communication system according 3GPP specifications, the new service requirements of FRMCS can simply be supported by the 3GPP solution.

The system performance is also an important aspect to evaluate the system suitability. The key performance metrics for trunking communication of B-TrunC and MCPTT system are listed in **Tables 1** and **2**. From the tables we can see that the QoS requirements of B-TrunC are stricter than those of MCPTT system. This is because that B-TrunC has simplified system architecture and technology realization.

### 5.3 Standardization

The standardization process of B-TrunC system is shown in **Fig. 3**. B-TrunC Release 1 was finished in 2014. At this stage, the broadband trunking functionality including group multimedia services is enhanced in a local network system, and the radio interface and interface between core network and dispatcher, i.e. Uu and D, are standardized respectively. The network interoperability and roaming architecture of B-TrunC is designed in Release 2, which was finished in the beginning of 2017. Besides that, more technical realizations for rail trans-

▼Table 1. Latency requirements of B-TrunC system

Performance metrics	Value
Group call access time	<300 ms
Full duplex unicast call access time	<500 ms
Half-duplex unicast call access time	<500 ms
Speak right apply time	<200 ms
Group call capacity	7.5 voice groups/(cell·MHz)

▼Table 2. Latency requirements of MCPTT system

Performance metrics	Value
MCPTT access time	<300 ms
End-to-end MCPTT access time	<1 000 ms
Mouth-to-ear latency	<300 ms
Maximum late call entry time (without application layer encryption)	< 150 ms
Maximum late call entry time (with application layer encryption)	<300 ms

MCPTT: Mission Critical Push to Talk

portation communication service requirements would be considered at this stage. In Release 3, the additional functionalities including device direct communication and cognitive radio had been considered by the end of 2017.

The standardization process of 3GPP MCPTT system is also shown in **Fig. 3**. 3GPP Release 13 was frozen in 2016; the functional architecture and signaling flows to support MCPTT services were finished and eMBMS functionality were enhanced. In Release 14, more mission critical video and data

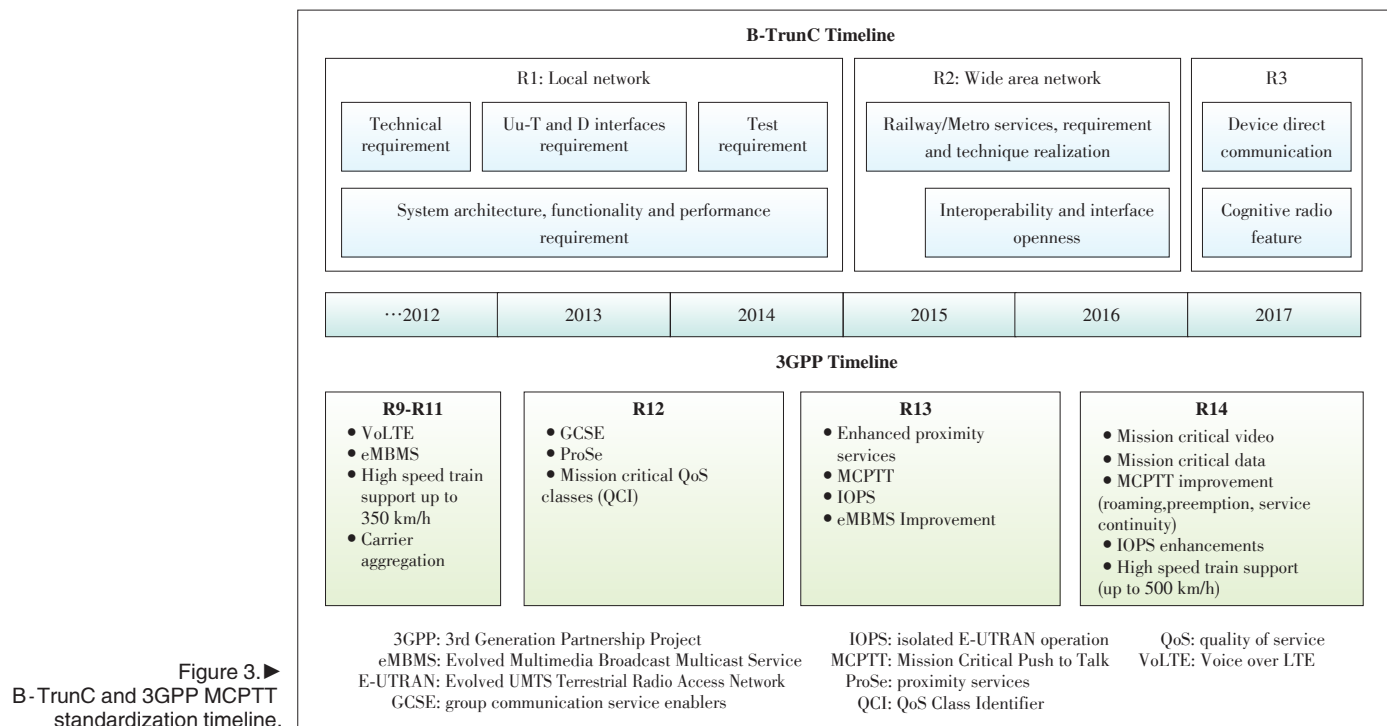


Figure 3. ▶  
B-TrunC and 3GPP MCPTT  
standardization timeline.



service requirements has been considered. The roaming, pre-emption, service continuity of MCPTT services were enhanced in this version.

In summary, in terms of multimedia trunking communication functionality, B-TrunC has realized the multimedia group communication at stage 1, so B-TrunC is developed ahead of 3GPP MCPTT. However, by the end of 2017, the two systems had similar functionalities.

### 5.4 Interoperability

In most of countries and regions, such as Europe and China, different rail lines are operated by different railway operators. For example, two railway operators, i.e. Guangzhou Railway and Hong Kong Railway Corporation Limited (MTR), manage the rails from Guangzhou to Shenzhen and to Hong Kong. The radio access subsystem in the Hong Kong section has to interconnect with the core network layer in the Guangzhou-Shenzhen section. Because the signaling of the two sections is provided by different vendors, the device interoperability and interface openness are mandatory requirements for the railway mobile communications. Therefore, GSM-R requires the interoperability as a mandatory feature, which should be inherited by LTE-R.

For 3GPP R13 specifications, besides the standardized interfaces, the interfaces of group communication entities are also standardized, such as the MB2 interface between the MCPTT server and Broadcast Multicast-Service Centre, M1 interface between the Evolved UMTS Terrestrial Radio Access Network (E-UTRAN) and multimedia broadcast/multicast service gateway (MBMS-GW), and GC1 interface between MCPTT client and MCPTT server.

For the B-TrunC system, the Uu-T interface between the Trunking UE and eNode-B and the D interface between Trunking core network and dispatcher are defined in Release 1. The enhanced S1 interface and the interfaces of Trunking core network are standardized in Release 2.

In summary, as the 3GPP is a global mobile telecommunications standardization organization, MCPTT specifications have better interoperability and interface openness.

### 5.5 High Mobility Support Capability

High speed trains will be operated up to 500 km/h in Europe and China in the near future. The severe Doppler shift of uplink and downlink signals and frequently handover have great impact on the reliability of services transmission, so LTE-R requires the mandatory high mobility support.

For LTE system, in terms of mobility, 3GPP Release 9 is designed aiming primarily at low mobility applications in the range of 0 to 15 km/h to show the highest system performance. The system is capable of working well at higher speeds from 15 km/h to 120 km/h and providing the basic functional support from 120 km/h to 350 km/h. In Release 14, the high speed railway is considered as a typical scenario, so the high mobility

support is required to achieve up to 500 km/h. In Release 14, the frequency offset correction algorithm and multi-RRU co-cell sharing technology are adopted to improve the system performance with high speed railway scenarios.

To evaluate the effects of high mobility on the B-TrunC system and MCPTT system, we set up the hard-in-the-loop simulation system in the lab. Because there is no MCPTT products currently, the MCPTT is designed based on the Frequency Division Duplex (FDD) system. Therefore, the FDD LTE system is used to illustrate the performance evaluation results of MCPTT and the simulation results are shown in **Table 3**. The simulation results show that the throughput and block error rate of FDD LTE are a little better than B-TrunC, and the two systems have similar performance in terms of ping delay. Therefore, the MCPTT system has better performance in a high mobility environment.

### 5.6 Backward Compatibility with GSM-R

LTE-R requires backward compatibility with GSM-R. On one hand, it means that the borne services can be automatic transferred between GSM-R network and LTE-R. On the other hand, it should be enabled to reuse the existing equipment of GSM-R which has not reached the end of its lifecycle, such as the towers, hardware of base stations, and on-board devices.

The two feasible solutions are designed based on the commercial LTE system, so the backward compatibility with GSM system can be guaranteed. GSM-R was designed by the ETSI but involved based on 3GPP specifications, so the backward compatibility of MCPTT with GSM-R is considered in 3GPP Release 14 and 15 specifications. For the B-TrunC system, there is no plan to investigate the backward compatibility with GSM-R. Besides, the most GSM-R macro base station products support a seamless upgrade to LTE eNode B that satisfies 3GPP specifications.

## 6 Conclusions

The development of LTE technology offers an excellent opportunity to improve both the performance and capabilities of railway mobile communication systems. In this work, we analyzed the suitability of the B-TrunC and 3GPP MCPTT on LTE-

▼Table 3. Performance evaluation of B-TrunC and MCPTT

Speed/(km/h)	Downlink throughput/(Mbit/s)	Uplink throughput/(Mbit/s)	Downlink BLER	Ping delay/ms
B-TrunC				
100	13.3	8.12	1.12%	15
300	11.1	7.76	3.06%	17
LTE FDD				
100	15.2	9.2	0.76%	17
300	12.3	7.1	2.31%	18

BLER: Block Error Rate FDD: Frequency Division Duplex LTE: Long Term Evolution

R. In terms of system functionality, performance, and standardization process, the B-TrunC system keeps ahead of MCPTT system. In terms of system architecture, interoperability, high mobility support capability, and backward compatibility with GSM-R, the MCPTT system performs better than B-TrunC. Besides, 3GPP MCPTT is an international standard, so LTE-R is designed based on the 3GPP MCPTT which is benefit for globalization of railways.

## References

- [1] International Union of Railways. Future Railway Mobile Communications System User Requirements Specification [EB/OL]. (2015). <http://www.uic.org/IMG/pdf/frmcuser-requirements.pdf>
- [2] DIAZ ZAYAS A, GARCIA PEREZ C A, GOMEZ P M. Third-Generation Partnership Project Standards: For Delivery of Critical Communications for Railways [J]. IEEE Vehicular Technology Magazine, 2014, 9(2): 58–68. DOI: 10.1109/MVT.2014.2311592
- [3] LI S, CHEN Z, YU Q, MENG W, TAN X. Toward Future Public Safety Communications: The Broadband Wireless Trunking Project in China [J]. IEEE Vehicular Technology Magazine, 2013, 8(2): 55–63. DOI: 10.1109/MVT.2013.2252277
- [4] China Academy of Telecommunication Research (CATR). The Progress of B-TrunC Broadband Trunking Standard Based on TD-LTE [EB/OL]. (2014). [http://enterprise.huawei.com/link/enenterprise/download/HW\\_328557](http://enterprise.huawei.com/link/enenterprise/download/HW_328557)
- [5] Ian Allan Publishing. Rail Radio Revolution [J]. Modern Railways, 2004, 61 (673): 65–67
- [6] EIRENE—Functional Requirements Specification Version 8.0.0 [EB/OL]. (2015). [https://uic.org/IMG/pdf/frs-8.0.0\\_uic\\_950\\_0.0.2\\_final.pdf](https://uic.org/IMG/pdf/frs-8.0.0_uic_950_0.0.2_final.pdf)
- [7] EIRENE—System Requirements Specification Version 16.0.0 [EB/OL]. (2015). [https://uic.org/IMG/pdf/srs-16.0.0\\_uic\\_951-0.0.2\\_final.pdf](https://uic.org/IMG/pdf/srs-16.0.0_uic_951-0.0.2_final.pdf)
- [8] GHOSH A, RATASUK R, MONDAL B, et al. LTE-Advanced: Next-Generation Wireless Broadband Technology [J]. IEEE Wireless Communications, 2010, 17 (3): 10–22. DOI: 10.1109/MWC.2010.5490974
- [9] PARKVALL S, FURUSKAR A, DAHLMAN E. Evolution of LTE Toward IMT-Advanced [J]. IEEE Communications Magazine, 2011, 49(2): 84–91. DOI: 10.1109/MCOM.2011.5706315
- [10] YE S G, WONG S H, WORRALL C. Enhanced Physical Downlink Control Channel in LTE Advanced Release 11 [J]. IEEE Communications Magazine, 2013, 51(2): 82–89. DOI: 10.1109/MCOM.2013.6461190
- [11] ASTELY D, DAHLMAN E, FODOR G, et al. LTE Release 12 and Beyond [J]. IEEE Communications Magazine, 2013, 51(7): 154–160. DOI: 10.1109/MCOM.2013.6553692
- [12] LEE J, KIM Y, KWAK Y, et al. LTE-Advanced in 3GPP Rel-13/14: an Evolution Toward 5G [J]. IEEE Communications Magazine, 2016, 54(3): 36–42. DOI: 10.1109/MCOM.2016.7432169
- [13] HOYMANN C, ASTELY D, STATTIN M, et al. LTE Release 14 Outlook [J]. IEEE Communications Magazine, 2016, 54(6): 44–49. DOI: 10.1109/MCOM.2016.7497765
- [14] KIM J, CHOI S W, SHIN W, et al. Group Communication over LTE: a Radio Access Perspective [J]. IEEE Communications Magazine, 2016, 54(4): 16–23. DOI: 10.1109/MCOM.2016.7452261

## Biographies

**SUN Bin** received the B.S. degree in electronic engineering and the M.S. degree in electronic engineering from Beijing Jiaotong University, China in 2004 and 2007, respectively. From 2007 to 2015, he was a R&D manager with Beijing LiuJie Technology Co., Ltd. He is currently an assistant researcher with National Research Center of Railway Safety Assessment, Beijing Jiaotong University. His main research interest is the interconnection and interworking of core network for dedicated railway mobile communication system.

**DING Jianwen** (jwding@bjtu.edu.cn) received his B.S. and M.S. degrees from Beijing Jiaotong University, China in 2002 and 2005, respectively. He is currently an associate researcher with National Research Center of Railway Safety Assessment, Beijing Jiaotong University. He received the second prize of progress in science and technology of the Chinese Railway Society. His research interests are broadband mobile communication and personal communication, dedicated mobile communication system for railway, and safety communication technology for train control system.

**LIN Siyu** received the B.S. degree in electronic engineering and Ph.D. degree in electronic engineering from Beijing Jiaotong University, China in 2007 and 2013, respectively. From 2009 to 2010, he was an exchange student with the Universidad Politécnica de Madrid, Spain. From 2011 to 2012, he was a visiting student with the University of Victoria, Canada. He is currently an associate professor with Beijing Engineering Research Center of High-speed Railway Broadband Mobile Communications, Beijing Jiaotong University. His main research interest is performance analysis and channel modeling for wireless communication networks, dedicated railway mobile communication system.

**WANG Wei** is the LTE-R technical director and a railway wireless communication system expert of ZTE Corporation, with rich experience of the GSM-R system design. He has a deep understanding of GSM-R and LTE-R and has undertaken several major railway-related projects on wireless communication systems.

**CHEN Qiang** is a LTE-R product manager and railway wireless communication system expert of ZTE Corporation, with rich experience of railway communications. He is familiar with international and domestic railway standards, service applications, and service processes.

**ZHONG Zhangdui** is a professor and supervisor of Ph.D. candidates at Beijing Jiaotong University, China. He is now a Chief Scientist of the State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University. He is also a director of the Innovative Research Team of Ministry of Education of China and a Chief Scientist of Ministry of Railways of China. He is an executive council member of the Radio Association of China and a deputy director of Radio Association of Beijing, China. His research interests are wireless communications for railways, control theory and techniques for railways, and GSM-R system. He has authored/co-authored seven books, five invention patents, and over 200 scientific research papers in his research areas. He received the MaoYiSheng Scientific Award of China, ZhanTianYou Railway Honorary Award of China, and Top 10 Science/Technology Achievements Award of Chinese Universities.



# **ZTE Communications Guidelines for Authors**

## **Remit of Journal**

*ZTE Communications* publishes original theoretical papers, research findings, and surveys on a broad range of communications topics, including communications and information system design, optical fiber and electro-optical engineering, microwave technology, radio wave propagation, antenna engineering, electromagnetics, signal and image processing, and power engineering. The journal is designed to be an integrated forum for university academics and industry researchers from around the world.

## **Manuscript Preparation**

Manuscripts must be typed in English and submitted electronically in MS Word (or compatible) format. The word length is approximately 3000 to 8000, and no more than 8 figures or tables should be included. Authors are requested to submit mathematical material and graphics in an editable format.

## **Abstract and Keywords**

Each manuscript must include an abstract of approximately 150 words written as a single paragraph. The abstract should not include mathematics or references and should not be repeated verbatim in the introduction. The abstract should be a self-contained overview of the aims, methods, experimental results, and significance of research outlined in the paper. Five carefully chosen keywords must be provided with the abstract.

## **References**

Manuscripts must be referenced at a level that conforms to international academic standards. All references must be numbered sequentially in-text and listed in corresponding order at the end of the paper. References that are not cited in-text should not be included in the reference list. References must be complete and formatted according to ZTE Communications Editorial Style. A minimum of 10 references should be provided. Footnotes should be avoided or kept to a minimum.

## **Copyright and Declaration**

Authors are responsible for obtaining permission to reproduce any material for which they do not hold copyright. Permission to reproduce any part of this publication for commercial use must be obtained in advance from the editorial office of *ZTE Communications*. Authors agree that a) the manuscript is a product of research conducted by themselves and the stated co-authors, b) the manuscript has not been published elsewhere in its submitted form, c) the manuscript is not currently being considered for publication elsewhere. If the paper is an adaptation of a speech or presentation, acknowledgement of this is required within the paper. The number of co-authors should not exceed five.

## **Content and Structure**

*ZTE Communications* seeks to publish original content that may build on existing literature in any field of communications. Authors should not dedicate a disproportionate amount of a paper to fundamental background, historical overviews, or chronologies that may be sufficiently dealt with by references. Authors are also requested to avoid the overuse of bullet points when structuring papers. The conclusion should include a commentary on the significance/future implications of the research as well as an overview of the material presented.

## **Peer Review and Editing**

All manuscripts will be subject to a two-stage anonymous peer review as well as copyediting, and formatting. Authors may be asked to revise parts of a manuscript prior to publication.

## **Biographical Information**

All authors are requested to provide a brief biography (approx. 100 words) that includes email address, educational background, career experience, research interests, awards, and publications.

## **Acknowledgements and Funding**

A manuscript based on funded research must clearly state the program name, funding body, and grant number. Individuals who contributed to the manuscript should be acknowledged in a brief statement.

## **Address for Submission**

<http://mc03.manuscriptcentral.com/ztecom>

# ZTE COMMUNICATIONS

中兴通讯技术(英文版)

**ZTE Communications has been indexed in the following databases:**

- Abstract Journal
- Cambridge Scientific Abstracts (CSA)
- China Science and Technology Journal Database
- Chinese Journal Fulltext Databases
- Index of Copernicus
- Inspec
- Ulrich's Periodicals Directory
- Wanfang Data

---

## ZTE COMMUNICATIONS

Vol. 17 No. 1 (Issue 65)

Quarterly

First English Issue Published in 2003

### **Supervised by:**

Anhui Publishing Group

### **Sponsored by:**

Time Publishing and Media Co., Ltd.

Shenzhen Guangyu Aerospace Industry Co., Ltd.

### **Published by:**

Anhui Science & Technology Publishing House

### **Edited and Circulated (Home and Abroad) by:**

Magazine House of ZTE Communications

### **Staff Members:**

Editor-in-Chief: WANG Xiyu

Associate Editor-in-chief: JIANG Xianjun

Executive Associate Editor-in-Chief: HUANG Xinming

Editor-in-Charge: ZHU Li

Editors: XU Ye and LU Dan

Producer: YU Gang

Circulation Executive: WANG Pingping

Assistant: WANG Kun

---

### **Editorial Correspondence:**

Add: 12F Kaixuan Building, 329 Jinzhai Road,  
Hefei 230061, P. R. China

Tel: +86-551-65533356

Email: magazine@zte.com.cn

Online Submission: <https://mc03.manuscriptcentral.com/ztecom>

**Annual Subscription:** RMB 80

### **Printed by:**

Hefei Tiancai Color Printing Company

**Publication Date:** March 25, 2019

**Publication Licenses:** ISSN 1673-5188  
CN 34-1294/ TN