www.zte.com.cn/magazine/English

ISSN 1673–5188
CODEN ZCTOAK

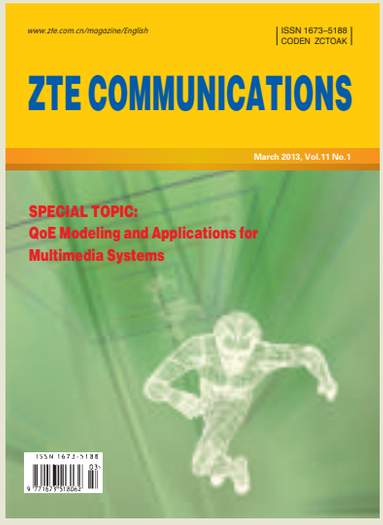
ZTE COMMUNICATIONS

March 2013, Vol.11 No.1

SPECIAL TOPIC:
QoE Modeling and Applications for Multimedia Systems

ISSN 1673-5188
03>
9 771673 518062

CONTENTS



Special Topic

QoE Modeling and Applications for Multimedia Systems

CONTENTS

ZTE COMMUNICATIONS
Vol. 11 No.1 (Issue 37)
Quarterly
First English Issue Published in 2003

Supervised by:
Anhui Science and Technology Department

Sponsored by:
ZTE Corporation and Anhui Science and Technology Information Research Institute

Editorial Correspondence:
Add: 12/F Kaixuan Building,
329 Jinzhai Road,
HeFei 230061, P. R. China
Tel: +86-551-65533356
Fax: +86-551-65850139
Email: magazine@zte.com.cn

Published and Circulated (Home and Abroad) by:
Editorial Office of *ZTE Communications*

Printed by:
Hefei Zhongjian Color Printing Company

Publication Date:
March 25, 2013

Publication Licenses:

ISSN 1673-5188
CN 34-1294/ TN

Advertising License:
皖合工商广字0058号

Annual Subscription Rate:
RMB 80

Research Papers

Roundup

QoE Modeling and Applications for Multimedia Systems
Wenjun Zeng and Weisi Lin

QoE Modeling and Applications for Multimedia Systems

▶ Wenjun Zeng

Wenjun Zeng (zengw@missouri.edu) is a professor and the director of the Mobile Networking and Multimedia Communications Lab in the Computer Science Department, University of Missouri. He received his BE degree from Tsinghua University, his MS degree from the University of Notre Dame, and his PhD degree from Princeton University. His research interests include mobile computing, social media analysis, semantic search, distributed source/video coding, 3-D analysis and coding, multimedia networking, and content and network security. He is the editor of *Multimedia Security Technologies for Digital Rights Management* (Elsevier, 2006) and has been granted 15 US patents.

Prior to joining the University of Missouri in 2003, he worked for PacketVideo Corp., Sharp Labs America, Bell Labs, and Panasonic Technology. He is an associate editor of *IEEE Transactions on Information Forensics and Security*, *IEEE Transactions on Circuits and Systems for Video Technology*, and *IEEE Multimedia Magazine*. He is also on the Steering Committee of *IEEE Transactions on Multimedia*. He is a fellow of the IEEE.

▶ Weisi Lin

Weisi Lin (wslin@ntu.edu.sg) received his PhD from King's College, London. He was the lab head of visual processing at Infocomm Research, Singapore, and also acting manager of the media processing department at the same institute. Currently, he is an associate professor in the School of Computer Engineering, Nanyang Technological University, Singapore. His research interests include image processing, perceptual quality evaluation, video compression, multimedia communication, and computer vision. He has published more than 200 refereed papers in international journals and conferences proceedings.

He is on the editorial boards of *IEEE Transactions on Multimedia*, *IEEE Signal Processing Letters*, and *Journal of Visual Communication and Image Representation*. In 2012, he was the lead guest editor of a special issue of *IEEE Journal of Selected Topics in Signal Processing* on perceptual signal processing. He chairs the IEEE MMTC Special Interest Group on Quality of Experience and is an elected Distinguished Lecturer of APSIPA (2012/3). He holds the PCM 2012 Lead Technical Program Chair and a Technical Program Chair for IEEE ICME 2013. He is a fellow of Institute of Engineering Technology and an honorary fellow of the Singapore Institute of Engineering Technologists.

Improving the quality and experience perceived by the user is fundamental when developing multimedia technologies, products, and services. Quality of experience (QoE) involves subjective perception, user behavior and needs, appropriateness, context, and usability of delivered content. Modeling QoE is critical for enhancing QoE in various multimedia applications. In this special issue, we present the latest developments, trends, challenges, and practices in QoE modeling and applications for multimedia systems. The seven expert papers in this special issue come from academia and industry. They present some of latest developments in QoE modeling and assessment for emerging scenarios such as 3D video, streaming and cloud systems, user generated content, and mobile user experience.

We start with two papers that address the problems in 3D QoE assessment and perceptually-driven compression. In "Methodologies for Assessing 3D QoE: Standards and Explorative Studies," Chen et al. describe the fundamentals of existing subjective video quality assessment methods that are the starting point for 3DTV QoE assessment. The authors discuss potential methods for assessing QoE in stereoscopic 3DTV, focusing mainly on multidimensional QoE indicators and common features of subjective assessment. In "3D Perception Algorithms: Towards Perceptually Driven Compression of 3D Video," Hu et al. highlight the differences in perceptual effects between 2D and 3D video. They then share their ideas about 3D video coding and transmission, taking into consideration 3D visual attention, 3D just-noticeable-difference, and 3D texture synthesis modeling. We hope that these two papers prompt further thinking about emerging 3D signal processing.

QoE estimation and modeling has been an important tool for improving user experience in multimedia communication systems. In "Estimating Reduced-Reference Video Quality for Quality-Based Streaming Video," Atzori et al. analyze reduced-reference algorithms for modeling signal distortion, modeling the human visual system, and analyzing the video signal source. The authors then discuss the practical use of these reduced-reference techniques for monitoring and controlling quality in streaming video systems. As the mobile cloud computing paradigm emerges, QoE has become a much more important issue to investigate. In "Human-Centric Composite-Quality Modeling and Assessment for Virtual Desktop Clouds," Xu et al. propose a novel reference architecture and discuss its use in modeling and assessing objective user QoE within virtual desktop clouds. This architecture avoids the need for expensive and time-consuming subjective evaluation.

With the widespread use of smartphones, digital cameras, imaging software, photo-sharing sites, and social networks, the amount of user-generated content has grown tremendously. In "Assessing the Quality of User-Generated Content," Winkler compares the traditional approaches to assessing quality of user-generated content with new approaches. Some sample applications are also discussed. In "An Improved Color Cast Detection Method Based on Ab-Chromaticity Histogram," Lu et al. propose a new method for evaluating the quality of an image in order to improve color cast detection. This is a necessary step before further image processing, such as white balance, is applied.

Energy consumption is a big issue for mobile devices and services. In "Battery Voltage Discharge Rate Prediction and Video Content Adaptation in Mobile Devices on 3G Access Networks," Mkwawa and Sun propose a way of performing visual content adaptation that saves energy. A regression model is used to predict the battery voltage discharge rate in VoIP applications. This is an interesting attempt. Optimizing user experience with a limited battery is challenging for practical system design (starting from algorithm development).

The guest editorial team would like to thank all authors for submitting their high-quality work to this special issue. We would also like to thank the reviewers whose hard work and expert contributions have ensured the quality of this issue. We hope you enjoy reading these fine quality papers.

Methodologies for Assessing 3D QoE: Standards and Explorative Studies

Wei Chen[1], Jérôme Fournier[2], Marcus Barkowsky[3], and Patrick Le Callet[3]
(1. Skype, Stockolm, Sweden;
2. Orange Labs, France Télécom, 4 rue du Clos Courtel, 35512 Cesson-Sevigne, France;
3. RCCyN UMR 6597 CNRS, Ecole Polytechnique del' Université de Nantes, rue Christian Pauc, La Chantrerie, 44306 Nantes, France)

Abstract

Mastering quality of experience (QoE) is key to the widespread adoption of stereoscopic 3DTV (S-3DTV). However, assessing QoE of S-3DTV is not straightforward. Methods for determining observer experience need to be clearly defined and sufficiently robust. In this paper, we present state-of-the-art subjective QoE assessment for S-3DTV. We present conventional standardized ITU recommendations for evaluating picture quality and discuss new ITU activities in the area of S-3DTV assessment. We also present and discuss explorative studies from the literature. We then introduce ways of using conventional quality assessment for S-3DTV QoE assessment. In discussing our proposal, we mainly focus on QoE indicators and common features of subjective assessment. Multidimensional QoE indicators need to be used in S-3DTV to highlight advantages and reveal problems. In the second part of our proposal, we discuss the requirements for adapting ITU-R BT.500, a conventional subjective QoE assessment method, ITU-R BT.500, for assessing QoE of S-3DTV are presented.

Keywords

stereoscopic 3DTV; quality of experience; subjective assessment

1 Introduction

Stereoscopic 3D television (S-3DTV) has created new technical challenges, especially in the provision of good quality of experience (QoE) along the delivery chain. S-3DTV has been rigorously marketed, and many people now have 3D-capable displays. However, the take-up of 3D content is still low. People still do not naturally prefer to watch 3D content. Mastering QoE is crucial for the widespread acceptance and success of S-3DTV.

QoE assessment is not only important in the selection of video bitrates, S-3DTV display techniques, and video encoders in the specification process; it is also important for producing 3D content that provides real added value compared with 2D. Evaluating good QoE in S-3DTV is an urgent and important task. In both academia and industry, subjective assessment has been the most direct way of evaluating QoE. This involves using well-defined methods to conduct experiments with observers. However, subjective assessments are mainly focused on picture or video quality. In S-3DTV, the criterion of picture quality mainly relates to the structural and textual characteristics of 3D pictures and does not, by itself, encompass all the visual characteristics that need to be taken into account to ensure good QoE. It does not include enhanced depth perception and visual comfort. In moving from 2D to 3D, testing conditions such as viewing distance, display calibration, and content selection need to be reviewed.

2 ITU Recommendations and the Foundations of Video Quality Assessment

Standardized subjective quality assessment has a long history. In 1974, the ITU published ITU-R BT.500 *Methodology for the subjective assessment of the quality of television pictures*. This recommendation has been revised several times and is still the most widely used recommendation on image quality assessment. In 2007, ITU published ITU-R BT.1788 *Methodology for the subjective assessment for video quality in multimedia application* [1]. This describes non-interactive subjective methods for evaluating the quality of multimedia and data broadcasting applications comprising video, audio, still pictures, text, and graphics. The main difference between ITU-R BT.500 and ITU-R BT.1788 is that ITU-R BT.500 is for subjective assessment of television pictures (large video format) and ITU-R BT.1788 is for subjective assessment of video quality for multimedia (reduced picture format).

ITU-R BT.500 specifies the common features and methods for subjective quality assessment (Table 1). Common features are the general conditions necessary to conduct subjective quality assessment. The assessment method refers to the protocol used to evaluate a particular question in a subjective quality assessment. ITU-R BT.1788 shares some specifications of ITU-R BT.500, but some features are adapted for multimedia application. For example, there is more flexibility with the viewing distance, which can be constrained or unconstrained.

2.1 ITU Common Features

To avoid unreliable results in subjective assessment, ITU-R BT.500 specifies the following:

Table 1. Specification of subjective quality assessment in ITU–R BT.500

Common features	General viewing condition
	Source signals
	Selection of test materials
	Range of conditions and anchoring
	Observers
	Instruction for the assessment
	The test session
	Presentation of the results
Assessment method	Particular method should be used to address particular assessment problems.

- general viewing condition. Environment luminance (room lighting and background chromaticity) screen luminance, display brightness and contrast calibration, display resolution review, viewing observation angle, and viewing distance are specified.
- source signals. These should be of optimum quality for the television standard used. To obtain stable results, it is crucial that there are no defects in the reference part of the presentation pair. The source signals are directly shown to the observer as the reference picture or they are input into the system being tested.
- selection of test materials. The number and type of test scenes are critically important for interpreting the results of the subjective assessment. New systems often depend heavily on the content of scenes or sequences. The number and type of test scenes should be selected to provide a reasonable generalization to normal programming. The spatial and temporal perceptual characteristics of a scene can be measured to determine the complexity of a scene.
- range of conditions and anchoring. Most assessment methods are sensitive to variation in the range and distribution of visible conditions; therefore, in viewing sessions, the full range of distortions being tested (or extreme examples as anchors) should be shown to cover the wide range in quality.
- observers. There should be at least 15 non–expert observers who are screened for visual acuity, color vision, and other visual anomalies prior to a viewing session.
- instruction for the assessment. Assessors should be carefully briefed on the method of assessment, types of impairment or quality factors likely to occur, grading scale, and timing. Training sequences should demonstrate the range and type of impairments being assessed. The training sequences should not be the same scenes as those used in the actual test but should have comparable content and degradation.
- test session. A test session should last up to half an hour. Dummy presentations should be used to stabilize the observer's opinion. If several sessions are necessary, the presentations should be random, but the under test conditions, the presentations should be ordered so that any effects on the grading of tiredness or adaption are balanced out from session to session.
- presentation of the results. This must include details of the test configuration, test materials, type of picture source and display monitors, number and type of assessors, reference system used, grand mean score for the experiment, original and adjusted mean scores, and 95% confidence interval.

2.2 ITU Quality-Assessment Methods and Scales

There are two classes of subjective assessment: quality assessment and impairment assessment. The former establishes the performance of a system in optimal conditions; the latter establishes the ability of a system to retain quality in non–optimal conditions.

ITU–R BT.500 also provides a collection of methods for different assessment problems. In general, four different methods are proposed to assess the quality of still images or short video sequences of 10 seconds. These methods are double–stimulus–continuous–quality–scale (DSCQS), double–stimulus impairment scales (DSIS), single–stimulus, and stimulus–comparison. The recommended rating scales for these methods are shown in Table 2.

In DSCQS, observers assess the overall image quality from a series of image pairs, each of which comprises an unimpaired (reference) and an impaired image (test). The two images are presented one by one, each for 10 seconds. This process is repeated twice. During the second run through, observers are asked to rate the overall quality of each image. The presentation structure is shown in Fig. 1. DSIS is similar to DSCQS but involves the use of impairment scales.

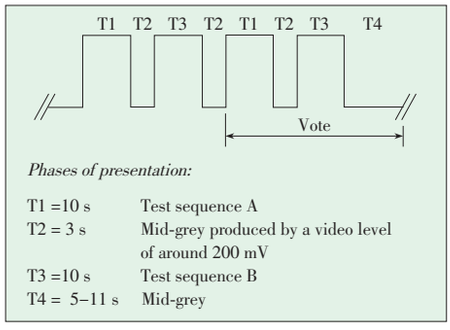
In the single–stimulus method, observers assess the quality of each image in the stimulus set individually. In stimulus–comparison scaling, a series of image pairs, including all possible combinations of the two images in the stimulus set or just a sample of all possible image pairs, are presented. Observers compare the two images in each image pair and assign a relationship using a comparison scale (Table 2).

For longer video sequences of between 60 s and 20 mins,

Table 2. ITU–R BT.500 recommendation rating scales

DSCQS Continuous Quality Scale		Comparison Scale of Stimulus–Comparison	
A B		–3	Much worse
Excellent		–2	Worse
Good		–1	Slightly worse
Fair		0	The same
Poor		+1	Slightly better
Bad		+2	Better
		+3	Much better

Single Stimulus Quality Scale		DSIS and Single Stimulus Impairment Scale	
5	Excellent	5	Imperceptible
4	Good	4	Perceptible, but not annoying
3	Fair	3	Slightly annoying
2	Poor	2	Annoying
1	Bad	1	Very annoying



◀Figure 1. Presentation structure of DSCQS and DSIS Variant II according to ITU–R BT.500.

single–stimulus continuous quality evaluation (SSCQE) and simultaneous double stimulus for continuous evaluation (SDSCE) methods are suggested.

In SSCQE, observers continuously assess the quality of a long video sequence by moving a handset slider. The slider is time sampled, typically at two samples per second. Its range is usually 0 to 100 and corresponds to the DSCQS continuous quality scales. SSCQE is used to assess video that contains scene–dependent and time–varying impairments.

SDSCE is similar to DSCQE, but two stimuli are presented at the same time. SDSCE is used to judge the difference in fidelity between the reference video sequence and the test sequence. When the fidelity is perfect, the slider should be at 100; when there is no fidelity, the slider should be at 0.

In ITU–R BT.1788, subjective assessment methodology for video quality (SAMVIQ) is proposed for assessing the video part of multimedia codecs or systems. SAMVIQ derives from DSCQS, which can be used to efficiently assess a large range of image qualities because it provides reliable discrimination at both high and low quality levels [3].

SAMVIQ allows both hidden and explicit references in a multi–stimulus test environment. Fig. 2 shows SAMVIQ test organization. All the stimuli are accessible in a multi–stimulus form. Besides the explicit reference, all the stimuli (with hidden reference and different algorithms) are randomly ordered. The observer can choose the order of viewing the stimuli, review them, and change ratings if necessary. Each stimulus is compared to an explicit reference in order to determine the best quality that can be achieved in the test. The observer gives a rating using a slider that is graded from 0 to 100 and corresponds to a rating of bad, poor, fair, good and excellent. A maximum of 15 s is necessary to get a stable, reliable quality score for each stimulus [3], [4]. The quality evaluation is carried out scene after scene.

3 Subjective QoE Assessments for S–3DTV: ITU Activities and Explorative Studies

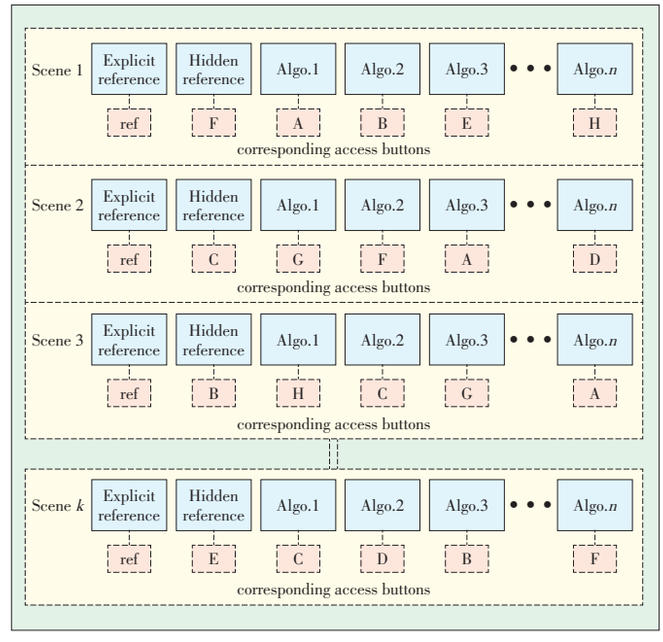
3.1 ITU Evolution Towards Quality Assessment of S–3DTV

The original ITU–R BT.500 specification does not cover S–3DTV assessment. In 2000, ITU published ITU–R BT.1438: *Subjective assessment of stereoscopic television pictures* [5]. This standard describes

- assessment factors. General factors such as resolution, color rendition, motion portrayal, overall quality, and sharpness, are assessed in monoscopic television pictures. To these are added new factors, such as depth resolution, depth motion, puppet theatre effect, and cardboard effect, which are specific to stereoscopic television.
- assessment methods. The methods of ITU–R BT.500 can be used for evaluating the quality of stereoscopic images or videos.
- viewing conditions. The display frame effect (i.e. windows violation), inconsistency between accommodation and convergence (minimum value of depth of focus as ±0.3 diopters), and camera parameters (camera separation, camera convergence angle, focal length of lens) should be taken into account when determining viewing conditions.
- observers. Besides vision tests mentioned in ITU–R BT.500, stereopsis test should be conducted to screen observers.
- test materials.

The ITU–R BT.1438 standard is still does not specify many new characteristics of S–3DTV and how to assess them. Thus, ITU–R SG6 WP6C and ITU–T SG9 have addressed Question Q.2 and Q.12, respectively, for finding a more adequate way to assess S–3DTV. The recent recommendations (draft) from ITU–R SG6 WP6C and ITU–T SG9 are listed in the Table 3[6].

The Video Quality Expert Group (VQEG) has been an active contributor to most of the questions of ITU–T SG9. VQEG established a new project called 3DTV to investigate how to subjectively assess 3DTV video quality. The most recent ITU rec-



▲Figure 2. SAMVIQ test organization [3].

▼Table 3. Recommendation for subjective assessment of S–3DTV

Recommendation	Title	Content
ITU–R BT.[3DTV SubMEth]	Subjective Methods for the Assessment of Stereoscopic Three–Dimensional Television (3DTV) systems	Recommendation covering subjective assessment methods for 3DTV
ITU–T P.3D–sam	Subjective assessment methods for 3D video quality	Recommendation regarding 3D assessment methods for the current 3D environment
ITU–T J.3D–fatigue	Assessment methods of visual fatigue and safety guideline for 3D video	Visual fatigue and safety assessment guideline for 3D video
ITU–T J.3D–disp–req	Display requirements for 3D video quality assessment	Requirements for displays used for 3D assessment testing

ommendation, ITU–R BT.2021: *Subjective methods for the assessment of stereoscopic 3DTV systems*, was published in August 2012 [7]. Compared with ITU–R BT.1438, ITU–R BT.2021 highlights primary perceptual dimensions (picture quality, depth quality, and visual (dis)comfort) as well as additional perceptual dimensions (naturalness, and sense of presence).

3.2 Explorative Studies

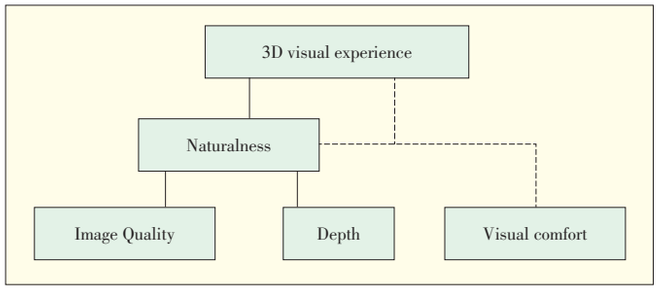
Besides international standardization activities, many explorative studies have been conducted over the past decade to better understand and assess the QoE of stereoscopic images.

In [8], the authors discuss the human factors in 3DTV. Subjective evaluation criteria were proposed to guide the development of 3DTV services. In [9], Wöpking conducted a subjective experiment to assess the annoyance caused by impairments in stereoscopic images. A single–stimulus impairment scale with nine different disparity levels and five levels of background resolution was used. In [10], Ijsselsteijn et al. investigated the effect of camera parameters and display duration on subjective evaluation of stereoscopic images. The authors used single–stimulus methods with a numerical scale from one to ten, where one is the lowest level and ten is the highest level of the attribute. Observers were asked to rate quality of depth and naturalness of stereoscopic images. In [11], Yano et al. used SSCQE with a quality scale to subjectively test visual comfort. Two 15–minute video sequences, (a 2D video and a stereoscopic video) were used as stimuli. In [12] and [13], Meester et al. identified underlying attributes of image quality and quantified the perceived strengths of each attribute. They described how the principles of quantitative quality measurement of 2D image quality can be applied to 3DTV. In [14], Kooi and Toet used the DSIS Variant I method and a five–level scale of discomfort to assess the visual discomfort created by visual asymmetries in stereoscopic images. This scale is: 1) equal viewing comfort; 2) slightly reduced viewing comfort; 3) reduced viewing comfort; 4) considerably reduced viewing comfort; 5) extremely reduced viewing comfort. In [15], Yano et al. used a five–level visual fatigue scale and changed accommodation and convergence to evaluate the viewer's subjective fatigue level after an hour of stereoscopic viewing. The scale in [15] is: 5) I am not tired; 4) I sense a little tired; 3) I am a little tired; 2) I am tired; 1) I am very tired. In [16], Emoto et al. proposed thát the change of fusional amplitude and accommodation response is a valid indicator of visual fatigue. In [17], Seuntiëns et al. used a single–stimulus assessment method with a five–level quality scale to assess the naturalness of viewing 3D images. In [18], the same authors investigate perceptual attributes of crosstalk in 3D images. The same single–stimulus assessment method with five–level scale was used to assess perceived image distortion and perceived visual strain. In [19], the same authors still used the single stimulus method but with different scales to assess the effects of symmetric and asymmetric JPEG coding and camera separation. Perceived overall image quality was rated according to the ITU's five–level quality scale, and the eye strain was rated according to the ITU's five–level impairment scale. Perceived sharpness and depth were rated using a numerical scale from one to five. No adjectives were used on the depth and sharpness scale. In his PhD thesis, Seuntiëns summarized all his studies and proposed a perceptual model for 3D visual experience (Fig. 3) [20].

In [21], a questionnaire on the five main factors for visual fatigue was proposed. In [22], an electroencephalography (EEG) signal was used to detect visual fatigue. In [23], image quality; naturalness, depth perception; and viewing experience for stereoscopic images with different camera baseline distances, blur levels, and noise levels were rated using a single–stimulus method and the ITU quality scale. In [24] and [25], Goldmann et al. established a stereo image and video database. They used a single–stimulus method with continuous quality scale to evaluate the quality of stereoscopic images in the proposed database. In [26], Strohmeier et al. used a method that combined psychoperceptual evaluation (acceptance of quality, overall satisfaction, 3D impression) and qualitative attribute elicitation (perceived overall image quality and perceived depth) to attain a holistic understanding of 3D audiovisual quality in mobile 3D devices. In [27], a paired comparison method and autostereoscopic display was used to understand the affect of depth rendering on QoE. In [28], the authors assessed the quality, depth, and naturalness perceived in the uncompressed and compressed stereoscopic images. They concluded that both perceived quality and perceived depth need to be known in order to assess 3D QoE. Naturalness was found to be highly correlated to quality. Table 4 summarizes all of the previously mentioned studies.

3.3 Discussion

Conventional ITU standards such as ITU–R BT.500 do not cover the new characteristics of S–3DTV. The adapted ITU–R BT.1438 only covers a limited number of S–3DTV characteristics. New questions about subjective assessment for S–3D video have been raised, and new ITU activities on evaluating QoE



▲Figure 3. Model of 3D visual experience.

for 3D video are now underway.

Explorative studies on assessing QoE for S–3DTV have resulted in three main observations:

1) In many studies, different indicators, or subjective attributes, were used to measure QoE of stereoscopic images. These attributes include amount of depth, quality of depth, texture quality and sharpness, visual comfort, visual fatigue, viewing experience (overall image quality or visual experience), naturalness, presence, and enjoyment [29]. There are no common definitions for some QoE indicators; for example, depth may refer to the amount of depth [23] or the quality of depth [10]. Image quality may refer to texture quality [23] or overall image quality [24], [25]. In fact, it is difficult to accurately compare studies; however, a common understanding of S–3DTV QoE assessment can be drawn from explorative studies. Conventional quality indicators are not sufficient to determine QoE for S–3DTV, and multidimensional QoE indicators are required.
2) The subjective test environment was different for each of the subjective experiments. For general viewing conditions, various types and sizes of S–3DTV display were used, often without specification of the calibration process and luminance. The rule of determining viewing distance varied. Occasionally, test materials were not precisely specified. Most of the studies did not follow the recommendations of ITU–R BT.500 and ITU–R BT.1438, perhaps because the general viewing conditions proposed by ITU–R BT.500 are not suitable for 3D applications. This also makes it more difficult to compare studies.
3) There are still no common methods to assess visual fatigue.

In the development of new standardized subjective QoE assessment methods, these three observations must be taken into account. Reliable specifications must be created to guide subjective assessment and achieve reliable, comparable, and repeatable subjective experiential results.

4 Towards Comprehensive Adaptation of Subjective QoE Assessment for S–3DTV

Conventional subjective quality assessment methodologies need to be adapted to S–3DTV. Because S–3DTV QoE is multidimensional, multiple QoE indicators are required. Moreover, when specifying common features for the assessment of S–3DTV images, new factors in S–3DTV need to be considered because they might affect QoE.

In this section, we propose and define multidimensional QoE indicators for S–3DTV. Then, we discuss new factors that need to be considered for comprehensive subjective assessment of S–3DTV QoE. The traditional way of evaluating QoE involves assessing overall visual quality; however, this is not sufficient for determining the advantages and disadvantages of stereoscopic images. Image quality does not encompass perceived depth and visual comfort. One of the common conclusions from the literature presented in the previous section is that S–3DTV QoE should be considered multidimensional. We propose the following QoE indicators to assess S–3DTV QoE:

- 2D image quality. This is the quality of texture rendering without regard to depth.
- depth quantity. This is the amount of perceived depth induced by the combination of monocular and binocular depth cues.
- visual discomfort. This is caused by eye strain, dry eyes, and fusion difficulties. Variation in visual comfort can be also perceived as the sensation of vision difficulties.
- depth rendering. This is the quality of the perceived depth and depends on the observer's preferred basic depth reconstruction criteria. It is mostly related to stretching or compression of the real scene in the reconstructed scene and also affects the shapes of objects.
- naturalness. This is an evaluation of whether the scene more or less represents reality.
- visual experience. This is the overall QoE of the images (in terms of immersion) and the overall perceived quality.

By the definition of the above six QoE indicators, we can separate these indicators into two levels (Fig. 4). The higher–level concept QoE indicators, such as visual experience, naturalness, and depth rendering, can be a complex combination of different cognition and perception decisions. The low-

▼Table 4. Overview of the explorative studies on QoE of S–3DTV

QoE Indicators	Methods	Scales	Studies
Texture quality and sharpness	Single stimulus	ITU–R quality scale with or without adjectives	[19], [20], [23], [28]
Amount of depth	Single stimulus	Numerical scale(0–5)	[10], [19], [23]
Quality of depth	Single stimulus, paired somparison	Numerical scale (0–10)	[26], [27], [28]
Visual comfort, Eye strain and Visual Annoyance	Single stimulus, SSCQE, DSIS	ITU–R impairment and quality scale, adapted impairment scale from ITU–R	[9], [11], [14], [19]
Visual fatigue	Questionnaire, objective measurement (e.g. EEG)		[15], [16], [21]
Viewing experience, overall image quality, visual experience	Single stimulus	ITU–R quality scale	[17]–[20], [23]–[26]
Naturalness	Single stimulus	Numerical scale(0–10), ITU–R quality scale	[10], [17], [18], [20], [23], [28]
Presence and enjoyment	Single stimulus	ITU–R quality scale	[20]

er–level QoE indicators comprise the basic QoE indicators, which may provide a direct link to the technical parameters, such as image quality, depth quantity, and visual comfort.

In our studies [30]–[32], we designed subjective QoE experiments to understand how varying basic QoE indicators affects other quality indicators. The results led to a proposal for modeling higher–level concepts, such as depth rendering, naturalness and visual experience. A 3D *QoE* indicator, denoted *QoE*, may be represented as a weighted sum of 2D image quality (*IQ*), depth quantity (*D*), and visual comfort (*VC*):

$$QoE = (\alpha \cdot IQ) + (\beta \cdot D) + (\gamma \cdot VC) \quad (1)$$

The above indicators are used to determine short–term or instant opinion of the QoE of stereoscopic images. Long–term of viewing of S–3DTV images may induce visual fatigue and affect QoE of S–3DTV. Thus, visual fatigue can be used as a long–term QoE indicator and is defined as a decrease in performance of the visual system. It is an objectively measurable criterion that is particularly valuable for determining long–term adaptive processes of the visual system.

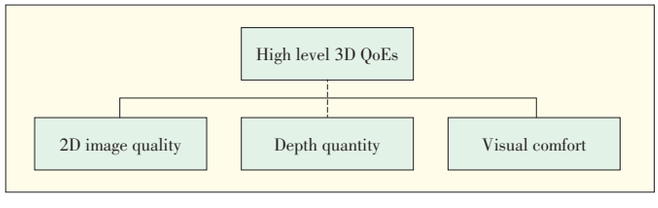
However, methods for measuring visual fatigue are being investigated, and no common methods currently exist.

4.1 New Factors Affecting Assessment of S–3DTV QoE

For subjective quality assessment, environmental setups, as those described in ITU–R BT.500, do not cover the new characteristics of S–3DTV. Thus, conventional methods need to be adapted to accommodate the new factors of S–3DTV. In this section, we discuss new factors that affect S–3DTV QoE assessment based on ITU–R BT.500 recommendation (Table 1).

1) General Viewing Conditions

- luminance and contrast ratio. Additional optical instruments for 3D viewing (e.g. glasses and filters) reduce luminance. We previously found that luminance reduces by up to 70% for 3DTV systems with active glasses and about 50–60% for polarization 3DTV systems [32]. This should be taken into account when measuring peak luminance. In [33], at least 30 cd/m^2 was suggested as the minimum luminance for S–3DTV displays in order to sustain the depth of focus and guarantee basic depth sensation. Moreover, crosstalk is not only an annoying artifact, but it also affects the final contrast ratio. Thus, the display measurement and calibration should be specified.
- background and room illumination. When the display is positioned too close to a wall, objects with uncrossed disparity in the screen may appear to be inside the wall. This may cause conflicts between the depth illusion from S–3DTV and the reality of the room. However, some researchers have also argued that this should not be a problem because people can recognize an S–3DTV display as a visual window. Further research is required to solve this problem. Moreover, room illumination may need to be defined more precisely for different 3DTV techniques. For example, the frequency of neon lighting depends on the local grid frequency. When using S–3DTV with active shutter solutions, interference between refresh frequency of the active shutter and the frequency of the neon light may induce serious flickering and eye stress.
- monitor resolution. Overall display resolution, per view resolution, and stereoscopic resolution should be considered as aspects of the monitor resolution. Spatially multiplexed S–3DTV displays have reduced spatial resolution. Moreover, the physical pixel distribution may not be uniform, and pixels belonging to the same view may not be positioned on a Cartesian grid. Time–multiplex displays have reduced temporal resolution. Temporal asymmetries and temporal luminance distribution problems can also occur. It is still an open question as to how the viewer perceives these changes in resolution. In [34], the depth resolution was assessed, and perceived depth voxels and perceived depth range were defined. In [35], stereoscopic resolution was defined as the number of planes of voxels within a certain depth range (± 100 mm around the display plane).
- viewing distance. Three times the height of the screen for HDTV and six times the height of the screen for SDTV were recommended in ITU standards BT.710 [36] and BT.500. Manufacturers often recommend a designed viewing distance (DVD) that differs from the ITU standards. In some cases, for example, autostereoscopic displays, 3D can only be viewed at the DVD. The preferred viewing distance (PVD) was recommended in BT–500 for 2D viewing in home environments. In [37], a subjective test shows that PVD is a function of different parameters, such as human visual acuity, screen size, picture resolution. In [33], perceived binocular depth is a function of binocular disparity scaling and viewing distance. Changing the viewing distance changes the binocular depth perception. Thus, depth perception should be added as a new component for the PVD function.
- viewing position. 3D geometrical distortions (e.g. shear distortion caused by a sideways movement of the observer [38]) can affect how a viewing position is chosen. Luminance reduces more severely when the observation angle increases. This also applies to motion parallax, which is seen on multiview autostereoscopic displays. The viewing position is limited to certain positions in front of the display. If viewers are not in the right position, left and right view images are not correctly perceived in the left and right eye. Crosstalk or reversal of left and right images may occur.



▲Figure 4. 3D QoE models.

- depth rendering. This is the way in which a display represents the perceived depth based on the input video. Depth rendering has been shown to significantly affect the QoE for autostereoscopic displays [27]. At the display side, depth rendering depends on viewing distance, content disparity, and display properties. Moreover, constraints cause by the comfortable viewing zone should be taken into account for depth rendering.

2) Source Signals

- video format. Various 3D representation formats are available in the literature. These formats include conventional stereo video, 2D–plus–depth, format, multiview video (MVC) and multiview video plus depth format (MVD), layer depth video (LDV), and depth–enhanced stereo (DES). For frame–compatible formats such as top–and–bottom and side–by–side, reducing resolution may affect quality. Our study [32] showed that side–by–side format provides better visual experience than top–and–bottom format for line–interleaved display, especially for interlaced scan content. To optimize 3DTV QoE, interaction with 3D display technique should be taken into account when selecting a 3D representation format. For formats based on depth maps, the quality of the rendered novel views is still not comparable to native stereo views. This even applies to the LDV format [39], [40]. Video format and view synthesis algorithm still need to be specified.
- video format conversion. Conversion between the previously mentioned video formats is lossy in most cases. For example, information for occluded objects is systematically lost if 2D–plus–depth–format with a single layer of depth is converted to conventional stereo video format [39]. The amount of loss depends on the implementation used. Minimum accuracy should be defined for the format conversion by providing a validation test set.

3) Selection of Test Materials

- video content complexity. For 2D video, ITU–T P.910 defines the spatial perceptual information (SI) and the temporal perceptual information (TI) as main elements of 2D video complexity [41]. Some new measurements, called depth perceptual information (DI), should complement these two measurements. With DI, spatial and temporal maximum disparity and average disparity in pixels may be considered. Adding a third dimension to the video content complexity also requires more standardized video sequences; for example, further shooting sessions are required to generate the new reference scenes with various complexity levels that take into account SI, TI, and DI.
- content acquisition and calibration. Stereoscopic distortion, such as puppet theater effect and cardboard effect [42], is an impediment to comfortable viewing and is a key factor that needs to be considered in content acquisition [43]. Moreover, view asymmetry, such as misalignment of camera positions, magnification between views, and desynchronization of color, may be induced by different sources. Because view asymmetries can induce visual artifacts and might result in visual discomfort, calibration of stereoscopic images is important [32].

4) Observers

- number. The number of observers depends on sensitivity and the required reliability of the experiments. In [44], individual differences in susceptibility are still unclear. The viewers´ opinion was reported to be not as stable for 3D as it was for 2D. Thus, an increase in the number of observers might be required to guarantee the reliability of the test. The minimum number of 15 observers recommended in ITU–BT.500 may not be sufficient.
- viewer's stereopsis performance. About 10–15% of the population cannot properly perceive binocular depth cues; therefore, additional optometric tests should be done to evaluate the viewer's binocular vision. ITU–R BT.1438 recommends different vision tests (VTs) for assessing binocular vision.

5) Test Session

- viewing duration. The reference in ITU–R BT.500 for short–duration 2D video samples is 10 s. For the transition to 3D, there are two conflicting viewpoints. One viewpoint is that because S–3DTV more closely resembles natural human viewing behavior, less time is needed to judge the quality. The other viewpoint is that more time is needed because more information is contained in the additional dimension of S–3DTV, and the viewer is used to 2D displays. For a short duration test, the presentation time had little effect on subjective evaluation results; however, only 5 s and 10 s were tested [10]. Further studies are required on viewing duration in subjective tests.

6) Analysis of Test Results

- viewer factor. A statistical analysis needs to be done in order to reject an incoherent viewer. For S–3DTV, subjective test results may be more sensitive to individual preferences; therefore, multimodal viewer distributions might need to be analyzed.
- multidimension indicator analysis. Using indicators such as QoE, depth sensation, and visual comfort for 3D requires new methods summarization and statistical analysis methods, and test results need to be carefully interpreted. of objective models for 3D video quality.

7) Test Methods

- visual fatigue. This is an objectively measurable quantity. Several approaches to assess visual fatigue have been investigated. Such approaches include optometric tests of visual function, electroencephalography (EEG) and event–related potential (ERP) [22], eye tracking considering visual interest, snapshots of visual discomfort (in the form of questionnaires before and after viewing [21]), and continuous assessment of comfort [11]. These efforts may lead to standardized procedures and recommendations.

- subjective QoE indicator. Multidimensional QoE indicators should be used to assess QoE of S–3DTV. Particular indicators should be used to assess particular problems in S–3DTV. Moreover, interactions between different QoE indicators should be well specified.

New factors affecting the subjective assessment of S–3DTV are summarized in Table 5. Further experiments are needed on most of these new factors.

5 Conclusion

In this paper, we have reviewed the current status of QoE assessment and have drawn several observations. First, conventional subjective quality assessment methods are not sufficient for evaluating the quality of stereoscopic images. ITU and VQEG are currently working on new subjective quality assessment methods for such contexts. Apart from standardization efforts, several explorative studies have been done on different topics considering very different QoE indicators. However, there are no common definitions for these QoE indicators. Moreover, the viewing environment and conditions vary between studies, and this makes it more difficult to draw comparisons. We investigated multidimensional QoE indicators, including 2D image quality, depth quantity, visual comfort, depth rendering, naturalness and visual experience, and visual fatigue. We discussed comprehensive adaptations of subjective QoE assessment for S–3DTV. New factors in 3D need to be considered when developing QoE assessment methods for S–3DTV. Such factors will help define new subjective QoE assessment methodologies for 3DTV stereoscopic images. These methods have already been successfully applied on the production side. Orange has developed a capture–monitoring system currently used by stereographers and post producers. This tool is successful mostly because it is based on carefully designed QoE experiments that follow the proposed framework in this paper. Nevertheless, this framework still needs to be challenged through use in other parts of the delivery chain.

▼Table 5. New factors affecting subjective assessment for S–3DTV

Feature	Factors	New factors
General viewing conditions	Luminance and contrast ratio	Luminance reduction caused by additional optical instrument, minimum luminance necessary to sustain DOF, contrast ratio affected by crosstalk
	Background and room illumination	Minimum distance between display and background necessary, technology of room illumination critical
	Monitor resolution	Recommendation of minimum values for spatial and temporal per view resolution and stereoscopic resolution
	Viewing distance	Designed viewing distance (DVD) fixed by display manufacturer and adding depth perception factor into preferred viewing distance (PVD)
	Viewing position	Avoidance of 3D geometrical distortion, luminance reduction, suboptimal viewing position for autostereoscopic displays
	Depth rendering	Upper bounds for Depth Of Focus and binocular disparity
Source signals	Video format	Requirements for depth representation formats
	Video format conversion	Specification of accuracy for conversion
Selection of test materials	Video content complexity	Measurement tools for depth complexity of content
	Content acquisition and calibration	Consider stereoscopic distortion and constrain of visual comfort in content acquisition, calibration of stereoscopic images to avoid view asymmetries
Observers	Number	Re–evaluation necessary to guarantee stability and reliability of results
	Viewer's stereopsis performance	Measurement of stereopsis, accuracy, ocular differences etc.
The test session	Viewing duration	Re–evaluation of duration for presentation, voting, session length
Test Results analysis	Viewer factors	Rejection criteria, detection of bimodal distributions
	Multidimension indicators analysis	Statistical methods for analysis, e.g. relation, interaction and combination of subjectively measured QoE indicators
Test method	Visual fatigue	Objective measurement of visual fatigue
	Subjective QoE indicator	Multidimensional QoE indicators

References

[1] ITU, "Methodology for the subjective assessment of video quality in multimedia applications," in Recommendation ITU–R BT.1788, International Telecommunication Union., 2007.
[2] ITU, "Methodology for the subjective assessment of the quality of television pictures," in Recommendation ITU–R BT 500–11, International Telecommunication Union., 2002.
[3] J. L. Blin, "New quality evaluation method suited to multimedia context SAMVIQ," in *The Second International Workshop on Video Processing and Quality Metrics for Consumer Electronic*, Phoenix, Arizona, 2006.
[4] F. Kozamernik, et al., "Subjective quality of Internet video codecs – Phase 2 evaluation using SAMIVQ," 2005.
[5] ITU, "Subjective Assessment of Stereoscope Television Pictures," in RECOMMENDATION ITU–R BT.1438, International Telecommunication Union., 2000.
[6] NTT, "3D quality assessment methods," NTT2011.
[7] ITU, "Subjective methods for the assessment of stereoscopic 3DTV systems," in Recommendation ITU–R BT. 2021, International Telecommunication Union., 2012.
[8] S. Pastoor, "Human factors of 3DTV: an overview of current research at Heinrich–Hertz–Institute Berlin," in *Stereoscopic Television, IEE Colloquium on*, 1992, pp. 11/1–11/4.
[9] M. Wöpking, "Viewing comfort with stereoscopic pictures: An experimental study on the subjective effects of disparity magnitude and depth of focus," *Journal of the Society for Information Display*, vol. 3, p. 3, 1992.
[10] W. A. IJsselsteijn, et al., "Subjective evaluation of stereoscopic images: effects of camera parameters and display duration," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 10, pp. 225–233, 2000.
[11] S. Yano, et al., "A study of visual fatigue and visual comfort for 3D HDTV/HDTV images," *Displays*, vol. 23, pp. 191–201, 2002.
[12] L. Meesters, et al., "A survey of perceptual quality issues in three–dimensional television systems," in *Stereoscopic Displays and Virtual Reality Systems X*, Santa Clara, CA, USA, 2003, pp. 313–326.
[13] L. M. J. Meesters, et al., "A survey of perceptual evaluations and requirements of three–dimensional TV," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 14, pp. 381–391, 2004.
[14] F. L. Kooi and A. Toet, "Visual comfort of binocular and 3D displays," *Displays*, vol. 25, pp. 99–108, 2004.
[15] S. Yano, et al., "Two factors in visual fatigue caused by stereoscopic HDTV images," *Displays*, vol. 25, pp. 141–150, 2004.
[16] M. Emoto, et al., "Changes in fusional vergence limit and its hysteresis after viewing stereoscopic TV," *Displays*, vol. 25, pp. 67–76, 2004.
[17] P. J. Seuntiens, et al., "Viewing experience and naturalness of 3D images," in *Three–Dimensional TV, Video, and Display IV*, Boston, MA, USA, 2005, pp. 601605–7.
[18] P. J. H. Seuntiëns, et al., "Perceptual attributes of crosstalk in 3D images," *Displays*, vol. 26, pp. 177–183, 2005.
[19] P. Seuntiëns, et al., "Perceived quality of compressed stereoscopic images: Effects of symmetric and asymmetric JPEG coding and camera separation," *ACM Trans. Appl. Percept.*, vol. 3, pp. 95–109, April 2006.

[20] P. Seuntiëns, "Visual experience of 3D TV," doctor doctoral thesis, Eindhoven University of Technology, 2006.
[21] O. L. Hyung-Chul, et al., "Method of Measuring Subjective 3-D Visual Fatigue: A Five-Factor Model," 2008, p. DWA5.
[22] H. C. O. Li, et al., "Measurement of 3D Visual Fatigue Using Event-Related Potential (ERP): 3D Oddball Paradigm," in *3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video*, 2008, 2008, pp. 213-216.
[23] M. Lambooij, et al., "Evaluation of Stereoscopic Images: Beyond 2D Quality," *IEEE Trans. Broadcasting*, vol. 57, pp. 432-444, 2011.
[24] L. Goldmann, et al., "Impact of acquisition distortions on the quality of stereoscopic images," in *Fifth International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM 2010)*, Scottsdale, Arizona, U.S.A., 2010.
[25] L. Goldmann, et al., "A comprehensive database and subjective evaluation methodology for quality of experience in stereoscopic video," San Jose, California, USA, 2010, pp. 75260S-11.
[26] D. Strohmeier, et al., "NEW, LIVELY, AND EXCITING OR JUST ARTIFICIAL, STRAINING, AND DISTRACTING : A Sensory profiling approach to understand mobile 3D audiovisual quality," presented at the *Fifth International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM 2010)*, Scottsdale, Arizona, U.S.A., 2010.
[27] M. Barkowsky, et al., "Influence of depth rendering on the quality of experience for an autostereoscopic display," presented at the *First International Workshop on Quality of Multimedia Experience*, San Diego, California, USA, 2009.
[28] K. Yamagishi, et al., "Subjective characteristics for stereoscopic high definition video," in *3rd Int. Workshop on Quality of Multimedia Experience (QoMEX 2011)*, pp. 37-42.
[29] W. A. IJsselsteijn, et al., "State-of-the-art in human factors and quality issues of stereoscopic broadcast television," Eindhoven University of Technology Department Technology Management ATTEST/WP5/01 - Advanced Three-dimensional Television System Technologies, August 2002.
[30] W. Chen, et al., "Quality of experience model for 3DTV," in *Stereoscopic Displays and Applications XXIII*, San Francisco, 2012.
[31] W. Chen, et al., "Exploration of Quality of Experience of Stereoscopic Images: Binocular Depth," in *Sixth International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM 2012)*, Scottsdale, Arizona, U.S.A..
[32] W. CHEN, "Multidimensional characterization of quality of experience of stereoscopic 3D TV," Doctoral dissertation, University of Nantes, Nantes, France, 2012.
[33] R. Patterson, "Human factors of 3-D displays," *Journal of the Society for Information Display*, vol. 15, p. 10, 2007.
[34] L. F. Hodges and E. T. Davis, "Geometric Considerations for Stereoscopic Virtual Environments," 1993.
[35] N. Holliman, 3D Display Systems: IOP Press, 2004.
[36] ITU, "SUBJECTIVE ASSESSMENT METHODS FOR IMAGE QUALITY IN HIGH-DEFINITION TELEVISION," in Recommendation ITU-R BT.710-4, International Telecommunication Union., 1998.
[37] M. Ardito, et al., "Influence of display parameters on perceived HDTV quality," *Consumer Electronics, IEEE Transactions on*, vol. 42, pp. 145-155, 1996.
[38] A. J. Woods, et al., "Image distortions in stereoscopic video systems," in *Stereoscopic Displays and Applications IV*, San Jose, CA, USA, 1993, pp. 36-48.
[39] P. Kauff, et al., "3D4YOU Deliverable D2.1.2: Requirement on post-production and formats conversion," Philips et al. , 07/25 2008.
[40] P. Kerbiriou, et al., "Comparative study and recommendations," 3D4YOU2010.
[41] ITU, "Subjective video quality assessment methods for multimedia applications," in Recommendation ITU-T P.910, ed: International Telecommunication Union., 1999.
[42] H. Yamanoue, et al., "Geometrical analysis of puppet-theater and cardboard effects in stereoscopic HDTV images," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 16, pp. 744-752, 2006.
[43] W. Chen, et al., "New requirements of subjective video quality assessment methodologies for 3DTV," in *Fifth International Workshop on Video Processing and Quality Metrics for Consumer Electronics - VPQM 2010*, Scottsdale, Arizona, U.S.A., 2010.
[44] K. Ukai and P. A. Howarth, "Visual fatigue caused by viewing stereoscopic motion images: Background, theories, and observations," *Displays*, vol. 29, pp. 106-116, 2008.

Manuscript received: January 23, 2013

Biographies

Wei Chen

Wei Chen recevied his PhD degree in signal and image processing from the University of Nantes in 2012. His thesis focused on characterizing the quality of exprience for stereoscopic 3D TV. This also led to his participation on the European projects 3D4YOU and 3Dlive and on the proposal for an ITU recommendation on subjective qualtiy assessment methodologies for 3D TV. From 2007 to 2009, he was a research engineer in the IVC team, IRCCyN Labs. In 2009, he joined France Telecom Research and Development, where he was a research engineer involved in the development of new subjective quality-assessment methodologies for 3D TV. In 2012, He joined Microsoft Skype Division as a video quality expert for real time video communication. His current research interests include subjective and objective quality assessment, video coding, and human visual system.

Jérôme Fournier

Jérôme Fournier received his PhD degree in signal and image processing from the University of Rennes in 1995. His thesis was on the subjective evaluation of stereoscopic television. He participated in the European project RACE DISTIMA. After that, he worked on video compression for video communications at LEP (Philips Research Laboratory in France) and participated in the standardization of H.263. In 1997, he joined France Telecom R&D and worked on the subjective evaluation of video sequences and on the implementation of standardized video codecs such as MPEG-4 Part 2 and H.264. From 2004 to 2006, he was in charge of HDTV activities at France Telecom. Since 2006, Jérôme has been working on the deployment of the Orange stereoscopic 3D TV. He has also been working on the assessment of innovative 3D TV depth-based video formats as part of the European project, 3D4YOU. He is currently involved in the subjective evaluation of ultra-HD video formats recently standardized at ITU-R as well as the new video compression technology called HEVC.

Marcus Barkowsky

Marcus Barkowsky received his Dipl.-Ing. degree in electrical engineering in 1999. He received his Dr.-Ing. degree in 2009 from the University of Erlangen-Nuremberg, Germany. In September 2010, he was an associate professor in the Department of Informatics, Polytechnical Faculty, University of Nantes, France. He also worked in the IRCCyN research lab. His PhD thesis focused on subjective and objective video quality assessment. He designed a reliable video quality measure for low bitrate scenarios with special emphasis on mobile transmission. Since November 2008, he has worked at the University of Nantes researching the influence of 3D TV on the human visual system, with a special emphasis on visual fatigue and human factors. He has established computational models for the influence of quality degradation on human perception in a spatiotemporal concept in 2D and 3D display conditions. He co-chairs the activities of the Video Quality Experts Group in 3D TV and Hybrid Video Quality Measurement towards the creation and modification of ITU recommendations.

Patrick le Callet

Patrick le Callet received his MSc degree and PhD degree in image processing from Ecole Polytechnique de l'université de Nantes. He was also student at the Ecole Normale Superieure de Cachan where he got the "Aggrégation" (credentialing exam) in electronics of the French National Education. He worked as an assistant professor from 1997 to 1999 and as a full time lecturer from 1999 to 2003 in the Department of Electrical Engineering, Technical Institute of the University of Nantes (IUT). Since 2003, he has been a full professor at Ecole Polytechnique de l'université de Nantes in the Department of Electrical Engineering and Computer Science. Since 2006, he has been the head of the Image and Video Communication Lab at CNRS IRCCyN, a group of more than 35 researchers. He mostly researches human vision modeling in image and video processing. His current research interests include 3D image and video quality assessment, watermarking techniques, and visual attention modeling and applications. He has co-authored more than 150 publications is the co-holder of 13 international patents on these topics. He has coordinated and is currently managing several National and European collaborative research programs representing grants of more than 3 million euros. I then Video Quality Expert Group, he is co-chairing the High Dynamic Range and 3DTV projects. He has been a member of the technical committee of several conferences, and he has previously reviewed journals such as *Signal Processing: Image Communications*, *IEEE Transactions on Broadcasting*, *IEEE transactions on Image Processing*, and *Journal of Electronic Imaging*. He is currently an associate editor of *IEEE Transactions on Circuit System and Video Technology*, *SPIE Journal of Electronic Imaging*, and *SPRINGER EURASIP Journal on Image and Video Processing*.

3D Perception Algorithms: Towards Perceptually Driven Compression of 3D Video

Ruimin Hu, Rui Zhong, Zhongyuan Wang, and Zhen Han
(National Engineering Research Center for Multimedia Software, School of Computer, Wuhan University, Wuhan 430072, China)

Abstract

In this paper, we summarize 3D perception–oriented algorithms for perceptually driven 3D video coding. Several perceptual effects have been exploited for 2D video viewing; however, this is not yet the case for 3D video viewing. 3D video requires depth perception, which implies binocular effects such as con fl icts, fusion, and rivalry. A better understanding of these effects is necessary for 3D perceptual compression, which provides users with a more comfortable visual experience for video that is delivered over a channel with limited bandwidth. We present state–of–the–art of 3D visual attention models, 3D just–noticeable difference models, and 3D texture–synthesis models that address 3D human vision issues in 3D video coding and transmission.

Keywords

3D perception; 3D visual attention; 3D just–noticeable difference; 3D texture–synthesis; 3D video compression

1 Introduction

3D TV provides an immersive visual experience, and the development of 3D TV technologies has hastened. New video formats such as multiview and multiview plus depth (MVD) were designed for 3D perception [1]. 3D introduces new requirements, such as disparity adaptation between different display screens, 2D to 3D conversion, 3D error concealment, and 3D rendering. All of these require 3D video perceptual processing algorithms [2]. The huge amount of 3D video data also has created challenges in compression and storage. Many proposed 3D video compression algorithms exploit the statistic redundancy of the 3D video; however, coding performance is improved by increasing the computational complexity, which eventually creates a bottleneck. Because human eyes are the final receivers of a stereoscopic scene, human perception plays a part in designing high–efficiency coding algorithms for 3D video.

Integrating human visual perception into the general 2D video coding framework is an open issue [3]. In [4], structure similarity and content saliency information was incorporated into the distortion metric, and 10.14% bit rate was saved with similar subjective perception. In [5], details of many perception–based coding methods are discussed. 2D perception–based coding algorithms are mature; however, 3D perception–based coding algorithms are in still in their infancy. In [6], a novel depth coding method was proposed. The authors took into consideration the fact that distortion around object edges leads to serious artifacts. In [7], the authors highlighted the importance of incorporating the quality of synthesized color video into the distortion metric when encoding the depth video. The 3D perception model should integrate the specific visual perception difference between 2D and 3D. In this paper, we analyze the features leading to perception difference for 3D and review existing works in which 3D perception is integrated into the coding framework.

The main difference between 3D and 2D perception is depth perception; stereoscopic vision is created via binocular cues such as conflicts, fusion, and rivalry [8]. Binocular conflict arises from inherently ambiguous sensory signals. Although 3D perception still exists when binocular conflict occurs, visual switching between left eye and right eye is uncomfortable [2]. Visual attention models can relieve discomfort by reducing the conflict in salient regions. Two images shot from different angles are displayed for each eye. When two monocular images have different luminance or contrast but share a common polarity, the fused average of the two monocular components leads to binocular fusion [8]. Binocular rivalry occurs when dissimilar monocular stimuli are presented to the corresponding retinal locations of the two eyes [9]. Because these binocular cues exist, the visual attention regions are shifted, and the just–noticeable–difference values in 3D video are different to those for 2D video human perception. For a comfortable 3D TV experience, 3D perception coding algorithms attempt to show those binocular effects accurately. Bandwidth is another critical factor that affects 3D TV experience. Depth–image–based rendering (DIBR) has been proposed to synthesize the virtual videos of different perspectives and save bit rate in 3D video encoding. However, there may be holes in the synthesized video because the occluded regions in the original view become visible in the synthesized view [10]. Approaches based on texture synthesis and texture masking are taken to recover holes in the synthesized video [11].

In section 2, we give an overview of state–of–the–art 3D per-

ception models and briefly discuss the usefulness of these models briefly. In section 3, we describe some 3D visual perception algorithms that perform well. In section 4, we analyze and compare the previously discussed 3D perception models. Section 5 concludes the paper.

2 3D Perception Algorithms

Achieving high-quality 3D TV is a hot research topic. As well as bandwidth and processing steps, depth perception also affects human 3D visual experience. 3D perception models are used to address human visual issues such as disparity adaptation between different display screens, 2D to 3D conversion, 3D error concealment, and 3D rendering [2]. Existing 3D perception models explain binocular effects from the angle of subjective experimentation and modeling. Here, we describe the current status of the 3D perception models.

2.1 Just-Noticeable Difference Models for 3D Video

Just-noticeable difference (JND) models for 3D images have recently been proposed to accurately estimate redundancy in visual perception. A depth JND model proposed [12] demonstrated why human beings are not sensitive to varied depth values. With the development of 3D image processing technologies, the depth JND model, which only measures the depth perception difference, is not sufficient. A 3D image JND model for describing the total stereoscopic perception is necessary. In [13], a binocular JND (BJND) model was proposed to describe the basic binocular vision properties of asymmetric noises in paired stereoscopic images. This was the first binocular JND model in which luminance adaption and contrast masking were taken into account. The model was verified in a formal psychophysical experiment, and the results showed that the JND values could be accurately obtained using the model. However, the model was constructed on the assumption that the disparity was zero; therefore, the model was not suitable for normal binocular stereo images with nonzero disparity. In [14], a joint JND (JJND) model was proposed to separately measure the sensitivity difference of occlusion and non-occlusion regions, taking into account the fact that occlusion regions at the object edges are more visually sensitive. This model addressed the problem caused by ignoring disparity, and more accurate JND values were assigned for human visual perception. However, JND values are affected by differing human visual sensitivity to different stimuli [15]. Using depth intensity as the only influencing factor does not result in precise visual sensitivity. Depth intensity and depth contrast, both of which significantly affect human visual perception, need to be explored when building a 3D image JND model.

2.2 3D Visual Attention Models

Region-of-interest algorithms can guide bit rate allocation during 3D video coding. Depth perception plays an important role in 3D video viewing, and this probably affects the location of the region of interest. In [16], the saliency region was determined using the scene depth derived from 2D saliency algorithms. According to the center-surround mechanism, a saliency map was created by extracting low-level features from the images. The depth map was treated as another low-level feature and was linearly integrated into the overall saliency model. However, the model was not validated by standard subjective experiments, and did not refer to the binocular effect. In [17], binocular rivalry in 3D perception is discussed. Directly adapting 2D saliency algorithms for use in 3D video introduces new problems; therefore, a region-of-interest map based on a hierarchical model can be generated from basic and special features [17]. Although the model gives the displacement of the region of interest based on binocular effects, the response of each eye is treated independently. A perceptual model for disparity is given in [18]; however, it is more accurate to calculate visual saliency based on depth perception. Wang constructed a model for quantifying depth bias for free viewing of still stereoscopic video [19]. In [20], a bottom-up visual saliency model was proposed for 3D video. The 3D visual hierarchical model was extended by treating the depth map as an extra clue. Depth was incorporated into the saliency map that was built by integrating color, orientation, and motion contrast features.

However, in this model, binocular rivalry, binocular combination, or binocular conflict were not taken into account. The author demonstrated the model by using eye tracking to analyze stereoscopic filmmaking [21]. The eye tracking mechanism allows the model to be compared with ground-truth results from 3D visual saliency models. In [22], a saliency model was created by solving the temporal coherence problem in 3D visual perception. However, in this paper, we focus on spatial consistency in 3D saliency models.

2.3 Texture-Synthesis Models

MVD or multiview video (MVV) generates a greater amount of data for transmission and storage compared with conventional 2D video. To address this problem, 3D video coding needs to have high compression efficiency. However, the difference between the 3D and 2D video features makes it difficult to use 2D encoding algorithms for 3D video. In general, it is advantageous to use the depth video to assist in the coding of the color video.

There are structural similarities between the depth image and color image, which means that objects at the same location in the images share the same motion information. In [23], a method of sharing motion information between the depth video and texture video was proposed. The motion vector of the texture video was split and recombined for motion compensation in the depth video. However, the method results in only slightly better coding performance in low-bitrate scenarios. In [24], view synthesis prediction was proposed for multiview video

coding, and rate distortion was optimized to guide the coding process. This optimization was based on view synthesis prediction and was shown to improve coding gain. Depth image–based rendering was done to synthesize the virtual videos of different perspectives and to reduce the coding bit rate. Encoding bit rate is reduced by increasing decoding complexity. Occluded regions in the original videos are properly displayed in the virtual videos. In [25], a novel non–parametric texture–synthesis–based approach was proposed to fill the holes in the synthesized video. The method takes into account the statistical dependencies of a sequence by a background sprite, and unknown regions are recovered using the image content.

3 Details of Some Algorithms that Perform Well

Here, we describe state–of–the–art models for 3D perception. 3D visual attention models can be integrated into the encoding framework as the guide for bit–rate assignment. 3D JND models are frequently used to filter encoding distortion. Models based on texture synthesis are proposed to fill the holes in synthesized video that arise a result of texture masking. The 3D–perception coding algorithms accurately describe binocular effects.

3.1 Just–Noticeable Difference Algorithms

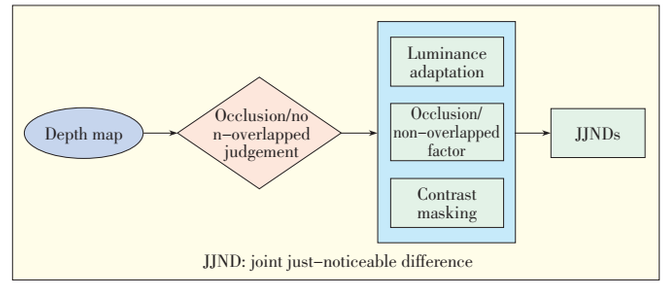
In the conventional 2D JND model in [26], luminance adaptation and contrast masking are non–linearly summed by weight to obtain the JNDs. The luminance adaptation describes the visibility threshold in terms of background luminance (Weber's law) [27]. Contrast masking arises because the visibility of a spatial object can be reduced in the presence of a neighboring object:

$$JND_{2d}(x, y) = LA(x, y) + CM(x, y) - C^{LC}(x, y) \cdot \min\{LA(x, y), CM(x, y)\} \quad (1)$$

where $JND_{2d}(x,y)$ is the 2D image JND; $LA(x,y)$ and $CM(x,y)$ are the visibility thresholds for luminance adaptation and contrast masking, respectively; and $C^{LC}(x,y)$ is the effect of overlapping of two factors for $0 < C^{LC}(x,y) \leqslant 1$. The factors of the 2D JND could also work in the 3D–image JND. A joint JND (JJND) model was built on the assumption that the occlusion introduces stronger depth perception and leads to smaller JNDs. Therefore, the JNDs were calculated by dividing the image into occlusion and non–overlapped regions [14]. The 3D JND model is shown in Fig. 1 [9] and is given by

$$JJND(x,y) = \begin{cases} JND_{2d}(x, y) \cdot \alpha & W_{\text{occlusion}} = 1 \\ JND_{2d} \cdot \beta(x, y) & \text{otherwise} \end{cases} \quad (2)$$

where α is set to 0.8, and $\beta(x, y)$ derives from the depth of a pixel [14]. When the pixel belongs to an occluded area, $W_{\text{occlusion}} = 1$. The method used to judge whether a pixel is in an occluded area is described in [28].

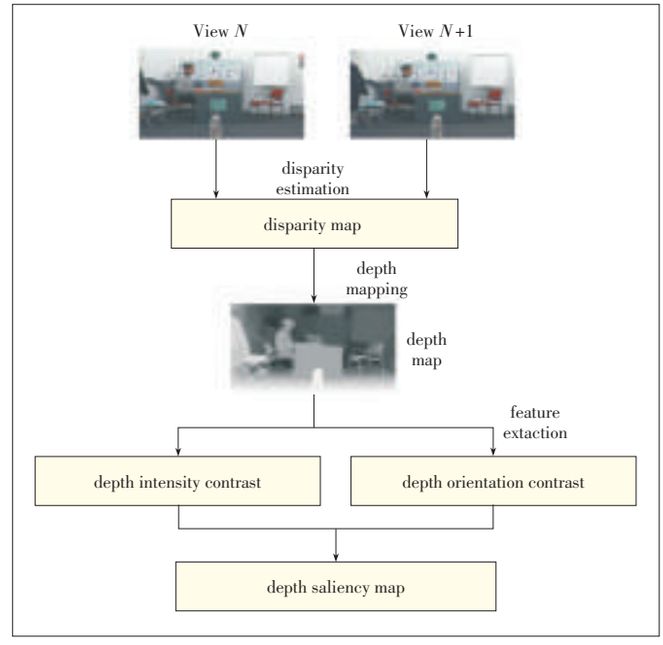


▲Figure 1. JJND model.

3.2 Building the Depth Saliency Model

The visual attention values of video content are calculated by simulating the human visual perception mechanism in traditional 2D video saliency models. The prominent difference between 2D and 3D imaging is depth perception. Depth affects saliency by making the pop–out areas of a 3D image more attractive to the human eye than concave regions and by making areas with inconsecutive depth or higher depth contrast more attractive to the human eye. All of these areas are more visually stimulating [20]. The depth saliency model in [20] is used to obtain the depth saliency map, which is the same size as the original image, and each pixel of the map corresponds to each depth attention value. First, we calculate the depth from the horizontal disparity map. Then, the depth intensity and depth contrast are weighed to obtain the final depth saliency map (Fig. 2).

In the first step, a stereo–matching algorithm based on color segmentation is used to calculate the disparity map that repre-



▲Figure 2. Generation of the depth–saliency map.

sents the relative depth between the two views. The vertical disparity is assumed to be zero. The next step is to translate the disparity map into a depth map:

$$Z = \frac{B \cdot F}{disp} \text{ for } disp \neq 0 \quad (3)$$

where F is the focal length of the camera, B is the baseline distance between adjacent cameras, $disp$ is the disparity of the corresponding object in the neighbor view video, and Z is the depth value of the distance between the object and camera in the scene.

The depth saliency can be calculated using (3). The intersection of the two cameras creates a zero-disparity plane that is the default screen for 3D TV. The pop-out objects correspond to negative disparity, and the concave objects correspond to positive disparity. Depth is inversely proportional to disparity. Then, depth is quantized as an 8-bit value, where 0 is the farthest object and 255 is the nearest object. The degree of saliency decreases monotonically with the distance of the objects; the nearer the objects, the more sensitive the human visual perception. Therefore, the depth is mapped into the range between a minimum and a maximum value through non-linear quantization [29]:

$$V = \left\lfloor 255 \cdot \frac{z^n}{z} \cdot \frac{z^f - z}{z^f - z^n} + 0.5 \right\rfloor \quad (4)$$

where $\lfloor a \rfloor$ is the integer less than or equal to a; z^f and z^n are the farthest and nearest depth values, respectively; $z^f = Bf/\min\{disp\}$; and $z^n = Bf/\max\{disp\}$. The non-linear mapping space of depth value is given by $v(x,y)$. Recent research suggests that the pop-out part of a stereo image and the part with inconsecutive depth or higher depth contrast are more attractive to the human eye. The depth contrast map is determined by the absolute center-surround difference (CSD) between different depth intensity channels [30]:

$$F_D = N\left(\oplus_{c=2}^{4} \oplus_{s=c+3}^{c+4} N(|v(c) \ominus v(s)|)\right) \quad (5)$$

where $\ominus$ is the cross-scale difference between two maps, and $\oplus$ is the cross-scale addition in which each map is reduced to a scale of four and point-by-point addition is performed [30]. The final feature map is created by fusing the depth contrast and orientation contrast. The orientation feature is obtained from the depth intensity through oriented Gabor filters [31] and is given by $O(\sigma,\theta)$, where $\sigma \in [0 \cdots 8]$ is the image scale at the different pyramid levels, and $\theta \in \{0, \pi/4, \pi/2, 3\pi/4\}$is the orientation. The depth orientation map F_0 is obtained from the absolute center-surround difference (CSD) between different depth orientation channels:

$$F_O = \frac{1}{4} \sum_{\theta \in \{0, \pi/4, \pi/2, 3\pi/4\}} N\left(\oplus_{c=2}^{4} \oplus_{s=c+3}^{c+4} N(|O(c,\theta) \ominus O(s,\theta)|)\right) \quad (6)$$

where $N(.)$ normalizes the values into a fixed range. The depth contrast and depth orientation maps are then summed with weights to create the final saliency map:

$$S_D = \frac{1}{2}\left(N(F_O) + N(F_D)\right) \quad (7)$$

This model was built with on assumption that the pop-out regions of the 3D image attract more attention than the concave regions, and the area with inconsecutive depth or higher depth contrast is more attractive to the human eye. In [31], the model exploits the special visual characteristics of a 3D image; however, the results are not compared with the ground-truth.

3.3 Building a Texture-Synthesis Model

For additional viewing perspectives, depth image-based rendering (DIBR) is proposed to synthesize the virtual video from the original color image and corresponding depth image. Unknown areas in the original images become visible in the virtual images. The texture-synthesis model is used to recover the unknown areas in [11], holes are classified according to size. Small holes are reconstructed via Laplace cloning, which is 10 times faster than texture synthesis. Holes larger than 50 samples are filled using patch-based texture synthesis in which the statistical properties of pixels of the known neighboring areas are calculated. Content in unknown areas is derived from known areas, and the filling position is deduced by a priority term. Two aspects of the algorithm in [32] are improved in [11]. The gradient is also obtained for initialized content, and filling occurs from background areas to foreground areas Finally, the best-matched patch for the current hole is sourced from neighboring patches by minimizing the cost function:

$$E = \sum_{i=1}^{K} \|x_i - z_i\|^2 + w_\alpha \sum_{j=1}^{K_\alpha} \|x_j - z_j\|^2 \quad (8)$$

where E is the cost energy, x_i is a patch in known regions, and z_i is a patch in the hole area. The number of patches belonging to known areas is K, and the number of patches belonging to known holes is K_α. The weight factor of the patches in holes is w_α. Post-processing is incorporated into the texture-synthesis framework to make the patch transition smooth. Texture synthesis allows for ameliorative virtual video based on texture masking. However, patch-based texture-synthesis algorithms have higher computational complexity.

4 Performance Analysis and Comparison

The JND model tolerates more additional noise without sacrificing subjective image quality [33]. Therefore, to evaluate the JND model accurately, the objective and subjective quality of the noise-injected image needs to be measured. Objective PSNR is used to calculate the amount of noise added to the images. In [12], JND was experimentally measured, and it was determined that a depth value change of 7% can result in noticeable difference. In [13], a JND model for 3D images is created

by experiment. The model and experimental results are helpful for theoretical study; however, the view condition constraint limits its applicability.

Compared to the 2D JND model in [26], the 3D JND model in [14] calculates visual perception more accurately. The PSNR of the images processed by the model in [14] is, on average, 1.03 dB lower than that of the images processed by the model in [26] when MOS scores are similar. This means that more noise can be added to the images guided by the model in [14]. Therefore, for 3D images, the model in [14] could explore more vision redundancies while keeping the 3D images at a similar subjective performance level.

In [17], the quality of the proposed model is not quantitatively evaluated. However, saliency was shown to be accurate for several images for simple geometric objects [17]. The problem with this model is that it is designed for one eye only and might not be suitable for binocular perception. The 3D saliency models are evaluated according to the efficiency of their bit rate allocation. The model in [13] can save more than 21.06–34.29% bit rate, which corresponds to 0.46–0.61 dB ROI PSNR gain with similar subjective video quality with as JMVM 7.0 [33]. The drawback of this method is that binocular effects are not taken into account. In [19], depth–bias feature is demonstrated using an eye–tracking experiment. Binocular effects are exploited while visual saliency is modeled and the saliency feature is described accurately.

In [34], a method is proposed in which motion information is shared between depth and color images. This reduces encoding complexity to 60% that of existing algorithms, and 1 dB PSNR gain against encoding two sequences separately at low bit rates. In [24], a rate–distortion optimization algorithm was created for multiview video coding. The algorithm achieved 0.3–0.8 dB PSNR gain at low to medium bit rates. In [25], the motion vector was predicted using view synthesis prediction. This approach saves 3.86–9.32% more bit rate than MVC. However, the models previously mentioned cannot guarantee high coding efficiency at high bit rates.

5 Conclusion

Unlike 2D video, 3D video has depth perception, which necessitates the development of 3D visual perception algorithms. New requirements include disparity adaptation between different display screens, 2D to 3D conversion, 3D error concealment, and 3D rendering. Considering that the final receivers of a stereoscopic scene are the human eyes, 3D perceptual models have been exploited so that visually comfortable 3D video can be transmitted over a channel of limited bandwidth. In this paper, we focus on state–of–the–art of 3D perception algorithms and analyze their potential application in 3D video coding. Experimental results show that higher coding efficiency and more satisfying 3D TV experience can be achieved by properly integrating 3D perceptual models. However, extending 3D perception algorithms and effectively incorporating 3D perceptual models into video compression requires further exploration.

References

[1] Philipp Merkle, Yannick Morvan and Aljoscha Smolic, "The effects of multiview depth video compression on multiview rendering", *Signal Processing: Image Communication*, vol.24,issues 1–2,pp.73–88,January 2009.
[2] Quan Huynh–Thu, M. Barkowsky and P. Le Callet, "The importance of visual attention in improving the 3D–TV viewing experience: overview and new perspectives", *IEEE Transactions on Broadcasting*, vol.57, no. 2, pp. 421–431, June 2011.
[3] Ndjiki–Nya, P., D. Doshkov, H. Kaprykowsky, F. Zhang, D. Bull, and T. Wiegand. "Perception–oriented Video Coding Based on Image Analysis and Completion: A Review." *Signal Processing:Image Communication*, vol.27, no. 6, pp.579–594,2012.
[4] Wang, X., L. Su, Q. Huang, and C. Liu. "Visual Perception Based Lagrangian Rate Distortion Optimization for Video Coding," *Proceedings of the IEEE International Conf. on Image Processing*, pp.1653–1656, 2011.
[5] H.R. Wu, K.R. Rao, "Digital Video Image Quality and Perceptual Coding", *Signal Processing and Communications*, CRC Press, Inc., Boca Raton, FL, USA, 2005.
[6] E. Bosc, L. Morin, and M. Pressigout, A Content Based Method for Perceptually Driven Joint Color/depth Compression[Online] , Available: http://hal.archives–ouvertes.fr/hal–00670373.
[7] G. Tech, H. Schwarz, K. Muller, and T. Wiegand. "3D Video Coding Using the Synthesized View Distortion Change," in *Picture Coding Symposium (PCS)*, Berlin, pp.25–28, 2012.
[8] Alais, D., Blake. R., "Binocular Rivalry," MIT Press, Nov 2004.
[9] Takase S, Yukumatsu S and Bingushi K., "Local binocular fusion is involved in global binocular rivalry," *vision research*, vol.48, no.17, pp.1798–803. Jun 2008.
[10] A. Smolic, K. Müller, K. Dix, P. Merkle, P. Kauff, and T. Wiegand, "Intermediate view interpolation based on multiview video plus depth for advanced 3D video systems", in *Proceedings of the IEEE International Conf. on Image Processing*, pp. 2448 – 2451, October 2008.
[11] Patrick Ndjiki–Nya, Martin Köppel and Dimitar Doshkov, "Depth Image–based Rendering with Advanced Texture Synthesis for 3D Video", *IEEE Transactions on Multimedia*, vol. 13 , no. 3, pp. 453–465, June 2011.
[12] D.V.S.X De Silva and W. A.C Fernando, "Just noticeable difference in depth model for stereoscopic 3D displays," in *IEEE International Conf. on Multimedia and Expo*, pp. 1219–1224, Jul. 2010.
[13] Y. Zhao and L. Yu, "Binocular just noticeable–difference model for stereoscopic images," *IEEE Signal Processing Letters*, vol. 18, no. 1, pp. 19–22, Jan. 2011.
[14] X. Li, Y. Wang and D. Zhao, "Joint just noticeable difference model based on depth perception for stereoscopic images," in *IEEE Conf. on Visual Communications and Image Processing*, pp. 1 – 4, Nov. 2011.
[15] C.–H. Chou and Y.–C. Li, "A perceptually tuned sub–band image coder based on the measure of just–noticeable–distortion pro fi le," *IEEE Trans. Circuits Syst. Video Technology*, vol. 5, no. 6, pp. 467–476, Dec.1995.
[16] N. Ouerhani and H. Hügli, "Computing visual attention from scene depth," in *Proc. Int. Conf. Pattern Recog.*, pp. 375–378, 2000.
[17] N. D. B. Bruce and J. K. Tsotsos, "An attentional framework for stereo vision," in *Proc. 2nd Canadian Conf. Comput. Robot Vis.*, pp. 88–95, May 2005.
[18] P. Didyk, T. Ritschel, E. Eisemann, K. Myszkowski, and H. Seidel, "A perceptual model for disparity," in *ACM Transactions on Graphics (Proceedings SIGGRAPH 2011, Vancouver)*, vol. 30, no. 4, 2011.
[19] J. Wang, P. Le Callet, V. Ricordel, and S. Tourancheau, "Quantifying depth bias in free viewing of still stereoscopic synthetic stimuli," in *16th European Conference on Eye Movements*, Marseille, France, 2011.
[20] Yun Zhang, Gangyi Jiang, Mei Yu, Ken Chen, and Qionghai Dai, "Stereoscopic visual attention based regional bit allocation optimization for multiview video coding," *EURASIP Journal on Advances in Signal Processing*, Vol.2010, pp.24 pages, 2010.
[21] C. Ramasamy, D. House, A. Duchowski, and B. Daugherty, "Using eye tracking to analyze stereoscopic filmmaking," in *SIGGRAPH'09*,New York,2009.
[22] J. Gautier and O. Le Meur, "A time–dependent saliency model mixing center and depth bias for 2D and 3D viewing conditions", *Cognitive Computation*, vol.

4, no. 2, pp. 141–156, June 2012.
[23] H. Oh, Y.–S. Ho, "H.264–based depth map sequence coding using motion information of corresponding texture video," *Lecture Notes in Computer Science*, pp.898–90, 2006.
[24] S. Yea, A. Vetro, "RD–optimized view synthesis prediction for multiview video coding," in *Proceedings of the IEEE International Conference on Image Processing*, vol. 1, pp.209–212, 2007.
[25] Kiran Nanjunda Iyer, Kausik Maiti and Bilva Navathe, "multiview video coding using depth based 3D warping," in *IEEE International Conf. on Multimedia and Expo*, pp. 1108–1113, 2010.
[26] X. K. Yang , W. Lin, Z. Lu, E. Ong and S. S. Yao, "Just noticeable distortion model and its applications in video coding," *Signal Processing: Image Commun.*, vol. 20, no. 7, pp. 662–680, 2005.
[27] Jianhong Shen, "On the foundations of vision modeling I. Weber's law and Weberized TV (total variation) restoration," *Physica D: Nonlinear Phenomena*, vol. 175, no. 3–4, pp. 241–251, 2003.
[28] G. Egnal and R. Wildes. "Detecting binocular half occlusions: empirical comparisons of five approaches," *PAMI*, vol.24,no.8, pp.1127–1133, 2002.
[29] M. Tanimoto, T. Fujii and K. Suzuki, "Improvement of depth map estimation and view synthesis," in *ISO/IECJTC1/SC29/WG11, M15090*, Antalya, Turkey, January 2008.
[30] L. Itti, C. Koch and E. Niebur, "A model of saliency–based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp.1254–1259, 1998.
[31] H.G. Feichtinger , T. Strohmer, "Gabor analysis and algorithms: theory and applications", *Series: Applied and Numerical Harmonic Analysis*, publisher: Birkhäuser, ISBN: 0–8176–3959–4, pp. 230–235, 1998.
[32] Chan, S.C., Heung–Yeung Shum and King–To Ng, "image–based rendering and synthesis", *IEEE Signal Processing Magazine*, vol. 24, no. 6, pp. 22–33, 2007.
[33] A. Liu, W. Lin, M. Paul, C. Deng, and F. Zhang, "Just noticeable difference for image with decomposition model for separating edge and texture regions," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 20, no. 11, pp. 1648–1652, Nov. 2010.
[34] A. Vetro, P. Pandit, H. Kimata, A. Smolic, and Y. K. Wang,"Joint multi–view video model (JMVM) 7.0," in *Tech. Rep. JVT–Z207, Joint Video Team of ITU–T VCEG and ISO/IEC MPEG*, Antalya, Turkey, January 2008.

Manuscript received: January 21, 2013

Biographies

Ruimin Hu

Ruimin Hu (hrm1964@163.com) (M'09–SM'10) received his BS and MS degrees in communication and electronic systems from Nanjing University of Posts and Telecommunications in 1984 and 1990. He received his PhD degree in communications and electronic systems from Huazhong University of Science and Technology, Wuhan, in 2000. He is currently a professor at Wuhan University. He is also the associate dean of the School of Computer Science, Wuhan University, and Director of the National Engineering Research Center on Multimedia Software. His research interests include audio and video signal processing, multimedia network communication, security and surveillance technology, digital multimedia content management, and protection.

Rui Zhong

Rui Zhong (zhongrui0824@126.com) received her MS degree in telecommunications engineering from Wuhan University, China, in 2008. She is currently a PhD student in the School of Computer Science, Wuhan University. Her research interests include video compression, 3D image processing, and multimedia communications.

Zhongyuan Wang

Zhongyuan Wang (wzy_hope@163.com) received his PhD degree in communication and information systems from Wuhan University, China, in 2008. He is currently an associate professor in the School of Computer Science, Wuhan University. He is directing two projects funded by the National Natural Science Foundation Program of China. His research interests include video compression, image processing, and multimedia communications.

Zhen Han

Zhen Han (hanzhen_2003@hotmail.com) received his PhD degree in communication and information systems from Wuhan University, China, in 2009. He is currently a lecturer in the School of Computer Science, Wuhan University and is directing a project funded by the National Natural Science Foundation Program of China. His research interests include image super–resolution, and multimedia communications.

New Member of *ZTE Communications* Editorial Board

Fuji Ren received his BE degree and ME degree from Beijing University of Posts and Telecommunications in 1982 and 1985. He received his PhD degree from Hokkaido University in 1991. From 1991 to 1994, he was a chief researcher at Computer Service Kabushiki–Kaisha (CSK) Japan. In 1994, he joined the Faculty of Information Sciences, Hiroshima City University, as an associate professor. From 1996 to 1997, he was a visiting professor at New Mexico State University, USA. Since 2003, he has been the president of the AIA International Advanced Information Institute. His research interests include natural language processing; artificial intelligence; affective computing; language understanding and communication; emotional robots; multilingual, multifunctional, multimedia intelligent systems; information retrieval; and cloud computing. He is a senior member of the IEEE, a member of ACL, NLP, AAMT, IPSJ, IEICE, IASTED, and JSISE. He is also the editor–in–chief of *the International Journal of Advanced Intelligence*, vice president of CAAI, and a fellow of the Japan Federation of Engineering Societies.

Estimating Reduced–Reference Video Quality for Quality–Based Streaming Video

Luigi Atzori, Alessandro Floris, Giaime Ginesu, and Daniele D. Giusto
(Department of Electrical and Electronic Engineering, University of Cagliari, Cagliari 09123, Italy)

Abstract

Reduced–reference (RR) video–quality estimators send a small signature to the receiver. This signature comprises the original video content as well as the video stream. RR quality estimation provides reliability and involves a small data payload. While significant in theory, RR estimators have only recently been used in practice for quality monitoring and adaptive system control in streaming–video frameworks. In this paper, we classify RR algorithms according to whether they are based on a) modeling the signal distortion, b) modeling the human visual system, or c) analyzing the video signal source. We review proposed RR techniques for monitoring and controlling quality in streaming video systems.

Keywords

reduced–reference quality estimation; video streaming; adaptive rate control

1 Introduction

The paradigm of internet anywhere, any time and the diffusion of powerful end–user multimedia devices such as smartphones, tablets, networked gaming consoles, and e–book readers have led to the proliferation of new multimedia services. Such services include social TV, immersive environments, mobile gaming, HDTV over mobile, 3D virtual worlds, electronic books and newspapers, social networking, and IPTV applications to name just a few.

Services such as smartphone multimedia apps and electronic newspapers and magazines have already achieved market success. This success has been achieved because the whole design process—from content production to service activation, content consumption, and service management and updating—has been user–centered. The quality of user experience, perceived simplicity of accessing and interacting with systems and services, and concealment of complex underlying technologies determine the success or failure of these novel services.

Optimizing and managing quality of experience (QoE) is crucial for the successful deployment of future services and products. While this may seem straightforward, it is difficult to do in real end–to–end systems and networks. QoE is difficult to model, evaluate, and translate. For more than a decade, researchers have not been able to fully deal with QoE because of its dynamic end–to–end nature across a range of networks, systems, and devices.

Assessing video quality is an important part of managing QoE because video is the most important type of content in many multimedia services. Full–reference (FR) methods are used when the full availability of the reference signal is assured. This happens when designing new systems (e.g. for coding, transmission, and processing content), where FR techniques are used to analyze the effect of algorithms on the quality perceived by the end user [1].

Unfortunately, it is impossible in practice to compute these metrics at the receiver because end–users do not have access to the original frames at their terminals. As an alternative to FR methods, no–reference (NR) and reduced–reference (RR) methods have been proposed in the literature. These allow quality to be estimated at the decoder, where there is no reference signal. In NR methods, distortion of the received frames is estimated only from the reconstructed video available at the receiver or from parameters extracted from the transmitted bitstream, and the original video is not accessed. NR methods are the best choice in a broadcasting scenario because no extra data is added to the bitstream. However, NR metrics are quite complex to develop, and without any information about the reference signal, it is difficult to determine which part of the received signal is distortion and which part is the reference signal. In RR methods, a small signature of the original content is added to the video stream and sent to the receiver. At the content–producer side, a compact feature vector is extracted and transmitted to the receiver, where it is used to estimate the visual quality of the received video stream. To produce perceptually significant estimates, the receiver approximates the quality metric between the original and received streams. The feature vector is assembled in such a way that it contains sufficient information to estimate the FR metric. The availability of this side information at the receiver allows for a significantly better estimation of the received video quality in an NR scenario. The trade–off is a moderate increase in required bandwidth.

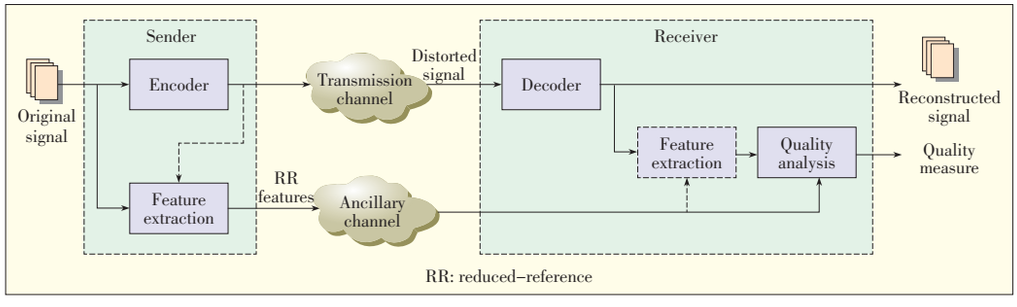
In this paper, we analyze techniques that have been proposed for quality monitoring and system control for streaming video. In section 2, we describe reference generalized

schemes. In section 3, we discuss RR techniques and how they can be used in reference scenarios. Proposed methods can be categorized according to whether they are based on modeling the signal distortion, modeling the human visual system, or analyzing the video signal source. In section 4, we discuss the use of these techniques for quality monitoring and control in practical streaming video streaming systems. We also discuss recently proposed approaches for 3D video. Section 5 concludes the paper.

2 Generalized Frameworks

We categorize RR systems as those related to measuring the quality of multimedia content and those that implement an RR quality measure in order to control the transmission bitrate or other streaming parameters.

In the former, RR quality assessment (RRQA) is used to predict quality degradation in either an image or video sequence where there is incomplete information about the reference signal. This prediction is given in the form of a set of RR features. RRQA is useful for monitoring quality in real-time visual communications over wired or wireless networks. Fig. 1 shows a generalized RRQA framework that includes a feature extraction process at the sender side and a feature extraction/quality analysis process at the receiver side. Typically, the extracted RR features (side information) have a much lower data rate than the visual data and are ideally transmitted to the receiver through an ancillary channel [2]. Although the ancillary channel is often assumed to be error-free, it may be merged with the distortion channel [3]–[7]. In such a case, the RR features would need stronger protection than the multimedia data during transmission. This protection might be achieved by stronger error-protection coding. Data is often hidden or watermarked in order to merge the features data into the media content. At the receiver side, the difference between the features extracted from the reference and the distorted images or image sequences is the quality degradation. Fig. 2 shows the RR quality assessment framework when only the distortion channel is present.



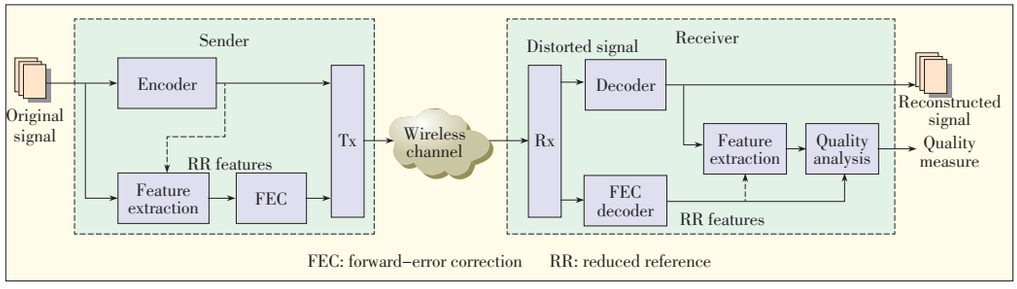
▲Figure 1. Generalized RR framework for quality assessment.

Fig. 3 shows the framework for video rate control based on RR quality metrics when typical contention-based wireless channels are used. The framework comprises a mobile station client that communicates through a wireless link with the video server. The server may be either a mobile station or a fixed system that is connected through a wired network to the access point (AP) of the wireless channel. In the later, we assume the wired part of the network is a high-throughput channel. The main subcomponents of both the server and client systems are shown in the figure. A typical video streaming scenario includes a video source, display, channel transceiver, encoder, decoder, and the buffers for each of these. The proposed architecture also comprises a rate-control module at the client side and a visual quality estimator at both the server and client sides. The rate-control module is the key component. In order to adjust the source bit rate, it monitors the channel throughput, playback buffer occupancy, and quality of the received signal computed by the visual quality estimators. The underlying encoder is capable of adjusting its encoding parameters to meet the required rate, which is computed by the rate-control algorithm.

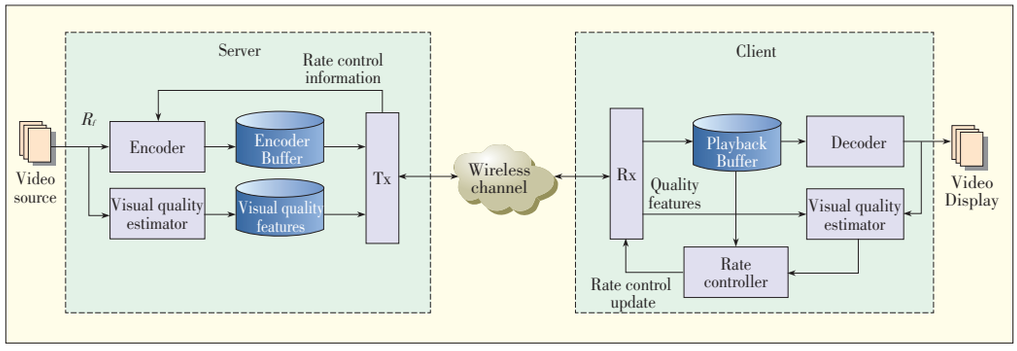
The video sequences can be generated in real time or retrieved from a video archive. When a video frame has been coded, it is segmented into one or more packets that are then delivered to the medium access control (MAC) layer and transmitted over the wireless link. In the proposed architecture, the receiver monitors the times at which it receives packets from the server. The interarrival time is the time needed for the server to conquer the channel in a multiaccess, contention-based network and to transmit the whole packet. These interarrival times are stored and processed so that information about network performance can be extracted. The received flow can be affected by errors that have not been corrected by the forward-error correction (FEC) mechanism because the mechanism has limited error-correction capabilities.

3 Reduced-Reference Quality Assessment

The general RRQA frameworks described in section 2 allow free selection of RR features, which is one of the main challenges in RRQA algorithm design. RR features should efficiently summarize the reference image, be sensitive to a variety of distortions, and relate to the visual perception of image quality. RR features should balance the data rate with accurate predictions about image quality. High data rates support the transmission of much information about the reference image, and may lead to more accurate estimation of image quality degradations. However, such data rates negatively affect transmission. Lower data rates make it easier to transmit RR informa-



▲Figure 2. Generalized RR framework for quality assessment with one transmission channel only.



▲Figure 3. Generalized RR framework for rate control.

tion, but the quality estimation is less accurate. The maximum allowed RR data rate is often given in practical implementations. When evaluating the performance of an RRQA system, consideration should be given to the tradeoff between accuracy and RR data rate.

There are three different but related types of RRQA algorithms: those based on signal distortion model, those based on the human visual system (HVS), and those based on signal source analysis. The latter two types can be used for general-purpose applications because the statistical and perceptual features being used are not limited to any specific distortion process. These two types are somewhat different to HVS, which is tuned for efficient statistical encoding of natural visual environments [8], [9]. In the following subsections, each algorithm is consigned to a best-fitting category, but many of these algorithms are based on a combination of models. For example, HVS considerations are often embedded in the other types of algorithms previously listed.

3.1 Reduced-Reference Based on Image Distortion Modeling

Algorithms based on image-distortion modeling are mostly developed for specific application environments. These algorithms provide useful, straightforward solutions when there is sufficient knowledge about the distortion process that the image or image sequence has undergone. When the distortion is standard image or video compression, a set of typical distortion artifacts, such as blurring, blocking or ringing, may be identified. Then, image features that are particularly useful for quantifying these artifacts can be determined [10], [11]. In [12], the tool for measuring compressed video quality is based on harmonic strength analysis of transmitted and received pictures at the edge. This analysis is needed to determine gain and loss information about the harmonics. The proposed metric is designed to detect blocking and blurring artifacts. The detected edges of the image are used to further process the image and extract different side information. In [13]–[15], a set of spatial and temporal features was effective for measuring distortion in standard compressed video. However, such features are limited in their generalization capability. In [16], a simple measure was designed for the perceptual qualities of MPEG-2 coded sequences. The tool uses the ratios of discrete cosine transform (DCT) coefficients. In [17], JPEG compression was considered. One block from the original image was used as an RR and was inserted over the whole image using digital watermarking. In [18], a quality index was used to evaluate perceived quality when watching video sequences on a tablet. The index was based on the results of subjective video quality assessments. In [19], artifacts (such as blurring and blocking) of the AVC/H.264 coded video sequence were determined and measured by the objective features. The measurements of these artifacts were combined into a single measurement of overall video quality. The weights of single features and the combination of these features were determined using methods based on multivariate data analysis. Generally, these methods cannot be applied beyond the distortions they are designed to capture. Table 1 shows the RR approaches discussed in this section. In the last two columns, a single value is the best performance with a combination of parameters, and a range shows the minimum and maximum performance for different datasets.

3.2 Reduced-Reference Based on Human Visual System Modeling

The second type of algorithm is based on modeling the HVS. Perceptual features that exploit computational models of low-level vision are extracted. These features are used to achieve a reduced description of the image and are not directly related to any specific distortion system. RRQA methods built on these features could be engineered for general purposes.

▼Table 1. RR approaches based on signal distortion modeling

Reference	Approach	RR Metric	Considered Distortions	Test Details	Subjective Model	Best Pearson Correlation	Other Performance Metric
[10]	Harmonic amplitude analysis	Local harmonic activity map (LHAM)	Blockiness blurriness	JPEG images	MOS	0.902	SROCC, 0.905
[12]	Discriminative analysis of local harmonic strength	Local harmonic strength (LHS)	Blockiness, blurriness	VQEG Test Phase-I video sequences	DMOS of VQEG Test Phase-I data sets	0.850	SROCC 0.860
[13]	Statistic extraction from spatial-temporal regions	Temporal, spatial, spatial-temporal and chrominance features	Blurriness, color artifacts edge noise, error blocks, frame repeats	H.261 and MPEG-2 video sequences	DMOS	0.990	NP
[14]	Statistic extraction from spatial-temporal regions	Spatial-temporal distortion metrics	Blockiness, blurriness	7 data sets of video clips	7 subjective data sets, ITU-R BT. 500 ITU-T P.910	0.780 – 0.930	NP
[15]	Feature extraction from spatial-temporal regions	10 kbit/s VQM	Blockiness, blurriness	18 data sets with 2651 video sequences, ITU-R BT.601	18 subjective data sets	NP	NP
[16]	DCT	Ratio between DCT coefficients	MPEG-2 compression	LIVE video database, MPEG-2 videos	Subjective test scores	0.950	NP
[17]	Block extraction from the original image	One 8x8 pixel block	JPEG distortion	LIVE video database, 1400 distorted images	NP	NP	Best wrong prediction ratio 24/1400
[18]	Linear function of transmission distortions	Quality index (Q.I.)	Overflow, packet loss rate, playout delay	216 H.264/AVC distorted video sequences	MOS ITU-T P.910	0.839 – 0.857	NP
[19]	Multivariate data analysis	Combination of objective video features	Blurriness, blockiness, noise	H.264/AVC video data set	2 subjective data sets, ITU-R BT. 500	0.851	SROCC 0.782

They may also be trained on different types of distortions and produce a variety of distortion-specific RRQA algorithms under the same general framework.

These methods are good for JPEG and JPEG2000 compression [20], [21]. In [22], an objective quality RR metric was proposed for color video. Because the size of the RR data is based on 12 features per frame with limited complexity, the metric is suitable for low-bandwidth transmission. The metric relies on psychovisual color space processing according to high-level HVS behavior. It also uses time-delay neural networks (TDNN). The quality criterion in [23] relies on extracting visual features from an image represented in a perceptual space. These features can be compared with perceptual color space, contrast sensitivity, psychophysical sub-band decomposition, and masking effect modeling used by the HVS. Then, a similarity metric computes the objective quality score of a distorted image by comparing the distorted image's extracted features with those extracted from the original. The performance is evaluated using three different databases and is compared with the results obtained using the three full reference metrics. The size of the side information can vary. The main drawback of this metric is its complexity. The HVS model, which is an essential part of the proposed image quality criterion, introduces a high degree of computational complexity.

In [24], a metric for RR objective perceptual image quality was proposed for use in wireless imaging. Specifically, a normalized hybrid image quality metric (NHIQM) and perceptual relevance weighted Lp-norm were proposed. HVS is trained to extract structural information from the viewing area. Image features are identified and measured according to the extent to which individual artifacts are present in a given image. The overall quality measure is then computed as a weighted sum of the features. The authors of [24] did not rely on public databases for performance evaluation but performed their own subjective tests. In [25], RR video quality was assessed by exploiting the spatial information loss and the temporal statistics of the interframe histogram. First, the change in energy of each encoded frame was measured, and the texture-masking property of the HVS was simulated. A generalized Gaussian distribution (GGD) function was then used to capture the natural statistics of the interframe histogram distribution. The distances of the histograms of the original and distorted images were modeled using the GGD functions and were computed using the city block distance measure. Finally, the spatial and temporal features were merged. In [26], the authors proposed a method that takes advantage of the HVS sensitivity to sharp changes in video. First, matching regions are determined in consecutive frames. Then, the quality of the matching regions is computed. Last, the quality of the video is calculated according to the parameters gathered in the spatial and temporal domains and using the motion activity density of the video as a controlling factor. In [27], a new metric was designed that combines the singular value decomposition (SVD) and HVS. Different singular values were extracted according to the characteristics of the video sequences. These values were used as the reference features. By comparing the original singular value (SV) with the processed value, different distortion types can be reliably measured. Furthermore, because the metric can reduce the bandwidth requirements of the system, it is suitable for measuring video quality in wireless applications. In [28], an RRQA metric was developed using a multiscale edge presentation technique in the wavelet domain. Multiscale decomposition techniques accurately simulate the psychological mechanisms of the HVS, which depends heavily on edges and contours to perceive surface properties and understanding scenes. In [29], the authors exploited contourlet transform, contrast sensitivity function (CSF), and Weber's law of just noticeable difference (JND) to define a new RRQA method. The contourlet transform is used to decompose images and extract features to mimic the multi-

channel structure of the HVS. The proposed framework is consistent with subjective perception values, and the objective assessment results accurately reflect the visual quality of images. This framework outperforms the standard PSNR and wavelet-domain image statistic (WDIS) metrics. In [30], a grouplet-based RRQA metric was proposed that makes use of the grouplet transform to efficiently characterize the image features and orientations. Then, the extracted features of the reference and distorted images were compared for quality. In [31], the authors proposed an algorithm based on the phase and magnitude of the 2D discrete Fourier transform. The phase and magnitude of the reference and distorted images were compared so that a quality score can be computed. However, the HVS has different sensitivities to different frequency components, so the frequency components are non-uniformly binned. This process also leads to reduced space representation of the image and opens up RR prospects for the proposed scheme. The phase usually conveys more information than the magnitude, so only the phase is used for RR quality assessment. Table 2 shows the RR approaches discussed in this section.

3.3 Reduced-Reference Based on Signal Source Modeling

The third type of algorithm is based on modeling natural image statistics. Because the reference image is not available in a deterministic sense, these models are often based on capturing a-priori low-level statistical properties of natural images. The basic assumption behind these approaches is that most real-world distortions disturb image statistics and make an image unnatural. This unnaturalness can be measured using models of natural image statistics and can be used to quantify degradation of the image quality. The model parameters provide a highly efficient way of summarizing the image information; thus, these methods often lead to RRQA algorithms with low RR data rates.

The Institute for Telecommunication Sciences/National Telecommunications and Information Administration (ITS/NTIA) developed a general video quality model (VQM) that, because of its performance, was selected by both ANSI and ITU as a video quality assessment standard [32]. However, the model requires a massive RR data rate to calculate the VQM value, and this prevents it from being used as an RR metric in practical systems. Spatial-temporal features and regions have been considered for trading-off between correlated subjective values and side-information overhead [13]. The proposed algorithm continually measures quality by extracting statistics from sequences of processed input and output video frames. These extracted statistics are communicated between the transmitter and receiver using an ancillary data channel of arbitrary bandwidth. Finally, individual video quality parameters are computed from these statistics. A low-rate RR metric based on the full reference metric was developed by the same authors [15]. The video quality monitoring system uses RRQA feature extraction techniques similar to those in the NTIA general VQM. A subjective data set was used to determine the optimal linear combination of the eight video quality parameters in the metric. In [33], an image quality assessment scheme using distributed source coding was proposed. The RR feature extractor comprises whitening, which is based on spread spectrum, and Walsh-Hadamard transform (WHT). The focus on reducing the bitrate of the feature vector by using distributed Slepian-Wolf source coding. In [34], an RR video quality measure was combined with a robust video watermarking approach. At the sender side, both intra- and inter-frame RR features are calculated using statistical models of natural video. The encoded features are embedded into the

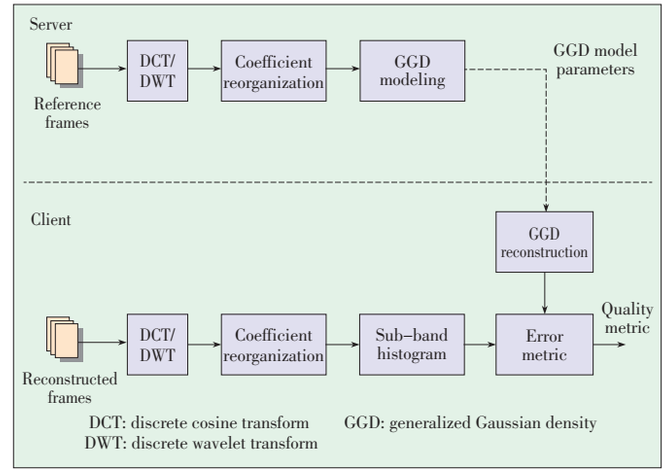
▼Table 2. Comparison of RR approaches that are based on modeling the HVS

Reference	Approach	RR Metric	Test Details	Subjective Model	Best Pearson Correlation Result	Other Performance Metric
[20]	Structural representation of the images	Structural features	LIVE image database, 45 JPEG and 45 JPEG2000	MOS	0.951 – 0.966	RMSE 0.490 – 0.800
[21]	Structural representation of the images	Global similarity Sk	IRCCyN/IVC and LIVE database	MOS	0.918– 0.961	NP
[22]	Psychovisual colorspace processing and TDNN	GHV, GHVP, P and B features	MPEG-2 video sequences from TDF	TDF subjective quality assessment, DMOS	0.942	RMSE 0.086
[23]	Psychophysical description of the images	Global similarity S	IVC, LIVE and Toyama image databases	MOS, VQEG Test Phase-I FR-TV	0.887 – 0.972	SROCC 0.887 – 0.953
[24]	Structural representation of the images	Normalized hybrid image quality metric (NHIQM) and Lp-norm	2 sets of 40 JPEG images	MOS, ITU-R BT. 500	NHIQM: 0.843 Lp-norm: 0.846	NP
[25]	Feature extraction from the spatial perspective	Visual quality index (VQI)	LIVE video database, MPEG-2 and H.264 videos	DMOS	0.555 – 0.757	SROCC 0.558 – 0.749 RMSE 6.722–8.586
[26]	Spatial-temporal assessment of quality (STAQ)	MeanS	LIVE video database, MPEG-2 and H.264 videos	Subjective results	0.720 – 0.913	SROCC 0.762 – 0.878
[27]	Singular value decomposition (SVD)	QImgSVD	H.264/AVC video sequences	Subjective scores, MSU VQMT, ITU-R BT. 500	0.814	SROCC 0.794 RMSE 0.115
[28]	Multi-scale analysis	Multi-scale modular maxima similarity (M3S)	LIVE image database, 233 JPEG and 227 JPEG2000	MOS, VQEG Test Phase-I FR-TV	NP	SROCC JPEG: 0.962 JPEG2000: 0.936
[29]	Contourlet transform, CSF and JND	Transformed city-block distance	LIVE image database, JPEG and JPEG2000	MOS	0.916–0.949	SROCC 0.9032–0.9309
[30]	Grouplet transform	Local similarity measure	LIVE image database, JPEG and JPEG2000	MOS	JPEG: 0.953 JPEG2000: 0.971	SROCC JPEG: 0.960 JPEG2000: 0.956
[31]	2D discrete Fourier transform (DFT)	Phase and magnitude of DFT	7 image databases, 2 video databases	9 subjective datasets, VQEG Phase-I/II test	0.759–0.954	SROCC 0.7393 – 0.956

same video signal using a robust angle-quantization-index, modulation-based watermarking method. At the receiver side, the RR features are extracted and decoded from the distorted video and are used to predict the perceptual degradation of the video signal.

In [35] and [36], the differences between entropies of the wavelet coefficients of the reference and distorted images were used to measure changes in the image information. The algorithm is flexible in terms of the amount of side information required from the reference, which may be as little as a single scalar per frame. In [37], used a color distribution was used to evaluate the perceived image quality. Descriptors based on the color correlogram were used to analyze alterations in color distribution that occur because of distortion. In [38], an RR approach was proposed to measure temporal motion smoothness of a video sequence. By examining the temporal variations in local phase structures in the complex wavelet transform domain, the proposed measure can detect a wide range of well-known distortions. In addition, the proposed algorithm does not require costly motion estimation and has a low RR data rate. This makes it much better for real-world visual communication applications. In [39], natural images were very specifically distributed in the gradient domain, so the authors measured the changes of image statistics in this domain. These changes in image statistics correspond to the degree and types of image distortion. The proposed method can be used for all distortion types. In [40] and [41], an RRQA method based on estimating the structural similarity index (SSIM) was proposed. In [42], the authors proposed an effective RRQA metric using statistics based on the divisive normalization transform (DNT) of the contourlet domain. The marginal histogram of the contourlet coefficients in each sub-band are fitted by Gaussian distribution after DNT. The standard derivations of the fitted Gaussian transform and fitted error are extracted as feature parameters.

In [43] and [3], the marginal distribution of the wavelet sub-band coefficients was modeled using a GGD function. GGD model parameters are used as RR features to quantify the variations of marginal distributions in the distorted image. This general-purpose approach has been successful because it does not require any training and has a low RR data rate. However, it still performs reasonably when tested with a wide range of image distortion types. In [44], the model was further improved by employing a nonlinear divisive normalization transform (DNT) after the linear wavelet decomposition. This improves quality prediction, especially when images with different distortion types are mixed together. In [24], results were compared with the RR metric and peak signal-to-noise ratio (PSNR). In [45] and [46], the perceived video quality was estimated on a frame basis. This perceived video quality is the distance between the distribution of DCT coefficients at the receiver side and the generalized GGD-modeled distribution of the same coefficients of the original signal, based on the framework presented in [47]. Fig. 4 shows a block diagram of the GGD-based technique.



▲Figure 4. GGD-based technique for estimating video quality.

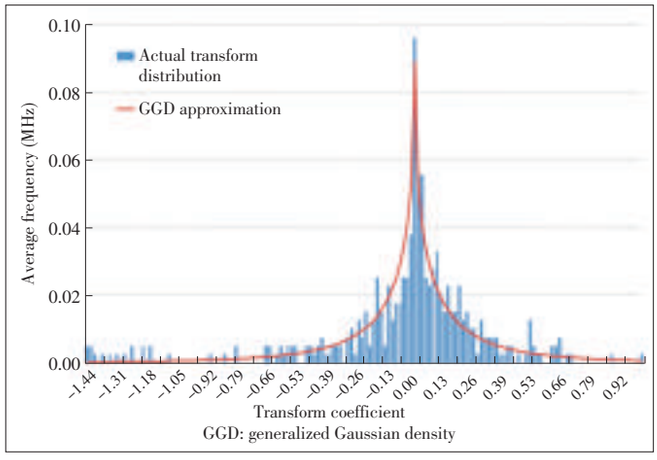
At the server (transmitter) side, the reference frames are first transformed using DCT or discrete wavelet transform (DWT). Then, the transform coefficients may be spatially rearranged so that those representing similar frequencies often are grouped in a dyadic way. Subsequently, a GGD function is used to model the coefficient distribution of each frequency sub-band:

$$P_{\alpha,\beta}(x) = \frac{\alpha}{2\beta \cdot \Gamma(1/\alpha)} exp\left[-(|x|/\beta)^{\alpha}\right] \quad (1)$$

where $\Gamma(\cdot)$ is the gamma function and α and β are the parameters of the GGD model. Fig. 5 shows an example of the GGD modeling. Only a few parameters are sent to the receiver for each frequency sub-band. These parameters are often α, β, and the city block distance between the actual sub-band distribution and its GGD approximation. At the receiver side, the received frames undergo similar processing; however, the GGD modeling is substituted by the distortion metric. Such an estimate is generally derived from the linear combination of the differences between the actual distribution of transform coefficients from the distorted frame and the distribution modeled by the received GGD parameters. Table 3 shows the RR approaches based on signal source modeling.

4 Use of RRQA in Streaming Video Applications

In this section, we give an overview of practical applications of RR metrics in streaming video frameworks. In some works, RR video quality estimation methods are used for video quality monitoring (VQM) of IPTV services. Complexity is one of the most important factors in VQM because of real-time constraints and hardware with limited capabilities. Set-top boxes, for example, have very limited computing and memory resourc-



▲Figure 5. Actual distribution of transform coefficients and its GGD approximation.

es. RR methods are used because they allow feature information to be transmitted with very little resource consumption and low uplink bandwidth. In [48], a VQM scheme uses a modified version of PSNR (called networked PSNR) that is based only on the RR visual rhythm data, which is used as feature information in the RR method. Two practical scenarios were proposed in which a quality monitoring server uses the proposed metric to evaluate the video impairments according to the packet loss experienced by the end user. In [49], an RR video quality estimation method makes use of the difference in activity values between the original video and received video. Temporal sub-sampling and partial bit transmission of activity values help accurately estimate the subjective video quality, and only a small amount of extra information is needed.

In [50], real-time video sequence matching and RR assessing techniques are proposed for IPTV. After the processing procedures have been completed, the QoE indicators, such as edge, block, blur, color, and jerkiness, are measured by the proposed RR real-time video measurement methods in order to define the VQM metric. In [51], the authors discuss color error, which is one of the most common artifacts in compressed and transmitted video through the IP network. They propose an FR or RR quality-assessment method based on a color error measure in order to monitor the color quality in IPTV services.

In [52], a quality of interest points (QIP) RR metric, described in [53], was used for JPEG 2000 wireless (JPWL) transmission over MIMO channels. Depending on the object's saliency, the interest points can predict a variation in the image. In the proposed scheme, QIP is used as a layer selector that is able to detect any reduction in the perceived quality while decoding an additional layer. All the possible configurations are decoded with JPWL robust decoder, and each configuration is evaluated by QIP giving a score from 0 (very bad quality) to 1 (excellent quality). Finally, the configuration with the best QIP score has the best QoE. Beyond basic quality assessment, RRQA metrics have been used in other scenarios. In [54], the authors propose a technique that uses image analysis to automatically detect camera anomalies. The technique allows good image quality in surveillance videos by correcting the field of view. The technique involves first extracting RR features from multiple regions in the surveillance image. Then, abnormal events in features are detected by analyzing image quality and field-of-view variations. Events are detected by statistically calculating accumulated variations in the temporal domain. In [55], the authors consider modeling the visibility of individual and multiple packet losses in H.264 videos. They propose a model for predicting the visibility of multiple packet losses and demonstrate its performance with dual losses (two nearby packet losses). To extract the factors affecting visibility, an RR method is used because it accesses the decoder's reconstructed video (with losses) and factors extracted from the

▼Table 3. Comparison of RR approaches that are based on modeling the signal source

Reference	Approach	RR Metric	Test Details	Subjective Model	Best Pearson Correlation Result	Other Performance Metric
[32]	Feature extraction from optimally sized spatial/ temporal regions of the video	General Video Quality Model (VQM)	11 video data sets, 1536 video sequences	11 subjective data sets, ITU-R BT. 500, ITU-T P.910, VQEG FR-TV Phase-II	0.865–0.980	NP
[35]	Spatial and temporal entropic differences	SRRED, RRED and STRRED indices	LIVE video database, 150 distorted videos	Subjective quality scores of LIVE video database	0.415–0.832	SROCC 0.385–0.819
[36]	Entropic differences	RRED	LIVE image database and Tampere image database (TID)	DMOS of the LIVE and TID video database	0.784–0.984	SROCC 0.277–0.978
[37]	Color distribution information	Descriptors based on the color correlogram	2nd release of the LIVE image database	DMOS of the LIVE video database	0.777–0.991	NP
[39]	Dual derivative priors of the image	Vertical derivative	7 datasets from the LIVE image database	MOS	0.882–0.980	SROCC 0.871–0.990
[40]	SSIM estimation	RR-SSIM	6 image databases	MOS	0.800–0.802	0.800–0.805
[41]	VSSIM approximation	RR-VSSIM	2 H.264/AVC video sequences	NP	0.770–0.970*	NP
[42]	Contourlet transform	Statistical features	LIVE image database	DMOS from LIVE image database	0.908 –0.969	SROCC 0.876–0.975
[43]	Wavelet-domain image statistic model	Statistical features	LIVE image database	MOS, VQEG FR-TV Phase-I	0.845–0.969	SROCC 0.833–0.947
[44]	Divisive normaliza-tion-based image representation	Statistical features	LIVE and Cornell-VCL A57 image databases	Subjective scores from LIVE and Cornell- VCL A57 databases	0.538–0.917	SROCC 0.511–0.929
[47]	Statistical modeling of DCT coefficient distributions	Visual distance Vdist	LIVE image database	VQEG HDTV test	0.845–0.931	SROCC 0.838 – 0.930 RMSE 6.790–13.50

*computed between RR-VSSIM and FR-VSSIM

encoded video.

In the literature, there are very few works on the implementation of an RR quality measure for controlling the transmission bitrate or other streaming parameters. There are two works the describe the implementation of RR measures in a process commonly known as rate-distortion optimization (RDO), which is used to convey the sequence of images with minimum possible perceived distortion within the available bitrate. In [56], the authors describe a computationally efficient video-distortion metric that can operate in FR or RR mode. The metric guides an RDO rate-control algorithm for MPEG-2 video compression. Specifically, it is used to generate spatial distortion maps that are summed into macroblock-level and frame-level distortion scores in order to optimize the frame rate allocations for an MPEG-2 video coder. The coded sequences produced by the algorithm have fewer visible macroblock edges (blockiness), and the textured areas are sharper. Furthermore, the proposed metric is well correlated with subjective scores. In [57], the proposed RDO scheme is based on a novel RR statistical SSIM estimation algorithm and a source-side-information combined-rate model for H.264/AVC video coding. The adaptive Lagrange multiplier method was used at both frame and macroblock levels to select the best coding mode and achieve the best-rate SSIM performance. Experiments showed that the proposed scheme significantly reduces the rate but maintains the same range of SSIM values. Compared with the RDO scheme, visual quality also improved. In [45], a scheme is proposed for controlling the source rate in streaming video sequences. The scheme relies on RR quality estimation and is the only one of its kind mentioned in the literature. The server extracts important features of the original video, and then these features are coded and sent through the channel along with the video sequence. They are then used at the decoder to compute the actual quality. The observed quality is analyzed to obtain information about the effect of the source rate for a given system configuration. At the receiver side, decisions are made on the optimal encoding rate to maximize the perceived quality at the user side. The rate is adjusted on a per-window basis to compensate low-throughput periods with high-throughput periods. This eliminates abrupt changes in video quality caused by sudden variations in the channel throughput. RRQA optimizes user-perceived video quality from the actual signal that is affected by all possible impairments. Experiments show that the RR quality metric allows perceived quality at the decoder side to be accurately estimated. It also correlates with other FR metrics. Three methods for evaluating the performance of the proposed algorithm in transmitting video sequences are compared: transmission at constant bitrate, control of the starvation probability, and the proposed method. The proposed method gives the best overall results for all quality metrics and is second best at avoiding occurrences of starvation. The proposed method is also applicable to any channel conditions and coding settings and does not require any a-priori knowledge of system configuration or transmission conditions.

Video quality metrics have only very recently been used to assess 3D video quality, and this field deserves a particular attention. Exploiting immersive video implies the transmission of huge amounts of data because of the multichannel nature of such a format and high video resolutions. RRQA becomes even more interesting because it theoretically allows for a reliable measuring of quality at the receiver side and adds little side information to the video data. 3D video data has great redundancy that can be used to optimize the extraction of relevant features. From a technical perspective, 3D RRQA differs from the other approaches in that it combines both intraframe and interframe aspects with depth information in order to extract RR features for quality assessment [58], [59]. Table 4 summarizes the RRQA approaches used in streaming video.

5 Conclusion

In this paper, we have reviewed reduced-reference (RR) quality metrics and categorized them according to whether

▼Table 4. Comparison of RR approaches used in video streaming applications

Reference	Application	Approach	RR Metric	Test Details	Subjective Model	Best Pearson Correlation Result	Other Performance Metric
[48]	VQM of IPTV services	Feature extraction from video frames	VR, PSNR and NMOS	5 H.264/AVC video sequences	MOS	0.700–0.930[1] 0.890–0.940[2]	NP
[49]	VQM of IPTV services	Activity difference frame analysis	Activity-difference values	MPEG-2 and H.264 video sequences	Subjective scores, ITU-T P.910	0.911	NP
[51]	VQM of IPTV services	Color error measure	Hue and saturation of the frames	2 MPEG-4 video sequences	MOS	NP	NP
[52] [53]	JPWL transmission over MIMO channel	Interest point and object saliency in color images	Quality interest point (QIP)	LIVE, Toyama and TID image databases	MOS, VQEG	JPEG: 0.987 JPEG2000: 0.977	NP
[54]	Automatic event detection for camera anomaly	Region-based edge energy	Feature's energy	75 MPEG-1 video sequences	NP	NP	Precision rate 88.9%
[56]	Rate distortion optimization	Visual model	Coefficients of luminance channel	120 distorted video sequences	DMOS, VQEG, ITU-R BT.500	0.896	NP
[57]	Rate distortion optimization	SSIM estimation model	RR-SSIM	H.264/AVC video sequences	NP	0.996[3]	NP
[45] [47]	Source rate control scheme for streaming video sequences over wireless channels	Statistical modeling of DCT coefficient distributions	Visual distance Vdist	LIVE image database	VQEG HDTV test	0.845–0.931	SROCC 0.838–0.930 RMSE 6.790–13.50
[58]	3D color plus depth video compression and transmission	Edges/contours extraction	Depth map and color image quality	H.264/AVC 3D color plus depth video sequences	MOS	0.927–0.979	RMSE 0.006–0.011

[1] computed between RR-NPSNR and FR-NPSNR [2] computed between NMOS and MOS [3] computed between RR-SSIM and FR-SSIM

they are based on modeling the signal distortion, modeling the human visual system, or analyzing the video signal source. We have reviewed studies on the implementation of RR techniques in practical systems to monitor and control quality in streaming video systems.

RR methods do not require full access to reference signals; they only need a small amount of information in the form of a set of extracted features. These methods are ideal for evaluating the quality of multimedia content at the receiver side—at the far end of the communication chain. Because of they are not complex and have a low features data rate, they can be used for quasi real-time or streaming video applications. On the other hand, their reliability in several implementations can be an issue. RRQA algorithms also fail to given an integrated estimate of subjective factors such as user device, environment conditions, and interface perception, all of which are typical in QoE approaches. Future work will probably be done on integrating sensor data produced by user devices in order to provide a more accurate and robust estimate of perceived quality. Further work is also needed on applications related to 3D images and video sequences so that RR features can measure the quality of such multimedia content while balancing the data rate of RR features with accurate quality prediction.

References

[1] G. Ginesu, F. Massidda, and D. D. Giusto, "A multi-factors approach for image quality assessment based on a human visual system model," *Signal Processing: Image Communication*, vol. 21, no. 4, pp. 316–333, Apr. 2006.
[2] Z. Wang and A. C. Bovik, *Modern Image Quality Assessment*, San Rafael, CA: Morgan & Claypool Publishers, Mar. 2006.
[3] Z. Wang, G. Wu, H. R. Sheikh, E. P. Simoncelli, En-Hui Yang, and A. C. Bovik, "Quality-aware images," *IEEE Trans. Image Processing*, vol. 15, no. 6, pp. 1680–1689, June 2006.
[4] B. Hiremath, Q. Li, and Z. Wang, "Quality-aware video," in *Proc. IEEE Int. Conf. Image Proc., San* Antonio, TX, Sept. 2007, vol. 3, pp. 469–472.
[5] K. Zeng and Z. Wang, "Quality-aware video based on robust embedding of intra- and inter-frame reduced-reference features," *17th IEEE Int. Conf. on Image Processing (ICIP)*, 26–29 Sept. 2010, pp. 3229–3232.
[6] S. Altous, M. K. Samee, and J. Gotze, "Reduced Reference Image Quality Assessment for JPEG Distortion," *ELMAR Proceedings*, Zadar, Sept. 2011, pp. 97–100.
[7] A. N. Avanaki, S. Sodagari, and A. Diyanat, "Reduced reference image quality assessment metric using optimized parameterized wavelet watermarking," *9th Int. Conf. on Signal Processing (ICSP)*, Beijing, Oct. 2008, pp. 868–871.
[8] H. B. Barlow, "Possible principles underlying the transformation of sensory messages," in *Sensory Communication*, W. A. Rosenblith, Ed. Cambridge, MA: MIT Press, 1961, pp. 217–234.
[9] E. P. Simoncelli and B. Olshausen, "Natural image statistics and neural representation," *Annual Review of Neuroscience*, vol. 24, pp. 1193–1216, May 2001.
[10] I. P. Gunawan and M. Ghanbari, "Reduced reference picture quality estimation by using local harmonic amplitude information," in *Proc. London Commun. Symp.*, Sept. 2003, pp. 137–140.
[11] T. M. Kusuma and H.-J. Zepernick, "A reduced-reference perceptual quality metric for in-service image quality assessment," in *Proc. 1st Workshop on Mobile Future and Symposium on Trends in Communications*. Fei Stu, Bratislava, Slovakia, Oct. 2003, pp. 71–74.
[12] I. Gunawan and M. Ghanbari, "Reduced-reference video quality assessment using discriminative local harmonic strength with motion consideration," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 18, no. 1, pp. 71–83, Jan. 2008.
[13] S. Wolf and M. Pinson, "In-service performance metrics for mpeg-2 video systems," in *Proc. Made to Measure 98 – Measurement Techniques of the Digital Age Technical Seminar, International Academy of Broadcasting (IAB)*, ITU and Technical University of Braunschweig, Montreux, Switzerland, 1998, pp. 12–13.
[14] S. Wolf and M. H. Pinson, "Spatio-temporal distortion metrics for in-service quality monitoring of any digital video system," *Proc. SPIE*, vol. 3845, pp. 266–277, 1999.
[15] S. Wolf and M. H. Pinson, "Low bandwidth reduced reference video quality monitoring system," in *Proc. Video Processing and Quality Metrics for Consumer Electronics*, Scottsdale, Arizona, Jan. 2005, pp. 23–25.
[16] S. Yang, "Reduced reference MPEG-2 picture quality measure based on ratio of DCT coefficients," *Electronics Letters*, vol. 47, no. 6, pp. 382–383, March 2011.
[17] S. Altous, M. K. Samee, and J. Gotze, "Reduced Reference Image Quality Assessment for JPEG Distortion," *ELMAR Proceedings*, Zadar, Sept. 2011, pp. 97–100.
[18] L. Atzori, G. Ginesu, D. D. Giusto, and A. Floris, "QoE Assessment of Multimedia Video Consumption on Tablet Devices," *IEEE Globecom Workshop on Quality of Experience for Multimedia Communications*, Anaheim, California, Dec. 2012.
[19] T. Oelbaum and K. Diepold, "Building a reduced reference video quality metric with very low overhead using multivariate data analysis," *J. Syst. Cybern. Informatics*, vol. 6, no. 5, pp. 81–86, 2008.
[20] M. Carnec, P. Le Callet, and D. Barba, "An image quality assessment method based on perception of structural information," in *Proc. IEEE Int. Conf. Image Processing*, Barcelona, Catalonia, Sept. 2003, vol. 3, pp. 185–188.
[21] M. Carnec, P. Le Callet, and D. Barba, "Visual features for image quality assessment with reduced reference," in *Proc. IEEE Int. Conf. Image Processing*, Genoa, Italy, Sept. 2005, vol. 1, pp. 421–424.
[22] P. Le Callet, C. Viard-Gaudin, and D. Barba, "Continuous quality assessment of MPEG2 video with reduced reference," in *Proc. Int. Workshop Video Process. Quality Metrics for Consumer Electron.*, Scottsdale, AZ, Jan. 2005.
[23] M Carnec, P Le Callet, and D Barba, "Objective quality assessment of color images based on a generic perceptual reduced reference," *Signal Processing: Image Communication*, vol. 23, no. 4, pp. 239–256, Apr. 2008.
[24] U. Engelke, M. Kusuma, H.-J. Zepernick, and M. Caldera, "Reduced-Reference Metric Design for Objective Perceptual Quality Assessment in Wireless Imaging," *Signal Processing: Image Communication*, vol. 24, no. 7, pp. 525–547, 2009.
[25] L. Ma, S. Li, and K. N. Ngan, "Reduced-Reference Video Quality Assessment of Compressed Video Sequences," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 22, no. 10, pp. 1441–1456, Oct. 2012.
[26] S. A. Amirshahi and M. Larabi, "Spatial-temporal Video Quality Metric based on an estimation of QoE," *3rd Int. Workshop on Quality of Multimedia Experience (QoMEX)*, Mechelen, 7–9 Sept. 2011, pp. 84–89.
[27] F. Yuan and E. Cheng, "Reduced-Reference Metric Design for Video Quality Measurement in Wireless Application," *11th IEEE Int. Conf on Communication Technology (ICCT)*, Hangzhou, Nov. 2008, pp. 641–644.
[28] G. Zhai, W. Zhang, X. Yang, and Y. Xu, "Image Quality Assessment Metrics Based on Multi-scale Edge Presentation," *IEEE Workshop on Signal Processing Systems Design and Implementation*, Nov. 2005, pp. 331–336.
[29] D. Tao, X. Li , W. Lu, and X. Gao, "Reduced-Reference IQA in Contourlet Domain," *IEEE Trans. on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 39, no. 6, pp. 1623–1627, Dec. 2009.
[30] A. Maalouf, M.-C. Larabi, and C. Fernandez-Maloigne, "A Grouplet-Based Reduced Reference Image Quality Assessment," *Int. Workshop on Quality of Multimedia Experience (QoMEx)*, San Diego, CA, July 2009, pp. 59–63.
[31] M. Narwaria, W. Lin, I. V. McLoughlin, S. Emmanuel, and L.-T. Chia, "Fourier Transform-Based Scalable Image Quality Measure," *IEEE Trans. on Image Processing*, vol. 21, no. 8, pp. 3364–3377, Aug. 2012.
[32] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Trans. Broadcasting*, vol. 50, no. 3, pp. 312–322, Sept. 2004.
[33] K. Chono, Y.-C. Lin, D. Varodayan, Y. Miyamoto, and B. Girod, "Reduced-reference image quality assessment using distributed source coding," *IEEE Int. Conf on Multimedia and Expo*, Hannover, Apr.2008, pp. 609–612.
[34] K. Zeng and Z. Wang, "Quality-aware video based on robust embedding of intra- and inter-frame reduced-reference features," *17th IEEE Int. Conf. on Image Processing (ICIP)*, Hong Kong, 26–29 Sept. 2010, pp. 3229–3232.
[35] R. Soundararajan and A. C. Bovik, "Video Quality Assessment by Reduced

Reference Spatio-temporal Entropic Differencing", *IEEE Trans. on Circuits and Systems for Video Technology*, no. 99, 2012
[36] R. Soundararajan and A. C. Bovik, "RRED Indices: Reduced Reference Entropic Differencing for Image Quality Assessment" *IEEE Trans. on Image Processing*, vol. 21, no. 2, pp. 517–526, Feb. 2012
[37] J. A. Redi, P. Gastaldo, I. Heynderickx, and R. Zunino, "Color Distribution Information for the Reduced-Reference Assessment of Perceived Image Quality", *IEEE Tran. on Circuits and Systems for Video Technology*, vol. 20, no. 12, pp. 1757–1769, Dec. 2010
[38] K. Zeng and Z. Wang, "Temporal motion smoothness measurement for reduced-reference video quality assessment", *IEEE Int. Conf on Acoustics Speech and Signal Processing (ICASSP)*, Dallas, TX, March 2010, pp. 1010–1013.
[39] G. Cheng and L. Cheng, "Reduced reference image quality assessment based on dual derivative priors," *Electronics Letters*, vol. 45, no. 18, pp. 937–939, Aug. 2009
[40] A. Rehman and Z. Wang, "Reduced-Reference Image Quality Assessment by Structural Similarity Estimation," *IEEE Trans. on Image Processing*, vol. 21, no. 8, Aug. 2012.
[41] A. Albonico, G. Valenzise, M. Naccari, M. Tagliasacchi, and S. Tubaro, "A reduced-reference video structural similarity metric based on no-reference estimation of channel-induced distortion," *IEEE Int. Conf on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, April 2009, pp. 1857–1860.
[42] X. Wang, G. Jiang, and M. Yu, "Reduced Reference Image Quality Assessment Based on Contourlet Domain and Natural Image Statistics," *5th Int. Conf. on Image and Graphics (ICIG)*, Xi'an, Shanxi, Sep. 2009, pp. 45–50.
[43] Z. Wang and E. P. Simoncelli, "Reduced-reference image quality assessment using a wavelet-domain natural image statistic model," in *Human Vision and Electronic Imaging X, Proc. SPIE*, San Jose, CA, Jan. 2005, vol. 5666, pp. 149–159.
[44] Q. Li and Z. Wang, "Reduced-reference image quality assessment using divisive normalization-based image representation," *IEEE Journal on Selected Topics in Signal Processing*, vol. 3, no. 2, pp. 202–211, 2009.
[45] L. Atzori, G. Ginesu, D. D. Giusto, and A. Floris, "Streaming Video over Wireless Channels: Exploiting Reduced-Reference Quality Estimation at the User-Side," *Signal Processing – Image Communications*, vol. 27, no. 10, pp. 1049– 1065, Nov. 2012.
[46] L. Atzori, G. Ginesu, D. D. Giusto and A. Floris, "Rate Control based on Reduced-Reference Image Quality Estimation for Streaming Video over Wireless Channels", *IEEE International Conference on Communications (ICC)*, Ottawa, ON, June 2012, pp. 2021–2025.
[47] L. Ma, S. Li, F. Zhang, and K.N. Ngan, "Reduced-Reference Image Quality Assessment Using Reorganized DCT-Based Image Representation," *IEEE Trans. on Multimedia*, vol. 13 no. 4, pp. 824 – 829, Aug. 2011.
[48] J. C. Kwon, S. H. Jang, Y. Chin, and S.-J. Oh, "A novel video quality impairment monitoring scheme over an IPTV service with packet loss," *2nd Int. Workshop on Quality of Multimedia Experience (QoMEX)*, Trondheim, June 2010, pp. 224–229.
[49] T. Yamada, Y. Miyamoto, and M. Serizawa, "End-user video-quality estimation based on a Reduced-Reference model employing activity-difference for IPTV services," *Digest of Technical Papers Int. Conf. on Consumer Electronics (ICCE)*, Las Vegas, NV, Jan. 2009, pp. 1–2.
[50] J. Kim and S. Kim, "Accurate matching and assessing methodology for distorted IPTV contents," *Int. Conf. on ICT Convergence (ICTC)*, Seoul, Sept. 2011, pp. 766–767.
[51] M. A. Hasan, W. Kim, C. Kim, J. Kim, H.-W. Lee, and W. Ryu, "Color Error Measure for IPTV Service Quality Evaluation," *10th Int. Conf on Advanced Communication Technology (ICACT)*, Gangwon-Do, Feb. 2008, pp. 1407–1412.
[52] J. Abot, M. Nauge, C. Perrine, C. Larabi, C. Bergeron, Y. Pousset, and C. Olivier, "A robust content-based JPWL transmission over a realistic MIMO channel under perceptual constraints," *18th IEEE Int. Conf on Image Processing (ICIP)*, Brussels, Sept. 2011, pp. 3241–3244.
[53] M. Nauge, M.-C. Larabi, and C. Fernandez, "A reduced-reference metric based on the interest points in color images," *Picture Coding Symposium (PCS)*, Nagoya, Dec. 2010, pp. 610–613.
[54] Y. K. Wang, C.-T. Fan, K.-Y. Cheng, and P. S. Deng, "Real-time camera anomaly detection for real-world video surveillance," *Int. Conf on Machine Learning and Cybernetics (ICMLC)*, Guilin, July 2011, pp. 1520–1525.
[55] S. Kanumuri, S. G. Subramanian, P. C. Cosman, and A. R. Reibman, "Predicting H.264 Packet Loss Visibility using a Generalized Linear Model," *IEEE Int. Conf. on Image Processing*, Atlanta, GA, Oct. 2006, pp. 2245–2248.
[56] M. Masry, S. S. Hemami, and Y. Sermadevi, "A Scalable Wavelet-Based Video Distortion Metric and Applications," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 16, no. 2, pp. 260–273, Feb. 2006.
[57] S. Wang, A. Rehman, Z. Wang, S. Ma, and W. Gao, "SSIM-Motivated Rate-Distortion Optimization for Video Coding," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 22, no. 4, pp. 516–529, April 2012.
[58] C.T.E.R. Hewage and M. G. Martini, "Edge-Based Reduced-Reference Quality Metric for 3-D Video Compression and Transmission, *IEEE J. of Selected Topics in Signal Processing*, vol. 6, no. 5, pp. 471–482, Sep. 2012
[59] G. Nur, G. B. Akar, H. Gokmen, "Reduced Reference 3D Video Quality Assessment based on Cartoon Effect", 2012 Nem Summit, 16–18 Oct. 2012.

Manuscript received: January 8, 2013

Biographies

Luigi Atzori

Luigi Atzori (l.atzori@diee.unica.it) has been an assistant professor at the University of Cagliari, Italy, since 2000. His main research interest is service management in next generation networks, with particular attention to QoS, service-oriented networking, bandwidth management, and multimedia networking. He has published more than 80 journal articles and refereed conference papers. Dr. Atzori has received the Telecom Italia award for an outstanding Master's thesis in telecommunications and has been awarded a Fulbright Scholarship (11/2003–05/2004) to work on video streaming in the Department of Electrical and Computer Engineering, University of Arizona. He is a senior member of the IEEE, a member of the IEEE Multimedia Communications Committee (MMTC), and co-chair of the MMTC IG on Quality of Experience. He has been the editor for Wireless Networks Journal, published by ACM/Springer, and guest editor of IEEE Communications Magazine, Monet Journal, and Signal Processing: Image Communications.

Alessandro Floris

Alessandro Floris (alessandro.floris@diee.unica.it) received his MSc degree in electronic engineering from the University of Cagliari in 2011. He was awarded a CNIT research grant from June 2011 to June 2012. He is currently research fellow in the Department of Electric and Electronic Engineering (DIEE), University of Cagliari, Italy. His research interests are QoE estimation for multimedia streaming using reduced-reference approaches.

Giaime Ginesu

Giaime Ginesu (g.ginesu@diee.unica.it) received his MSc degree in electronic engineering in 2001. His thesis was on thermal image processing and pattern recognition. In 2005, he received his PhD degree in electronic engineering from the University of Cagliari, Italy. In 2001, he worked at the Institute for Telecommunications, Technical University of Braunschweig, Germany. There he worked on thermographic image processing with Professor V. Maergner. In 2003, he was a visiting scholar at the Rensselaer Polytechnic Institute, New York, and worked on volumetric data coding with Professor W. A. Pearlman. He is currently an adjunct professor at the University of Cagliari, Italy. Since 2007, he has been involved in ICT project management at the DG for Technological Innovation, Regione Autonoma della Sardegna. His research interests include signal processing, standards, and transmission. He is a member of IEEE.

Daniele D. Giusto

Daniele D. Giusto (ddgiusto@unica.it) has been a full professor of telecommunications at the University of Cagliari and director of CNIT Multimedia Communications Lab since 2002. His research interests include image and video processing and coding, multimedia systems, digital television, pictorial databases, and personal communications. Professor Giusto is a senior member of IEEE, the recipient of the 1993 AEI Ottavio Bonazzi Best Paper Award, and co-recipient of the 1998 IEEE Chester Sall Best Paper Award. Since 1999, he has been the head of the Italian delegation in the ISO-JPEG Standardization Committee. In 2007, he was appointed to the IEEE Standard Activities board (RevCom).

Human–Centric Composite–Quality Modeling and Assessment for Virtual Desktop Clouds

Yingxiao Xu[1], Prasad Calyam[2], David Welling[3,4], Saravanan Mohan[3,4], Alex Berryman[3,4], and Rajiv Ramnath[4]
(1. Fudan University, Shanghai 200433, China;
2. University of Missouri, MO 65201, USA;
3. Ohio Supercomputer Center/OARnet, OH 43212, USA;
4. The Ohio State University, OH 43210, USA)

Abstract

There are several motivations, such as mobility, cost, and security, that are behind the trend of traditional desktop users transitioning to thin–client–based virtual desktop clouds (VDCs). Such a trend has led to the rising importance of human–centric performance modeling and assessment within user communities that are increasingly making use of desktop virtualization. In this paper, we present a novel reference architecture and its easily deployable implementation for modeling and assessing objective user quality of experience (QoE) in VDCs. This architecture eliminates the need for expensive, time–consuming subjective testing and incorporates finite–state machine representations for user workload generation. It also incorporates slow–motion benchmarking with deep–packet inspection of application task performance affected by QoS variations. In this way, a "composite–quality" metric model of user QoE can be derived. We show how this metric can be customized to a particular user group profile with different application sets and can be used to a) identify dominant performance indicators and troubleshoot bottlenecks and b) obtain both absolute and relative objective user QoE measurements needed for pertinent selection of thin–client encoding configurations in VDCs. We validate our composite–quality modeling and assessment methodology by using subjective and objective user QoE measurements in a real–world VDC called VDPilot, which uses RDP and PCoIP thin–client protocols. In our case study, actual users are present in virtual classrooms within a regional federated university system.

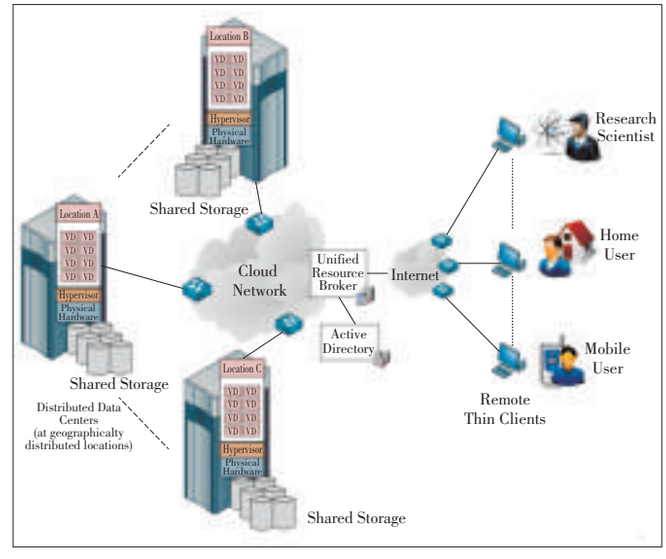
Keywords

virtual desktops; quality modeling and assessment; performance benchmarking; thin–client protocol adaptation; objective QoE metrics

1 Introduction

Motivations such as mobility, cost, security are behind the trend of traditional desktop users transitioning to virtual desktop clouds (VDCs) based on thin clients [1], [2]. With the increase in mobile devices with significant computing power and connections to high–speed wired and wireless networks, thin–client technologies for virtual desktop (VD) access are being integrated into these devices. In addition, users are increasingly consuming data–intensive content in scientific data analysis applications and multimedia streaming (e.g. IPTV). Thin clients are needed for these applications, which require sophisticated server–side computation platforms such as GPUs. Further, using a thin client may be more cost effective than using a full PC because a thin client requires less maintenance, and operating system, applications, and security upgrades are centrally managed at the server side.

Fig. 1 shows the various system components in a VDC. At the server side, a hypervisor framework, such as ESXi or Xen, is used to create pools of virtual machines (VMs) that host user VDs. These VDs have popular applications, such as Excel, Internet Explorer and Media Player, as well as more advanced applications, such as Matlab and Moldflow. Users of a common desktop pool access the same set of applications but maintain distinctive, personal datasets. The VDs at the server side have common physical hardware and attached storage drives. At the client side, users connect to a server–side unified resource broker via the Internet using various thin–client devices based on TCP (e.g. RDP) and UDP (e.g. PCoIP) The unified resource broker handles all connection requests by using Active Directory or other directory service lookups to authenticate users. It allows authorized users to access their VDs, and appropriate resources are allocated between distributed data centers.



▲Figure 1. Virtual desktop cloud system.

To allocate and manage VDC resources for large-SSsscale user workloads and maintain satisfactory QoE, VDC service providers (CSPs) need to suitably adapt cloud platform CPU, memory, and network resources. This also ensures that user-perceived interactive response times (timeliness) and streaming multimedia quality (coding efficiency) are satisfactory. CSPs need to ensure satisfactory user QoE when user workloads are bursty as a result of flash crowds or boot storms and when users access VDs from remote sites with varying end-to-end network path performance. CSPs need tools for VDC capacity planning in order to avoid overprovisioning system and network resources and to ensure QoE. CSPs also need frameworks and tools to benchmark VD application resource requirements so that adequate resources can be provisioned to meet user QoE expectations (e.g. less than 500 ms to open MS Office applications). When excess system and network resources are provisioned, a user does not perceive the benefit. For example, a user does not perceive any difference between an application open time of 250 ms and an application open time of 500 ms. Overprovisioning can become expensive, even at the scale of tens of users, given that each VD requires substantial resources (e.g. 1 GHz CPU, 2 GB RAM, and 2 Mbps end-to-end network bandwidth). Hence, CSPs need frameworks and tools that eliminate overprovisioning and result in benefits such as reduced data center costs and energy saving.

CSPs also need frameworks and tools to continuously monitor resource allocation and detect and troubleshoot QoE bottlenecks. To a large extent, CSPs can control CPU and memory resources and correctly provision them on the server side; however, frameworks and tools are critically needed to detect and troubleshoot network health issues on the Internet paths between the thin client and server-side VD. CSPs can use frameworks and tools to pertinently analyze network measurement data and adapt resources. Such resource adaptation involves selecting appropriate thin-client protocol configurations that are resilient to network health degradation and that provide optimum user QoE.

It is important to note that resources should be continually monitored without expensive, time-consuming subjective testing that involves actual VDC users. In terms of the thin client, remote display protocols are sensitive to network health and consume as much end-to-end network bandwidth as is available. They use different underlying TCP-based or UDP-based protocols, which have varying levels of QoE robustness when network conditions are poor [3]. Also, thin-client protocol configuration and associated VD application performance depends on the characteristics of the application content (i.e. characteristics of the text, images or video). Improper configuration can greatly affect user QoE in the form of lagging screen updates and poor keyboard and mouse responsiveness [4]–[7].

It is evident from the above CSP needs that frameworks and tools for capacity planning, thin-client protocol selection, and bottleneck troubleshooting have to be based on the principles of human-centric performance modeling and assessment so that users maximize their productivity and are highly satisfied. In this paper, we present a novel reference architecture that can be used to model and assess objective user QoE in VDCs. This architecture eliminates the need for expensive, time-consuming subjective testing. It involves offline benchmarking of VD application tasks, such as the time taken to open an Excel application or the time for a video to play back, in ideal network conditions. Performance degradation is modeled for different thin-client configurations in a broad range of deteriorated network conditions.

In our offline benchmarking methodology, we leverage finite-state machine representations to characterize the states of user workload tasks, and we use slow-motion benchmarking for deep-packet inspection of VD application task performance affected by QoS variations [4]. To define VD application task states and identify them in network traces during deep-packet inspection, we use marker packets that are instrumented in the traffic between the thin client and server-side VD ends. Our framework is implemented in the form of a VDBench benchmarking engine, which can be easily deployed in an existing VD hypervisor environment such as ESXi, Hyper-V, or Xen. VDBench can be used to instrument a wide variety of thin clients based on Windows and Linux platforms, such as embedded Windows 7, Windows/Linux VNC, Linux Thinstation, and Linux Rdesktop [8], [9]. The engine monitors VD user QoE by jointly analyzing the system, network, and application performance.

By using our offline benchmarking methodology in a closed-network testbed, we derive a novel composite-quality metric model of user QoE and show how the model can be customized for particular user-group profiles with different application sets. The model can be used during online monitoring to a) identify dominant performance indicators for numerous factors affecting user QoE and to troubleshoot bottlenecks, and b) obtain both absolute and relative objective user QoE measurements for pertinent selection/adaptation of thin-client encoding configurations. Absolute objective user QoE measurements allow the performance of a thin-client protocol (e.g. RDP) to be compared with that of another thin-client protocol (e.g. PCoIP) when there is latency and packet loss in the path between the thin client and server side. Relative objective user QoE measurements allow the performance of a thin-client protocol in degraded QoS conditions to be compared with the performance of the same protocol in ideal QoS conditions (where there is low latency and no loss).

We determine the effectiveness of our composite-quality modeling and assessment methodology by taking subjective and objective user QoE measurements in a real-world VDC in which RDP and PCoIP thin-client protocols are used. The actual end users are faculty and students in a virtual classroom lab within a federated university system. The high correlation between subjective and objective user QoE from this re-

al-work test allows us to determine the most suitable thin-client protocol for the use cases. It also allows us to verify that the configuration of the VDC infrastructure for the use cases had no inherent bottlenecks and delivered satisfactory user QoE.

The remainder of this paper is organized as follows: In section 2, we describe related work. In section 3, we present the reference architecture and its component functions and interactions, particularly user workload generation and slow-motion benchmarking. In section 4, we describe how our human-centric, composite-quality metric is formulated through closed network testbed experiments. In section 5, we validate the composite quality metric by using it in a real-world VDC. Section 6 concludes the paper.

2 Related Work

There are several works that outline the general architecture and requirements (e.g. isolation, scalability, dynamism, and privacy) of an end-to-end system-and-network-monitoring framework in cloud environments [10]-[13]. In most cases, the authors have proposed reusing traditional tools (e.g. active measurement tools, such as Ping, and passive measurement tools, such as Wireshark) as well as server-side methods to monitor QoS-related factors. The authors of [11] only instrument the server-side virtual machine with measurement-collection scripts.

In contrast, we emphasize human-centric quality assessment, and our method involves instrumenting both the thin clients and server-side VDs with measurement-collection scripts. These scripts feed measurements taken by traditional active and passive measurement tools into a benchmarking engine in order to correlate QoS and QoE measurements and to build a corresponding historical monitoring data set. Moreover, our method is very similar to that in [13]. Performance bottlenecks are detected during runtime or online, when QoE-related metrics exceed known benchmarks that are obtained through offline testing and analysis. Similar to the approach in [2], our approach involves the use of an historical monitoring dataset in the framework so that resources can be flexibly allocated; specifically, we improve user QoE and avoid overprovisioning resources.

The importance of human-centric quality-assessment frameworks for cloud environments has been highlighted in works such as [14]. The authors of [14] suggest that offline assessment and online monitoring should be based on profiles of user workloads and corresponding user QoE expectations. Our workload-generation methodology is based on the realistic emulation of user-group profiles, which are themselves based on application sets and corresponding tasks. We use the hierarchical-state-based workload-generation method described in [15]. In this method, the behavior of the actual user of the system represents the workload characteristics at a high level. This, in turn, results in a sequence of workload requests at the lower level that can be customized for a particular user-group profile and that are distinguishable (in our case, with marker packets) in network traces.

Along with workload generation, we use slow-motion benchmarking for network trace analysis in order to select suitable thin-client protocol configurations for particular user-group profiles. Our motivation for using slow-motion benchmarking for thin-client performance assessment is as follows: In advanced thin-client protocols (e.g. RDP and PCoIP), the server does all the compression and sends "screen scraping with multimedia redirection," which opens up a separate channel between the thin client and server-side VD. In this way, multimedia content can be sent in its original format to be rendered in the appropriate screen portion at the thin client. As a result, traditional packet-capture techniques using TCPdump or Wireshark do not directly allow VD application task performance to be measured. To overcome this challenge, we use and significantly expand on the slow-motion benchmarking technique originally developed in [4] and [5] for legacy thin-client protocols such as Sun Ray. In this technique, artificial delay is introduced between screen events of the tasks being benchmarked, and this allows isolation and full rendering of the visual components of the benchmarks. Alternative approaches have been taken in thin-client performance benchmarking toolkits [6], [7]. Such approaches involve recording and playing back keyboard and mouse events at the client side. However, in early thin-client benchmarking toolkits, server-side events and thin-client objective user QoE for degraded network conditions were not modeled and mapped in an integrated way (as proposed in our approach).

In earlier studies, different thin-client protocols, such as RDP and PCoIP, have been compared using metrics such as bandwidth/memory consumption in order to determine the suitability of these protocols for different application tasks [16]. Our implementation of the framework is based on our earlier work on the VDBench toolkit, which can be used for offline benchmarking of VDCs offline to create user group files based on system and network resource consumption [3]. Online monitoring was not taken into consideration for the VDBench toolkit. Compared with these works, this paper deals more with the framework architecture, framework implementation, and comparison and selection of thin-client protocols in VDCs (a process that is based on offline benchmarking and online monitoring). We propose a novel composite-quality metric function that maps QoS and QoE metrics related to any of the thin-client protocols, especially dominant metrics for profiles of VD application sets of user groups.

In other works, such as [17] and [18], neural networks and regression techniques were used to establish a relationship between QoS and QoE in specific multimedia delivery contexts. We use curve-fitting techniques to obtain absolute and relative objective user QoE, which are used together to compare

thin–client protocols in a broad range of degraded QoS conditions under different interactive operations. We validate these protocols by correlating them with subjective user QoE scores. The authors in [19] emphasize the need for new metrics to quantify user QoE. To date, metrics related to user QoE have not been precise, especially at the cloud scale.

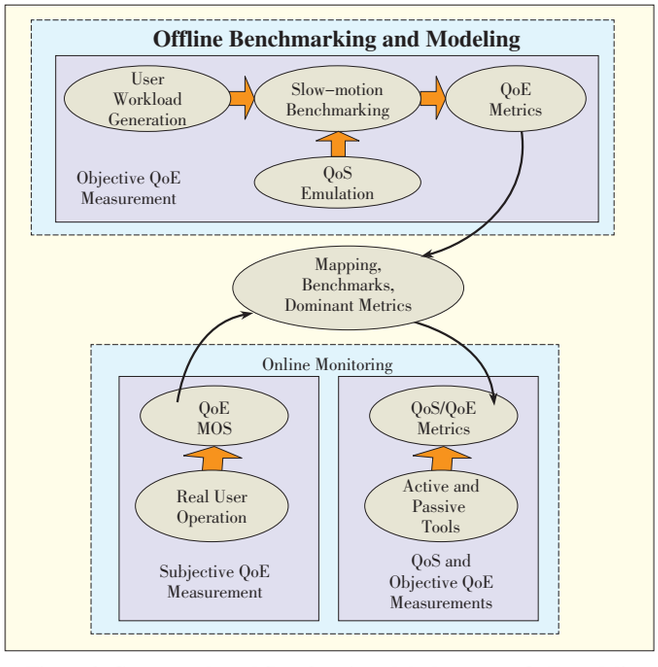
3 Reference Architecture

In this section, we first describe the conceptual workflow of our offline/online objective user QoE modeling and assessment steps for VDCs. Then, we describe in detail the user workload generation and slow–motion benchmarking approaches used within these steps.

3.1 Conceptual Workflow

Fig. 2 shows the two main steps in our VD user QoE modeling and assessment: 1) offline benchmarking and modeling and 2) online monitoring. In the offline step, we collect a set of benchmarks by generating user workflow and performing slow–motion benchmarking under different emulated QoS conditions where the VDC configurations are controlled for testing. QoS emulation involves varying the delay and loss levels between the thin client and server–side VD over a broad sample range. The benchmarks are collected by averaging the results from multiple experimental runs in the form of objective QoE metrics that correspond to VD application tasks (e.g. time taken to open an Excel application, time for a video file to play, or quality of a video file playback).

The benchmark data collected for a particular user group profile with a certain application set allows a CSP to determine the VD application performance in ideal conditions. The benchmark data also provides a model for mapping how performance may degrade for different thin–client configurations in deteriorated network conditions, which are measured using common active or passive measurement tools. By analyzing VD application performance in degraded conditions, metrics that are most sensitive to network fluctuations can be identified as dominant metrics. Dominant metrics are key performance indicators, as in the case of online monitoring. They can be relied on to confirm that the VDC is functioning well and also in some cases, to diagnose the cause of unsatisfactory user QoE feedback given by actual users while using the VDC. Subjective user QoE is measured using the popular mean opinion score on a scale from one to five [20]: [1, 3) is poor; [3, 4) is acceptable; and [4, 5] is good. Hence, models can be obtained (as closed–form expressions) for objective QoE metrics. These models are composite quality functions of dominant metrics, and higher weights are assigned to more dominant metrics. Objective QoE metrics can, in turn, be used to correlate performance with actual user QoE mean opinion scores.



▲Figure 2. Conceptual workflow for virtual desktop user QoE modeling and assessment.

3.2 User Workload Generation

Fig. 3 shows a finite–state machine used in user workload generation. It has various VD application states that correspond to actions performed by the user: productivity start (application is opened), productivity progress (application functions are in use), and productivity end (application is closed). The VD application states can be scripted using Windows GUI frameworks such as AutoIT, and these scripts can be launched on VDs within the VDC [21]. The progress of these states on the server side can be measured by recording and analyzing timestamps for different actions. This can be useful for understanding the interaction response times of VD applications perceived at the thin–client side. Controllable delays caused by user behaviors, such as think time, can be introduced for more

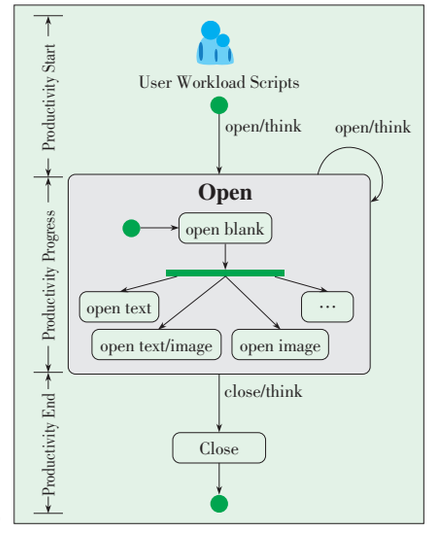


Figure 3. ▶ Finite–state machine for generating user workload.

realistic user workload generation.

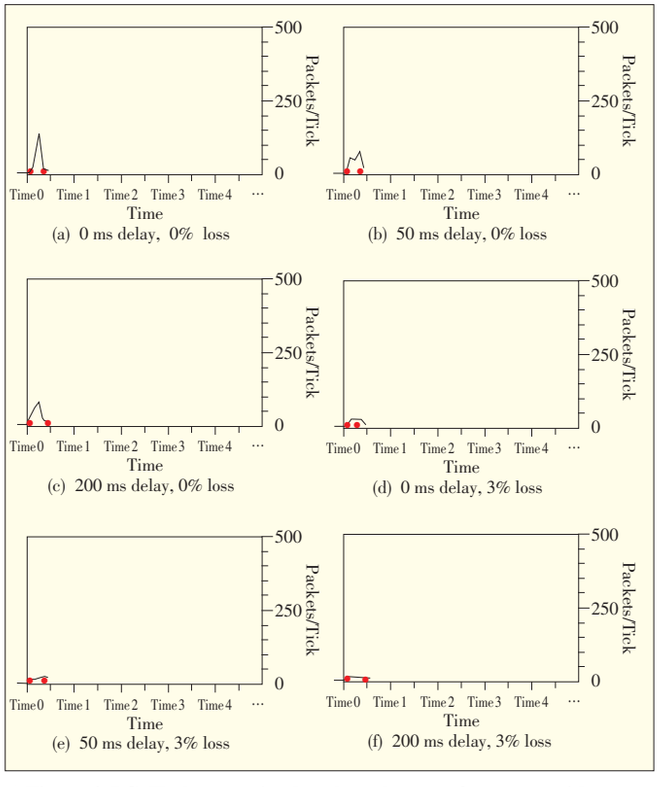
Workload templates can be created by combining a series of individual finite-state machines into a parent-state machine. These templates are used to orchestrate different VD applications, such as Microsoft Excel, Internet Explorer, and Windows Media Player, in random sequences. Depending on the workload performance measurement needs on the VDs, the state machines can be modified and redistributed on the VDs by using appropriate "marker" packets. The marker packets are needed to identify the different tasks within the network traces and are sent from the workload scripts between the server-side VD and the thin client. The marker packets are sent for subsequent filtering (as part of post-processing) via ports that are not standard in protocol communications. Marker packets contain information that describes the application being tested (e.g. Internet Explorer), the activity (e.g. type of webpage being downloaded), the event start/stop timestamps, and other metadata (e.g. network conditions emulated in the test).

3.3 Slow-Motion Benchmarking

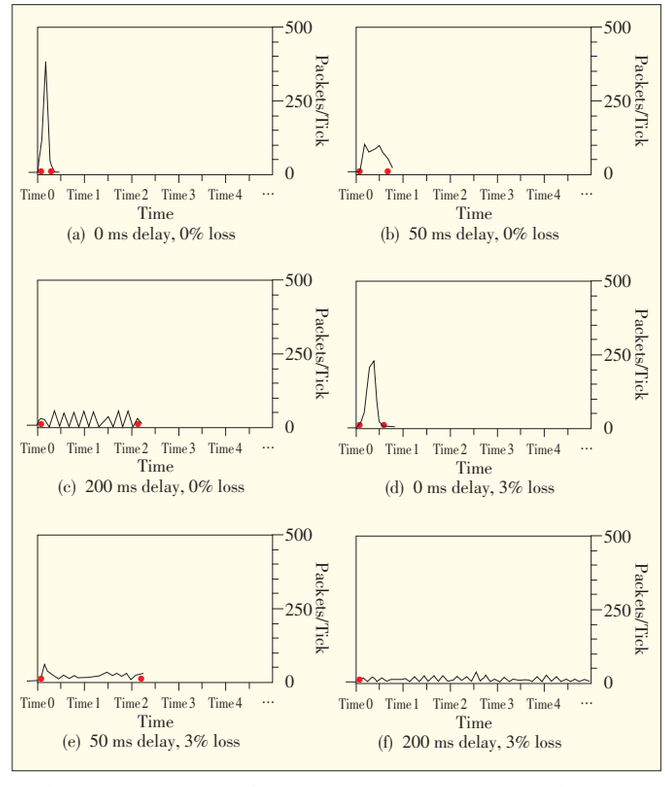
Figs. 4 and 5 show the slow-motion benchmarking packet traces with marker packets (red dots) for PCoIP and RDP, respectively. Wireshark is used to view the packet traces and marker packets for a page with low-resolution images that is loaded in Internet Explorer. Through deep-packet inspection, we observe that artificial delays are introduced between screen events of VD application tasks, mainly to ensure that the visual components of the tasks are isolated and fully rendered on the thin client. The network packet traces are analyzed both at the client and server sides to measure VD task performance in ideal and degraded network conditions. Figs. 4 and 5 show how VD task performance degrades at the thin-client side across a broad range of network conditions.

For the systematic emulation of network conditions, we use a Netem network emulator to introduce combinations of 0 ms, 50 ms, and 200 ms delay and 0% and 3% losses between the client side and server side [22]. These combinations are typical for well-managed, moderately well-managed, and poorly managed network paths on the Internet and are used hereafter as a representative sample space of network conditions to describe our proposed composite-quality modeling approach. They also are sufficient for showing the best-case and worst-case performance of the VD application tasks. More detailed samples of network conditions may be considered in order to obtain finer-grained composite-quality models. However, such data collection and modeling is beyond the scope of this paper.

Performance differences are apparent in the crispness of network utilization patterns and in the higher bandwidth consumption (indicated by the packet count along the y axis) and lower task times in ideal conditions. In degraded conditions, bandwidth consumption is lower and task times are longer. In addi-



▲Figure 4. PCoIP slow-motion benchmark traces for a page with low-resolution images loaded by Internet Explorer.



▲Figure 5. RDP slow-motion benchmark traces for a page with low-resolution images loaded by Internet Explorer.

tion, Figs. 4 and 5 show how the popular PCoIP and RDP thin–client protocols handle the degraded conditions while completing a particular application task (e.g. loading a page with a low–resolution image in Internet Explorer) and applying protocol–specific compensations. The compensations manifest in the transmission of the visual components of the VD applications to the client side and affect remote user consumption and productivity. Ultimately, the compensations affect user QoE.

PCoIP provides the same satisfactory user QoE as RDP in ideal conditions but consumes less bandwidth. Compared with RDP, PCoIP has a tighter rendering time, even in the worst cases (i.e. delay 200 ms and loss 3%), in our sample space. If the values for delay and loss are greater than those we chose for the worst–case network degradation, user QoE is always poor (as is evident in the slow–motion benchmarking traces). Hence, higher sample values were not used in our setup. This difference is due to the fact that PCoIP uses UDP as the underlying transmission protocol, whereas RDP is based on TCP and retransmits when network conditions are lossy.

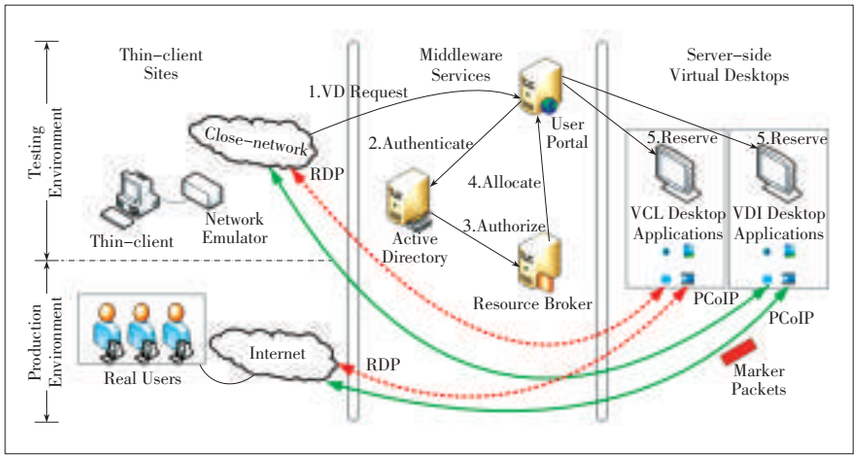
4 Composite Quality Formulation

In this section, we describe the closed–network setup and experiments used to formulate the composite quality functions for PCoIP and RDP thin–client protocols.

4.1 Closed–Network Testbed

Fig. 6 shows the reference architecture described in section 3 implemented as a VDBench benchmarking engine in a VDC at VMLab [23]. The physical components are set up and interactions are organized in three layers: thin–client sites, middleware services, and server–side VDs. At the thin–client sites, an offline, closed–network testing environment is used to formulate the composite quality functions. In the validation experiments described in section 3, the same infrastructure is used for the middleware services and server side; however, they connect via the Internet from geographically distributed locations. We set up two VD environments, one with the open–source Apache VCL [24] and the other with VMware VDI [25]. The default thin–client protocols for VD access in the Apache VCL environment was RDP, and the default thin–client protocol for VD access in the VMware environment was PCoIP.

The VDBench benchmarking engine automates and orchestrates 1) initial workflow steps 1 to 5 shown in Fig. 6, 2) the use of RDP or PCoIP to set up a thin–client session, and 3) the final objective user QoE measurements obtained by integrating user workload generation and slow–motion benchmarking in various network conditions. Our implementation can be easily deployed in any existing VD hypervisor environment (e.g. ESXi, Hyper–V, and Xen) and relies on APIs to authenticate and authorize VD requests and allocate and reserve resources. Utilities such as psexec are used to remotely invoke workload–generation scripts; hence, our implementation can also be used to instrument existing thin–clients based on Windows or Linux platforms, such as embedded Windows 7,Windows/Linux VNC, Linus Thinstation, and Linux Rdesktop) [8], [9]. Moreover, our implementation allows for the storing of performance measurements and for the joint analysis of system, network, and application performance in order to measure and monitor VD user QoE online.



▲Figure 6. User QoE modeling and assessment framework: Physical component setup and interactions.

4.2 Closed–Form Expressions

Table 1 shows the objective QoE metrics collected by the VDBench benchmarking engine. Although we collected and analyzed more than 20 metrics, we only chose seven (M_1 to M_7, Table 1) that were observed to significantly affect user QoE as network conditions degraded for both the RDP and PCoIP protocols. We ignored the metrics of tasks whose packet count variations over task times were not significant in the best and worst network conditions. Such metrics are not helpful in identifying QoE bottleneck scenarios in VD applications.

Metrics M_1 to M_6 are obtained by subtracting the timestamps between the marker packets and comparing them with the ideal QoS values. However, M_7 is calculated differently: A video is first played back at 1 frame per second (fps) in an atomic manner, and network trace statistics are captured. The video is then replayed at full speed a number of times in an aggregate manner using the thin–client protocol being tested and in various network conditions. A challenge in comparing the performance of UDP–based and TCP–based thin–client protocols in terms of video quality is to derive a normalized metric. The normalized metric should account for fast completion times with image impairments in UDP–based thin–client protocols (as opposed to long completion times in TCP–based thin–clients with no im-

▼Table 1. Notation

Notation	Alternative Notation	Definition
T^{excel}_{open}	M_1	Excel open time is the time in seconds taken for Excel application to open
T^{excel}_{render}	M_2	Excel render time is the time in seconds taken for Excel application to render sample text
$T^{low\_img}_{load}$	M_3	Low-resolution image load time is the time in seconds taken for Internet Explorer to load a webpage with a sample low-resolution image of the US Constitution
$T^{high\_img}_{load}$	M_4	High-resolution image load time is the time in seconds taken for Internet Explorer to load a webpage with a sample high-resolution image of the US Constitution
T^{text}_{load}	M_5	Text load time is the time in seconds taken for Internet Explorer to load a webpage with sample text
T^{video}_{play}	M_6	Media playback time is the time in seconds taken for Windows Media Player to play back a sample video
Q^{video}_{play}	M_7	Media playback quality is the playback quality of the sample video in Windows Media Player
CQS		The composite-quality score is given by (2).
RCQS		The relative composite-quality score is the normalized CQS obtained from (3).
ACQS		The absolute composite-quality score is the normalized CQS obtained from (4).
AMOS		The absolute mean opinion score is the average score given by real users during subjective QoE testing.

pairments but long frame freezes). To meet this challenge, we use the video quality metric given by (1) and originally developed in [5]. The M_7 metric relates slow-motion atomic playback to full-speed aggregate playback in order to determine how many frames are dropped, merged, or otherwise not transmitted.

$$Q^{video}_{play} = \frac{\left(\frac{(data\ transferred\ (aggregate\ fps)/render\ time\ (aggregate\ fps)}{ideal\ transfer\ (aggregate\ fps)}\right)}{\left(\frac{(data\ transferred\ (atomic\ fps)/render\ time\ (atomic\ fps)}{ideal\ transfer\ (atomic\ fps)}\right)} \quad (1)$$

Of these seven metrics, those that were most sensitive to change in network conditions are the more dominant metrics that affect user QoE. For other application sets and user expectations of VD application performance, the number of metrics may vary. In any case, the composite quality function used to predict user QoE given n metrics $\{M_1, M_2, \ldots, M_n\}$ can be calculated using a general form, $W_1M_1 + W_2M_2 + \ldots + W_nM_n$, where each metric M_i for $i \in \{1, n\}$ has a corresponding weight W_i for $i \in \{1, n\}$. Each corresponding weight is based on how sensitive the metric is to network degradation or how much of a key performance indicator the metric is. We assign a weight by calculating the change in the measurement value of the metric per unit change in QoS (i.e. the change in delay and loss). Table 2 shows the normalized weights of the dominant metrics derived for PCoIP and RDP protocols from the packet traces in our closed-network testbed experiments that involved the systematic emulation of network conditions described in section 3.3. By combining these weights and metrics, we derive a closed-form expression (2) that can be used to predict the composite-quality function for our setup:

$$CQS = \sqrt{W_7M_7} - (W_1M_1 + W_2M_2 + W_3M_3 + W_4M_4 + W_5M_5 + \frac{W_6M_6}{5}) \quad (2)$$

Equation (2) is derived through trial and error with increasing complexity in relation to (1) in order to obtain closed-form expressions that best fit the training data curve. Because the dominant metrics in our case were the same for both RDP and PCoIP, the same equation can be used to estimate CQS for RDP and PCoIP. Note that M_7 was found to be a highly dominant metric compared to the rest of the metrics; hence, we used the square root of this metric to balance its influence in relation to M_1 to M_6 .

To compare the thin-client protocols in relative and absolute terms, we give RCQS and ACQS as

$$RCQS = \frac{(CQS_{curr} - CQS_{200,3})}{(CQS_{0,0} - CQS_{200,3})} \times 100 \quad (3)$$

$$ACQS = \frac{CQS_{curr} - \min\ (CQS^{RDP}_{200,3}, CQS^{PCoIP}_{200,3})}{\max(CQS^{RDP}_{0,0}, CQS^{PCoIP}_{0,0}) - CQS_{200,3}} \times 100 \quad (4)$$

RCQS and ACQS are normalized between 0 and 1 and are given as percentages. They are calculated using the CQS for the current, ideal, and worst network conditions for RDP and PCoIP thin-client protocols. In our experiments, CQS_{curr} , the composite-quality score being calculated for a given network condition, is compared with the CQS for ideal (0 ms, 0%) and worst-case (200 ms, 3%) delay and loss. This allows us to compare how well user QoE performance conforms to ideal performance (the relative objective QoE score) for a thin-client protocol being tested. Moreover, it allows us to compare the thin-client protocols that may have different weights applied to the same metrics in degraded conditions (the absolute objective QoE score). Tables 3 and 4 respectively show RDP and PCoIP composite-quality calculations in relation to the RCQS and ACQS values obtained in our closed-network testbed. RDP and PCoIP protocols perform differently in different application contexts and network conditions.

5 Validation of Results

In this section, we validate our composite-quality modeling

▼Table 2. Normalized weights showing dominant metrics

Metric	RDP Weight	PCoIP Weight
T^{excel}_{open}	0.08	0.06
T^{excel}_{render}	0.03	0.01
$T^{low\_img}_{load}$	0.13	0.26
$T^{high\_img}_{load}$	0.14	0.22
T^{text}_{load}	0.08	0.07
T^{video}_{play}	0.30	0.09
Q^{video}_{play}	0.21	0.26

▼Table 3. RDP composite quality calculations in closed–network testbed

Delay (ms)	Loss (%)	T_{open}^{excel}	T_{render}^{excel}	$T_{load}^{low\_img}$	$T_{load}^{high\_img}$	T_{load}^{text}	T_{play}^{video}	Q_{play}^{video}	CQS	RCQS	ACQS
0	0	1.12	20.63	0.32	0.52	0.49	7.00	7.66	–0.13	100.00	89.67
50	0	0.94	20.85	0.98	1.13	0.51	50.65	1.37	–3.72	79.48	71.38
200	0	1.09	20.70	0.42	0.41	0.95	220.65	0.26	–14.33	18.78	17.07
0	3	1.21	20.64	0.46	1.04	0.50	6.65	6.07	–0.35	98.74	88.55
50	3	1.13	20.81	1.30	1.76	0.43	62.54	0.89	–4.70	73.90	66.50
200	3	1.40	20.58	0.42	0.68	0.41	273.58	0.21	–17.61	0.00	0.00

▼Table 4. PCoIP composite quality calculations in a closed–network testbed

Delay (ms)	Loss (%)	T_{open}^{excel}	T_{render}^{excel}	$T_{load}^{low\_img}$	$T_{load}^{high\_img}$	T_{load}^{text}	T_{play}^{video}	Q_{play}^{video}	CQS	RCQS	ACQS
0	0	0.91	20.51	0.18	0.34	0.26	5.98	22.78	1.86	100.00	100.00
50	0	0.88	20.56	0.23	0.46	0.29	8.80	11.45	1.05	64.47	95.84
200	0	0.96	20.46	0.39	0.71	0.39	9.91	10.49	0.85	55.45	94.47
0	3	0.77	20.52	0.17	0.36	0.22	6.00	0.80	–0.12	12.62	89.94
50	3	0.91	20.60	0.24	0.43	0.27	8.45	0.69	–0.25	7.02	89.16
200	3	0.90	20.50	0.35	0.77	0.51	6.03	0.45	–0.41	0.00	88.34

and assessment methodology in a real–world VDC. We describe how the validation testbed was set up for user trials. Following this, we give the results of our analysis of subjective and objective user QoE measurements.

5.1 VDPilot Testbed

We used the production environment shown in Fig. 6, which is similar to the testing environment (without the network emulator), as a VDPilot testbed for our validation experiments. The testbed uses RDP and PCoIP thin–client protocols, and actual users are present in virtual classrooms within a federated university system. A total of 36 users registered in the VDPilot, most of whom were faculty and students from diverse selection of Ohio–based universities. These universities included Ohio State University, University of Dayton, University of Akron, Ohio University, Denison University, Walsh University, Sinclair University, Ashland University, and Baldwin Wallace College.

As part of subjective testing in the VDPilot testbed, participants were asked to compare Apache VCL (configured with default RDP) with VMware VDI (configured with default PCoIP) remote thin clients while using Excel, Windows Media Player, and Internet Explorer to complete tasks in VDs. After completing the subjective tests, participants were asked to complete an online survey to provide feedback about their perception of QoE while accessing VD applications with Apache VCL and VMware VDI.

Subsequently, the participants were asked to perform objective QoE tests to more objectively compare user QoE at the different sites and eliminate any outlier biases or mood effects of the participants that may have affected their subjective judgment of VD application QoE. As part of the objective QoE testing, participants downloaded, installed, and ran the OSC/OARnet VDBench software shown in Fig. 7 for both Apache VCL and VMware VDI remote thin clients. The installation prerequisites included the latest Java runtime environment, Wireshark, and two clients in the form of JAR files (one for Apache VCL and the other for VMware VDI). The VDBench client software implements the client–side aspects of our user workload generation and slow–motion benchmarking methodologies explained in sections 3.2 and 3.3. It can run on both Windows and Linux platforms and is capable of NIC selection to initiate tests. It interacts with the VDBench benchmarking engine at the server side through messages encoded in marker packets. It can also be securely used in VDCs because it requires a participant to input a username and password that is valid in the Active Directory on the server side.

The VDBench client executes a series of automated tests over several minutes. These tests are performed in the participant's remote thin client in order to simulate or mimic the actions performed by participants during subjective testing over the Internet. While performing the tests within an Apache VCL or VMware VDI instance, the software records interactive application response times and video playback quality metrics (M_1 to M_7, Table 1). This quantitative performance information can be used to identify bottlenecks and can be correlated with subjective user QoE ratings, that is, the mean opinion scores of participants. Measurements taken during a test run in an instance of either Apache VCL or VMware VDI are displayed to the user in the VDBench client user inter-

▲Figure 7. Java VDBench client.

face, and a copy of these measurements is automatically stored in a database on the server side.

5.2 Correlating Subjective QoE Measurements

Table 5 shows the subjective and objective user QoE results from testing in the VDPilot testbed. Absolute mean opinion scores are the averages of the user QoE feedback. The absolute mean opinion scores for both RDP and PCoIP were greater than four (i.e. PCoIP was 4.74 and RDP was 4.21), which is in the "good" range of user QoE. In addition, the ACQSs were high and correlated with the baseline results from the closed–network testbed for a network with approximately 50 ms delay and 0% packet loss. Such correlation was expected given that the participants were dispersed over a regional wide–area network. Thus, we were able to determine that the VDC infrastructure configuration used in the VDPilot case had no inherent usability bottlenecks and could provide satisfactory user QoE for Ohio–based universities.

▼Table 5. Correlation results for subjective and objective QoE measurements

Protocol	T_{open}^{excel}	T_{render}^{excel}	$T_{load}^{low\_img}$	$T_{load}^{high\_img}$	T_{load}^{text}	T_{play}^{video}	Q_{play}^{video}	CQS	ACQS	AMOS
RDP	7.66	23.79	0.71	1.01	1.12	17.54	5.50	–1.81	81.11	4.21
PCoIP	2.19	20.79	0.22	0.41	0.37	7.65	11.46	0.99	95.76	4.74

We also observed that the AMOSs closely correlated with the objective user QoE measurements given by ACQS and calculated using (4) for both the RDP and PCoIP protocols. Further, we can conclude that the PCoIP thin–client protocol is more suitable than the thin–client protocol for the virtual classroom lab use cases. The PCoIP ACQS was 95.76, and the RDP ACQS was 81.11. ACQS is a more relevant objective QoE metric than RCQS for making a comparison with AMOS because we are comparing the performance of two different thin–client protocols.

6 Conclusion

In this paper, we have presented a novel, human–centric reference architecture and described how it can be used to model and assess objective user QoE in VDCs without the need for expensive, time–consuming subjective testing. The architecture incorporates finite–state machine representations for user workload generation and also incorporates slow–motion benchmarking, in which deep packet inspection is used to determine the performance application tasks affected by QoS variations. In this way, a composite–quality metric model of user QoE can be derived. We have shown how this metric can be customized to a particular user–group profile with different application sets and can be used to a) identify dominant performance indicators and troubleshoot bottlenecks and b) obtain both absolute and relative objective user QoE measurements needed for pertinent selection of thin–client encoding configurations in VDCs.

Our framework and its implementation in the form of a VDBench benchmarking engine on the server side and Java–based VDBench client on the thin–client side can be used by CSPs within existing VD hypervsior environments (e.g. ESXi, Hyper–V and Xen) and can be extended to instrument a wide variety of existing thin clients based on Windows and Linux platforms (e.g. embedded Windows 7, Windows/Linux VNC, Linux Thinstation, and Linux Rdesktop). CSPs can use our framework and implementation to monitor VD user QoE by jointly analyzing system, network, and application contexts. This ensures that a CSP's VD users are satisfied, VD applications are highly productive, and VDC support costs are reduced because performance is more transparent.

We validated our composite–quality modeling and assessment methodology by using subjective and objective user QoE measurements in a real–world VDC called VDPilot, uses RDP and PCoIP thin–client protocols. In our case study, actual users were present in virtual classrooms in a regional, federated university system. There was high correlation between the subjective and objective user QoE results from testing, and this allowed us to determine that PCoIP was the more suitable thin–client protocol for the virtual classroom case. We also determined that the configuration of the VDC infrastructure for this case had no inherent bottlenecks and could provide satisfactory user QoE over the Internet at a regional level.

Acknowledgements

This material is based on work supported by VMware and the National Science Foundation under award numbers CNS–1050225 and CNS–1205658. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of VMware or the National Science Foundation.

References

[1] P. Calyam, R. Patali, A. Berryman, A. Lai, and R. Ramnath, "Utility– directed resource allocation in virtual desktop clouds," *The International Journal of Computer and Telecommunications Networking*, vol. 55, no. 18, pp. 4112–4130, Dec. 2011.
[2] L. Deboosere, B. Vankeirsbilck, P.Simoens, F. De Turck, B. Dhoedt, and P. Demeester, "Cloud–based desktop services for thin clients," *IEEE Internet Computing*, vol. 16, no. 6, pp. 60–67, Nov. – Dec. 2012.
[3] A. Berryman, P. Calyam, M. Honigford, and A. Lai, "VDBench: A benchmarking toolkit for thin–client based virtual desktop environments," *Proc. of IEEE CloudCom*, Indianapolis, IN, 2010, pp. 480–487.
[4] J. Nieh, S. Yang, and N. Novik, "Measuring thin–client performance using slow–motion benchmarking," *ACM Transactions on Computer Systems*, vol. 21, no. 1, pp. 87–115, 2003.
[5]A. Lai and J. Nieh, "On the performance of wide–area thin–client computing," *ACM Transactions on Computer Systems*, vol. 24, no. 2, pp. 175–209, 2006.
[6] J. Rhee, A. Kochut, and K. Beaty, "DeskBench: flexible virtual desktop benchmarking toolkit," *Proc. of Integrated Management (IM)*, Long Island, NY, 2009, pp. 622–629.

[7] N. Zeldovich and R. Chandra, "Interactive performance measurement with VNCplay," *Proc. of USENIX Annual Technical Conference*, Anaheim, CA, 2005, pp. 189–198.
[8] *Thinstation: Open-Source Thin-Client Operating System* [Online]. Available: http://www.thinstation.org
[9] *Wyse Thin Client Products*[Online]. Available: http://www.wyse.com
[10] P. Hasselmeyer and N. d'Heureuse, "Towards holistic multi-tenant monitoring for virtual data centers," *Proc. of IEEE/IFIP NOMS Workshop*, Osaka, Apr. 2010, pp. 350–356.
[11] S. De Chaves, R. Uriarte, and C. Westphall, "Toward an architecture for monitoring private clouds," *IEEE Communications Magazine*, vol. 49, no. 12, pp. 130 – 137, 2011.
[12] S. Clayman, A. Galis, C. Chapman, et. al., "Monitoring service clouds in the future Internet," *Towards the Future Internet – Emerging Trends from European Research*, IOS Press ISBN 978-1-60750-538-9, 2010.
[13] V. Emeakaroha, M. Netto, et. al., "Towards autonomic detection of SLA violations in cloud infrastructures," *Future Generation Computer Systems*, vol. 28, no. 7, pp. 1017–1029, 2012.
[14] J. Shao, H. Wei, Q. Wang, and H. Mei, "A runtime model based monitoring approach for cloud," *Proc. of IEEE Conference on Cloud Computing (CLOUD)*", Miami, FL, Jul. 2010, pp. 313–320.
[15] H. Hlavacs and G. Kotsis, "Modeling user behavior: a layered approach," *Proc. of IEEE MASCOTS*, College Park, MD, 1999, pp. 218–225.
[16] J. Kouril and P. Lambertova, "Performance analysis and comparison of virtualization protocols, RDP and PCoIP," *Proc. of International Conference on Computers (ICCOMP)*, 2010, vol. 2, pp. 782–787.
[17] S. Mohamed and G. Rubino, "A study of real-time packet video quality using random neural networks," *IEEE Transaction on Circuits and Systems for Video Technology*, vol. 12, no. 12, pp. 1071–1083, 2002.
[18] M. Fiedler, T. Hossfeld, and T. Phuoc, "A generic quantitative relationship between quality of experience and quality of service," *IEEE Network*, vol. 24, no. 2, pp. 36–41, 2010.
[19] L. Spracklen, B. Agrawal, R. Bidarkar, and H. Sivaraman, "Comprehensive user experience monitoring," *VMware Technical Journal*, Mar. 2012.
[20] M. Pinson, S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Transactions on Broadcasting*, ISSN: 0018-9316, Vol. 50, Pages 312–22, 2004.
[21] *AutoIT Windows GUI Scripting Framework* [Online]. Available: http://www.autoitscript.com
[22] *The Linux Foundation Netem* [Online]. Available: http://www.linuxfoundation.org/collaborate/workgroups/networking/netem
[23] P. Calyam, A. Berryman, A. Lai, and M. Honigford, "VMLab: Infrastructure to Support Desktop Virtualization Experiments for Research and Education", *VMware Technical Journal*, Dec. 2012.
[24] M. Vouk, A. Rindos, S. Averitt, et. al., "Using VCL technology to implement distributed reconfigurable data centers and computational services for educational institutions", *IBM Journal of Research and Development*, vol. 53, no. 4, pp. 509–526, 2009.
[25] *VMware Virtual Desktop Infrastructure and VMware View* [Online]. Available: http://www.vmware.com

Manuscript received: January 23, 2013

Biographies

Yingxiao Xu

Yingxiao Xu (xuyx@fudan.edu.cn) received his BS and MS degrees in mechanical engineering from Southeast University, China, in 1993 and 1996. He received his PhD degree in computer science and engineering from Fudan University, China, in 2002. He is currently a lecturer at Fudan University. From 2010 to 2011, he was a visiting scholar at the Ohio Supercomputer Center/OARnet, The Ohio State University. His research interests include software reuse, social computing, and computer and network management.

Prasad Calyam

Prasad Calyam (calyamp@missouri.edu) received his BS degree in electrical and electronics engineering from Bangalore University, India, in 1999. He received his MS and PhD degrees in electrical and computer engineering from The Ohio State University, in 2002 and 2007. He is currently an assistant professor at the University of Missouri-Columbia. His research interests include distributed and cloud computing, computer networking, networked-multimedia applications and cyber security.

David Welling

David Welling (dwelling@osc.edu) is pursuing a BS degree in computer science and engineering at The Ohio State University. His research interests include web architectures, reconfigurable software engineering, and systems performance assessment.

Saravanan Mohan

Saravanan Mohan (smohan@osc.edu) received his BE degree in computer science from PSG College of Technology, India, in 2008. He is currently pursuing his MS degree in computer science and engineering at The Ohio State University. His research interests include cloud monitoring and desktop virtualization.

Alex Berryman

Alex Berryman (aberryman@osc.edu) is currently pursuing a BS degree in aeronautical and astronautical engineering at The Ohio State University. His research interests include network monitoring, desktop virtualization, and cyber security.

Rajiv Ramnath

Rajiv Ramnath (ramnath@cse.ohio-state.edu) is director of practice at the Collaborative for Enterprise Transformation and Innovation (CETI). He is also associate director of the Institute of Sensing Systems and associate professor of practice in the Department of Computer Science and Engineering, The Ohio State University. His research interests include wireless sensor networking, pervasive computing, enterprise architecture, software engineering, and work-management systems. He received his MS and PhD degrees in computer science from The Ohio State University in 1983 and 1989. He received his BS degree in electrical engineering from the Indian Institute of Technology, New Delhi, in 1981.

Assessing the Quality of User–Generated Content

Stefan Winkler
(Advanced Digital Sciences Center (ADSC), Singapore 138632)

Abstract

With the widespread use of digital cameras, imaging software, photo–sharing sites, social networks, and other related technologies, media production and consumption patterns have become much more multifaceted and complex than they used to be. User–generated content in particular has grown tremendously. As a result, quality of experience (QoE) and related quality assessment (QA) methods must also be looked at from a different angle. This paper contrasts some of the traditional quality assessment approaches with newer approaches designed for user–generated content. It also describes some sample applications we have developed.

Keywords

image quality; photo collections; social media; photowork; summarization

1 Introduction

Quality assessment (QA) for images dates back to the 1970s, when the first studies were done on the visual cortex, vision modeling, and digital imaging. Algorithms based on models of the human visual system now compete with more pragmatic, image– or feature–based methods [1]. Video QA has a similar, albeit shorter, history [2].

Ubiquitous and affordable digital cameras (also in the form of embedded image capture devices) now enable users to take pictures and videos almost anywhere, anytime. This has led to an explosion in the amount of picture material produced by both amateurs and professionals.

However, traditional media and user–generated content are fundamentally different on many levels, especially from a quality assessment perspective (Table 1).

1.1 QA for Traditional Media

Traditional QA methods have focused mainly on the processing and distribution chains of broadcast media, that is, the compression, transmission, and enhancement of images and video. In many cases, there is an explicit reference (e.g. the source image or video), that passes through the system and undergoes certain changes (e.g. loss of fidelity, compression artifacts, packet loss, noise removal).

For such high–value, professional content, QA is typically done manually during the production process. Afterwards, the content is prepared for and distributed to many paying consumers via channels such as cinema and broadcast TV. It is very much a linear process in which a single high–quality source passes through various processing steps that may change or affect the quality of the content. The entities concerned with quality throughout this process are typically encoder manufacturers, content providers, service providers, and operators. Furthermore, traditional media is designed for a wide audience, and as a consequence, the average user and mean opinion score (MOS) are the quality unit and benchmark of choice.

Most QA algorithms [1], [3], [4]; databases [5]; standards [6]; and products have so far have focused on this stage, where fidelity (i.e. how closely the processed image/video resembles the original source content) is of primary importance.

1.2 QA for User–Generated Content

With user–generated content, fidelity is secondary. The criteria for QA and enhancement are not only image– or content–specific (e.g. impairments or scene composition) but also user–centric (i.e. what is most relevant to the user in a collection). This is contrary to traditional QA approaches.

Automated QA for user–generated content is useful primarily in the production process for a number of reasons:

- Even if QA could be done manually by the user, it would be too time–consuming for most. Besides, the average user needs guidance to produce good–quality content.
- Processing and distribution are simple and are largely hidden from or opaque to the user (e.g. compression in the camera or uploading content to a website).
- Quality becomes a much more personal concept because it is mainly the user and the circle of people they may share

▼Table 1. Traditional media vs. user–generated content

Stages	Traditional Media	User–generated Content
Production	Professional quality, premium content	Amateur content/quality
Processing	Encoding, transcoding, multiplexing, etc.	Minimal user intervention, or hidden from user
Distribution	Real–time streaming, many users, high network demands	Sharing with friends, typically downloads

the content with who matter the most. Indeed, personalization has not received much attention so far despite its importance for user-generated content.

We discuss these aspects in more detail, using the example of photo collections. However, they also apply to other types of media.

2 Photo Collections

The most common type of user-generated content today is digital photos. Collections typically comprise pictures taken during an event or trip, possibly by multiple users using different devices. They may also comprise images shared in social networks, images on websites, and images stored in third-party repositories. Devices may include single-lens reflex (SLR) cameras, point-and-shoot cameras, and camera phones.

It has become so easy to take lots of pictures that users regularly have to deal with large photo collections. The role of QA here is primarily the selection of the best and most representative pictures from a collection. This task can be broken down into two basic steps: screening and summarization. In screening, the best photo is selected from a group of similar photos (typically multiple shots of the same scene) and enhancements are applied if necessary. In summarization, a subset of pictures is chosen for an album. Often, the purpose is to tell a story or share an experience.

Intelligent user interface design and personalization are essential in both steps because of the importance of user-specific criteria, tastes, and preferences. It is also difficult to fully automate these processes to the satisfaction of users.

2.1 Screening

With digital cameras, it has become common practice to take multiple pictures of the same scene. Users typically take two or three shots per scene on average, and for certain scenes and situations, between eight and ten shots [7]. Selecting the best picture from a group of pictures typically involves evaluating lighting, exposure, and white balance; framing and perspective; postures, actions and faces of people in the scene; and basic image quality [8].

Typically, the quality of a picture is assessed by comparing it to a reference with the same content but without impairments (full-reference comparison). It may also be assessed on its own (no-reference comparison). When there is no reference image, traditional full-reference methods do not apply. No-reference methods can be used in principle, but they generally work best along a single impairment dimension, for example, quantization or blur, of the same image.

The problem here is more general and revolves around comparing pictures that have similar, related (but not identical) content and different quality/impairment dimensions and levels (one image may be blurred while another may be underexposed). We need to choose the best of the pictures, and this requires a good understanding of the effects that different impairment dimensions have on perception.

Much of the existing work done in this area has been focused on the aesthetic aspect of quality. Features used to estimate the aesthetic value or classify aesthetic categories of consumer photographs are color and illumination, composition, depth of field and perspective, and subject matter [9]. Such features have even been used to provide automatic feedback to photographers when composing a shot [10].

Although aesthetic aspects are no doubt important, they become secondary for the rather large class of "family photos" that most amateur photographers are concerned with. For family photos, human factors, such as facial expressions, pose, activity, and interaction, are by far the most important factors that determine the value of an image. If there are people in a scene, a human observer will immediately focus their attention on them and their faces and largely ignore the other characteristics of the image [11]. Consequently, assessing human factors is of paramount importance to intelligently process family photo collections.

Unfortunately, human factors are much harder to measure than aesthetic or other low-level factors. Problems such as face detection or recognizing people, poses, activities, and expressions are still some of the most challenging problems in computer vision, especially for images captured in uncontrolled conditions.

2.2 Summarization

Selecting the most representative pictures from a set is similar to storytelling or summarization; the key is to identify which scenes the user considers to be important in the story. There may not be a unique set of pictures that can fully represent a collection because of the large number of possible subsets and different possible themes.

An effective summary should have certain properties: quality, the selected photos have to be interesting and attractive; diversity, there should be no duplication or redundancy; and coverage, important people or events should appear in the summary [12].

Criteria that people use to choose pictures from a collection have been studied previously. Such criteria include specific people, variety of places, and general image quality [8]. These can be used to guide the (semi)automatic selection process. Furthermore, it can be helpful that some events, such as weddings, in certain cultures follow a specific sequence of events. There may be a number of important milestones that need to be included.

It may also be desirable to find pictures that are not be part of the initial set but that are nevertheless relevant to the story and can be sourced from external collections. Examples of this are a map of places visited or a better picture of a popular sight if the ones present in the collection are not satisfactory.

Finally, the purpose of summarization is not necessarily to

produce a static album or set of photos; instead, it can be used for dynamic browsing of photo collections. Of particular interest is "associative" browsing, which refers to any method that assists users to discover, browse, or navigate large data libraries in a more intuitive way. With associative browsing, a user is guided towards other similar data that is relevant to the data currently being viewed. Naturally, the human factor plays a big role in the associations a person has when looking at photos. This makes it difficult not only to develop approaches for meaningful summarization but also to evaluate their effectiveness. Furthermore, associations are highly subjective, which brings us to the topic of personalization.

2.3 Personalization

Personal and social factors are much more important for user-generated content because such content is often only meaningful to the user and their family and friends. Consequently, generic models for appeal may be even more short-lived than those for aesthetics [13]. For personalization to be effective, it must be carefully tailored toward learning personal or situational preferences. Personalization implies that the criteria for selecting images are not those of the average user (as typified by the traditional MOS) but of the specific user. This approach is quite different to the way the topic is usually approached.

Pictures selected from a personal collection by a random person are unlikely to be the meaningful or relevant pictures for the owner. A given person may have certain preferences in terms of perspective, lighting, color, enhancements, subjects, expressions, and poses. Any QA system for such content should be able to offer personalized suggestions according to the user's individual taste and preferences. Image content and visual characteristics alone are likely insufficient, and image metadata such as tags, geographical information, time, and date can greatly help with personalization tasks.

3 Examples

3.1 Interactive Photo Screening

People often take multiple shots of the same scene and then select the best picture(s) from the set afterwards. This is especially common for photos that involve people, for example, family photos with babies and kids or photos of certain important events, such as weddings or graduations. In these cases, we want to capture the best moments when the subjects of the photos have the most memorable poses. Then, we want to share our favorite photos with family and friends.

Photo screening (triaging) is one of the most common photowork activities. Existing photo software provides very limited computational or interface support for photo triaging; in many cases, this basic task still relies on flipping through the photos and viewing them one by one, which is a primitive interaction method. Using thumbnail images as an alternative does not work well either because the relevant image features to be compared are often too small in a thumbnail view. Thus, details such as facial expressions are not easily recognizable. This is especially problematic on mobile devices with limited screen space and resolution.

We therefore propose an effective and easy-to-use brush-and-drag interface that allows the user to interactively explore and compare photos within a broader scene context (Fig. 1) [7]. First, the user brushes an area of interest on a photo. Our tailored segmentation engine automatically determines corresponding image elements among the photos. Then, the user can drag the segmented elements from different photos across the screen to explore them simultaneously and further use simple finger gestures to interactively rank photos, select favorites to sharing, or remove unwanted photos. This focus + context design allows the user to choose any area or object of interest by brushing (focus) while retaining the overview photo (context). The photo triaging process becomes more flexible and user-centric.

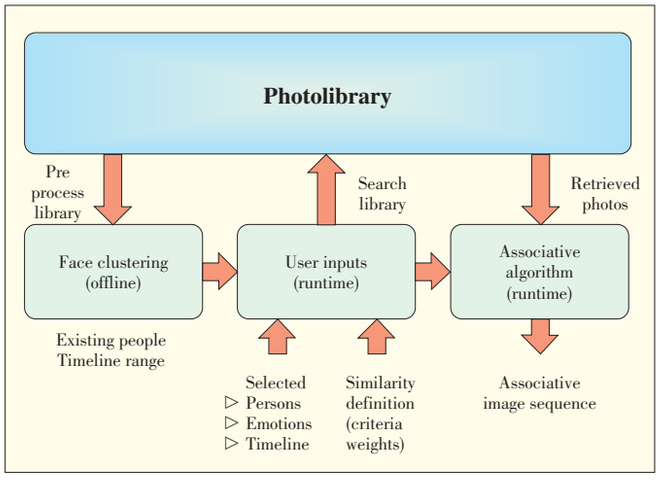
We implemented our interface on an Apple iPad 2 and evaluated it with a number of users. According to both objective and subjective measurements, our brush-and-drag interface is better than the conventional method of browsing photos by flipping. The participants preferred our interface in terms of ease of use and were able to select favorite photographs from groups of similar images more quickly [7].

3.2 People-Centric Summarization

There is a considerable body of research on slideshows and even some commercial products (e.g. Apple iPhoto) for automatic face annotation in personal photo-albums. However, there are no existing systems that can identify people and their emotions in photo libraries and use this information, along with other similarity features, to form an associative chain of image transitions or browsing suggestions. The majority of existing techniques that estimate human emotions do not take into account the human factor; they mainly focus on other global or local image features.

We have developed a method of creating people-centric slideshows that takes into account people and their emotions [14]. Fig. 2 shows how this system operates. The user specifies the person(s) that they wish to include along with the importance that they assigns to different similarity criteria. The system automatically scans the photolibrary for photos of the requested person(s) and performs face recognition and emotion

▲Figure 1. The brush-and-drag interface for photo screening.



▲**Figure 2. People-centric slideshow creation.**

estimation [15]. The retrieved images are then automatically arranged into a meaningful sequence, taking into consideration the importance values assigned by the user. The resulting image sequence can be displayed as a slideshow or to give browsing suggestions to the user. The current similarity criteria include facial emotions/expressions, timeline, color, and scene characteristics.

The system has a flexible design, and new similarity attributes can be easily added. It is also adaptable to the user's preferences. Different degrees of importance can be defined for the similarity attributes, making it a useful tool for personalized associative browsing or slideshow creation. The proposed system also takes into account emotions, which makes it useful for filtering out undesirable expressions such as angry faces.

4 Conclusions

We have contrasted traditional media and user-generated content in terms of their requirements for quality assessment. Using the example of photo collections, we have outlined a number of relevant research areas for novel quality assessment approaches. We have also highlighted two sample applications for photo screening and summarization that address issues such as user interface design and personalization that are important for user-generated content.

Acknowledgements

This research is supported by the Advanced Digital Sciences Center (ADSC) under a grant from the Agency for Science, Technology and Research of Singapore (A*STAR).

References

[1] W. Lin, C.-C. J. Kuo: "Perceptual visual quality metrics: A survey." *Journal of Visual Communication and Image Representation*, vol. 22, no. 4, pp. 297–312, May 2011.
[2] S. Winkler, P. Mohandas: "The evolution of video quality measurement: From PSNR to hybrid metrics." *IEEE Transactions on Broadcasting*, vol. 54, no. 3, pp. 660–668, Sept. 2008.
[3] S. Chikkerur, V. Sundaram, M. Reisslein, L. J. Karam: "Objective video quality assessment methods: A classification, review, and performance comparison." *IEEE Transactions on Broadcasting*, vol. 57, no. 2, pp. 165–182, June 2011.
[4] M. Vranješ, S. Rimac-Drlje, K. Grgic: "Review of objective video quality metrics and performance comparison using different databases." *Signal Processing: Image Communication*, vol. 28, no. 1, pp. 1–19, Jan. 2013.
[5] S. Winkler. "Analysis of public image and video databases for quality assessment." *IEEE Journal on Selected Topics in Signal Processing*, vol. 6, no. 6, pp. 616–625, Oct. 2012.
[6] S. Winkler: "Video quality measurement standards-current status and trends." In *Proc. 7th International Conference on Information, Communications and Signal Processing (ICICS)*, Macau, Dec. 7–10, 2009.
[7] S. J. Kim, H. Ng, S. Winkler, P. Song, C.-W. Fu. "Brush-and-Drag: A multi-touch interface for photo triaging." In *Proc. ACM SIGCHI International Conference on Human-Computer Interaction with Mobile Devices and Services (Mobile HCI)*, San Francisco, Sept. 21–24, 2012.
[8] A. E. Savakis, S. C. Etz, A. Loui: "Evaluation of image appeal in consumer photography." In *Proc. SPIE Human Vision and Electronic Imaging*, vol. 3959, San Jose, CA, Jan. 2000.
[9] C. Li, T. Chen, "Visual aesthetic quality assessment of digital images." In R. Lukac (ed.), *Perceptual Digital Imaging: Methods and Applications*. Chap. 4, pp. 91–122, CRC Press, 2012.
[10] L Yao, M. Qiao, P. Suryanarayan, J. Z. Wang, J. Li: "OSCAR: On-site composition and aesthetics feedback through exemplars for photographers." *International Journal of Computer Vision*, vol. 96, no. 3, pp. 353–383, 2012.
[11] E. Birmingham, W. F. Bischof, A. Kingstone: "Saliency does not account for fixations to eyes within social scenes." *Vision Research*, vol. 49, pp. 2992–3000, 2009.
[12] P. Sinha, S. Mehrotra, R. Jain: "Summarization of personal photologs using multidimensional content and context." In *Proc. ACM International Conference on Multimedia Retrieval (ICMR)*, Trento, Italy, April 17–20, 2011.
[13] D. Joshi, R. Datta, Q.-T. Luong, E. Fedorovskaya, J. Z. Wang, J. Li, J. Luo, " Aesthetics and emotions in images: A computational perspective." *IEEE Signal Processing Magazine*, vol. 28, no. 5, pp. 94–115, Sept. 2011.
[14] V. Vonikakis, S. Winkler. "Emotion-based sequence of family photos." In *Proc. ACM Multimedia Conference*, Nara, Japan, Oct. 29–Nov. 2, 2012.
[15] V. Vonikakis, S. Winkler. "System for creating slideshows based on people and their emotions." In *Proc. ACM Multimedia Conference*, Nara, Japan, Oct. 29–Nov. 2, 2012.

Manuscript received: January 16, 2013

Biography

Stefan Winkler

Stefan Winkler (stefan.winkler@adsc.com.sg) is principal scientist and director of the Interactive Digital Media Program at the University of Illinois Advanced Digital Sciences Center (ADSC) in Singapore. He has previously co-founded a start-up, worked in several large corporations, and held faculty positions at two universities. Dr. Winkler received his PhD degree from the Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland, and his MEng. degree from the Technische Universität Wien, Austria. He has published more than 80 papers and authored a book called *Digital Video Quality*. He is an associate editor of *IEEE Transactions on Image Processing* and *IEEE Signal Processing Magazine* (standards column). He has also contributed to video quality standards in VQEG, ITU, ATIS, VSF, and SCTE. His research interests include video processing, computer vision, perception, and human-computer interaction.

An Improved Color Cast Detection Method Based on an AB–Chromaticity Histogram

Ping Lu, Xia Jia, and Tirui Wu
(Pre–Research Department of ZTE Corporation, Nanjing 210012, China)

Abstract

AB–chromaticity histogram analysis works well most of the time, but it may not work well when the color cast is not severe. To overcome this problem, we propose an improved, two–step automatic cast–detection method. First, we compute the RGB color variance to evaluate the quality of the input image. If this variance is very small, we extract near–neutral color areas and compute the local ab–chromaticity histogram. We use this local ab–chromaticity histogram to evaluate the quality of the input image. This method has been tested in ZTE's video surveillance system. The results show that the proposed method produces better results based on subjective evaluation and is more efficient in various conditions.

Keywords

color cast; AB–chromaticity histogram; near–neutral color areas

1 Introduction

Many image processing algorithms are currently in use thanks to the Internet and the growing popularity of smart devices. Because of the differences between human eyes and digital camera lenses, it is necessary to develop techniques for automatic focusing, exposure, color adjustment, and white balance in order to improve the qualitiy of color in captured images.

A number of color–adjustment algorithms have been proposed in the literature. The most widely used of these algorithms is white balance, which adjusts color cast caused by light sources. There are two categories of white balance algorithm: gray world assumption [1] and max white [2]. Gray world assumption calculates weighting values for color correction by matching the color average with a gray reference value. However, the algorithm may fail if the image only contains a few color elements. Max white derives weighting values for color correction by changing the white point in the image to a white reference point. It fails if the white point is not found in the image.

There are also a number of other methods for color adjustment. For example, neural networks can be used to determine the type of illuminant and to correct images after a vast amount of illuminant data has been collected [3]. In the color–by–correlation method, illuminant data is first collected in order to set up a correlation matrix that is used to determine the type of illuminant and correct the image [4]. In the illuminant voting method, a probability and voting process is used to determine the type of illuminant [5]. The neural network, color–by–correlation, and illuminant–voting methods are all complicated because they require heavy computation in advance.

Color–cast detection protects images without color cast before the white balance algorithm is applied. In [6], the nonlinear classification function of a neural network is used to classify the input images as real cast, no cast, or intrinsic cast. The neural network uses seven features for detecting real cast and no cast and six features for detecting intrinsic cast. Other methods include the threshold method [7] and histogram method [8]–[10]. Color–cast detection methods can be used when

- images uploaded to or downloaded from the internet have color cast
- images have color cast caused by improperly adjusting the automatic white balance in the digital camera
- images have color cast from other devices, such as scanners or monitors
- old and stained photos are recaptured by a digital camera
- images have uncertain color cast.

In [11], a very effective method was proposed to detect color cast in images. The method is based on the assumption that when there is more color cast in an image, the distribution of the ab–chromaticity histogram is more centralized. Fig. 1(a) shows an image that has color cast, and Fig. 1(b) is the corresponding ab–chromaticity histogram. The distribution of the histogram is extremely centralized. This method works well most of the time, but it may not be suitable when the color cast is not severe. When evaluated over the entire image, the subjective assessment index may become very small because of dilution, and this leads to misclassification. To overcome this problem and evaluate the input image, we propose a method based on the local ab–chromaticity histogram of the near–neutral regions of the image.

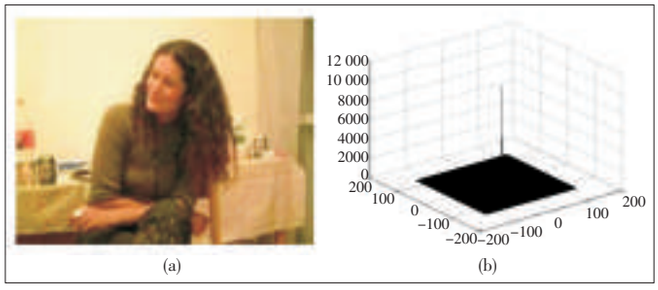
2 Improved Color–Cast Detection Method

2.1 Original Ab–Chromaticity Histogram Method

There are two basic types of color cast: intrinsic, which is

An Improved Color Cast Detection Method Based on an AB–Chromaticity Histogram

Ping Lu, Xia Jia, and Tirui Wu



▲**Figure 1. (a) An image with color cast and (b) the corresponding ab–chromaticity histogram.**

caused by a dominant color such as sky blue; and real, which is caused by the failure of the capturing device or by unusual lighting conditions. These two types of color cast are difficult to distinguish, so we simply treat them both as color cast. Here, we briefly describe the original ab–chromaticity histogram method [11].

Most color images are represented in the RGB color space; that is, a color image comprises a red channel, a green channel, and a blue channel. However, the RGB color space is not suitable for image proccesing because the correlations between the three channels are large, and the three channels contain a large amount of redundant information. Therefore, the original ab–chromaticity histogram method is used in the Lab color space. Compared with the RGB color space, the Lab color space has some useful attributes that help us detect color cast. The more color cast in the image, the more centralized the distribution of the ab–chromaticity histogram. In the original method, RGB is converted to Lab color; then, the 2D ab–chromaticity histogram H_{ab} is computed. Once this histogram has been computed, the following equation is used to provide statistics that help us analyze the histogram:

$$\begin{cases} \mu_k = \sum_k k H_{ab}(a, b) \\ \sigma_k = \sqrt{\sum_k (\mu_k - k)^2 H_{ab}(a, b)} \end{cases} \quad (1)$$

where $k = a, b$ (channel); μ_k is the mean levels of chromaticity in channels a and b; and σ_k is the variance of the ab–chromaticity histogram along the axis a and b, respectively.

Because a histogram is not an intuitive way to display the statistics from (1), we can use an equivalent circle. The center of the circle is (μ_a, μ_b), the radius is $\sigma = \sqrt{\sigma_a^2 + \sigma_b^2}$. We can define the minimum distance between the circle and the center of the ab–chromaticity plane ($a = 0$, $b = 0$) as

$$d = \mu - \sigma \quad (2)$$

where $\mu = \sqrt{\mu_a^2 + \mu_b^2}$. Then, we compute the color–cast coefficient (CCC):

$$d_\sigma = d / \sigma \quad (3)$$

Because d measures how far the whole histogram lies from the neutral axis ($a = 0$, $b = 0$) and σ is the spread of the histogram, d_σ can be used to quantify the strength of the cast.

2.2 Proposed Method

AB–chromacity histogram analysis may not work well when the color cast is not severe. To overcome this problem, we use the local ab–chromaticity of neural regions (defined in section 2.2.1) rather than the ab–chromaticity histogram of the whole image to evaluate the image quality. Furthermore, in order to speed up our method, we add an initial decision step. Fig. 2 shows the flow, which can be summarized follows:

Input: color image

Output: Boolean value(to indicate whether color cast exists or not)

Procedure:

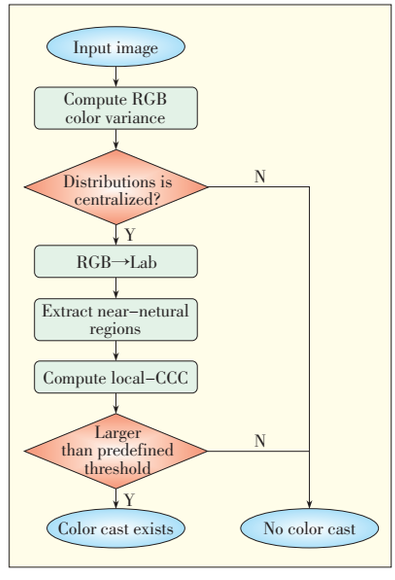
1. Compute RGB color variance $\sigma_{RGB} = \sum_i [(R_i - \bar{R})^2 + (G_i - \bar{G})^2 + (B_i - \bar{B})^2]$, where i is the pixel index and $\bar{R}$, $\bar{G}$, and $\bar{B}$ are the mean gray levels of red, green and blue channels, respectively.
2. If the σ_{RGB} is smaller than a predefined threshold $T_{\sigma RGB}$, go to step 3, else return false.
3. Convert the input image from RGB color space to Lab color space.
4. Extract near–netural regions
5. Compute local–CCC using equations (1)–(3).
6. If local–CCC greater than predefined threshold $T_{local-CCC}$, return true, else return true.

2.2.1 Extraction of Near–Neutral Regions

We propose using the local–CCC of near–neutral regions as the subjective color–cast index. We define a near–neutral region as pixels that have an ab–chromatictiy value near the center of the ab–chromacity plane; that is, the pixels are grayless. Fig. 3 shows some experiemental results for extraction of near–neutral regions.

3 Experimental Results

In our experiments, $T_{\sigma RGB} = 30$ and $T_{\text{local–CCC}} = 0$. Fig. 4 shows some results from the experiments. The left column shows the input image; the middle column shows the equivalent circle of the global ab–chromaticity histogram; and the right column is the equivalent circle of the local ab–chromaticity histogram. When there is no color cast, the difference between global



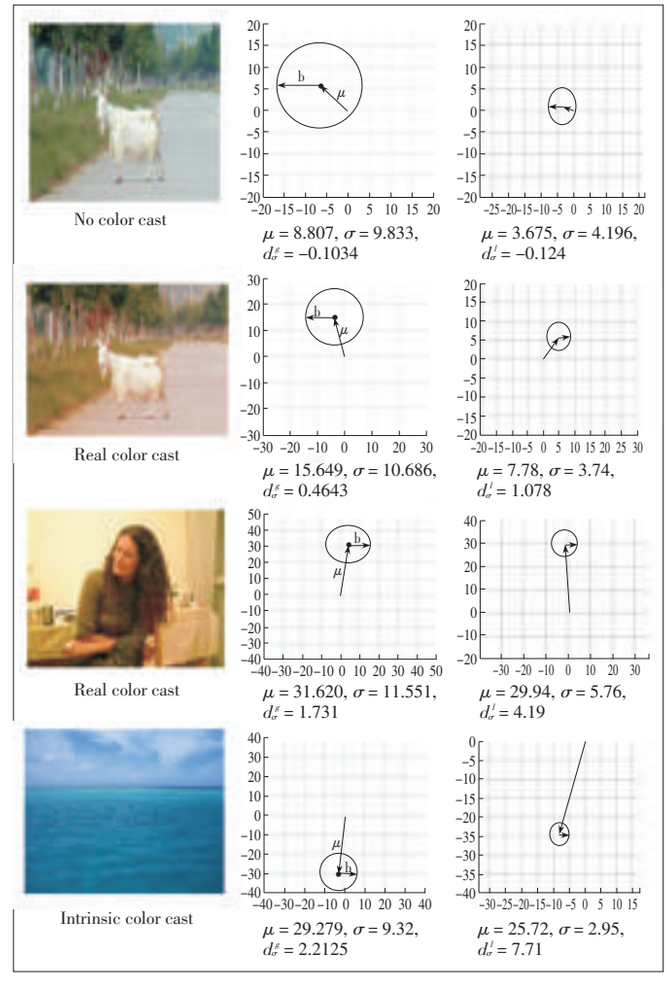
▲**Figure 2. Flow of the proposed method.**

◀Figure 3. Extraction of near–neutral regions. Left: input image; right: near–neutral region.

CCC d_σ^g and local CCC d_σ^l is small. In Fig. 3(a), $abs(d_\sigma^g - d_\sigma^l) = 0.0206$ (nearly zero). However, when ther is color cast, the difference between d_σ^g and d_σ^l is larger. In Fig 3(b), (c) and (d), the differences between d_σ^g and d_σ^l are 0.6137, 2.459, and 5.4975, respectively. That is, if there is color cast, the difference between the global CCC and local CCC increases. Local CCC is a more distinguishable subjective index for color–cast evaluation.

4 Conclusion

In this paper, we have proposed an improved color–cast detection method in which local CCC of near–neutral regions is used to evaluate the input image and determine whether color cast exists in that image. Our method can overcome the limitations of using the ab–chromaticity histogram method when the input image does not have severe color cast. This method has been used in ZTE's video surveillance system to evaluate video quality and is effective in a range of environments.



▲Figure 4. Experiental results for the proposed method.

References

[1] K. Barnard, V. Cardei, and B. Funt, "A comparison of computational color constancy algorithms. I: Methodology and experiments with synthesized data," *IEEE Trans. Image Processing*, vol. 11, no. 9, pp. 972–984, 2002.

[2] K. Barnard, L. Martin, A. Coath, and B. Funt, "A comparison of computational color constancy algorithms. II. Experiments with image data," *IEEE Trans. Image Processing*, vol. 11, no. 9, pp. 985–996, 2002.

[3] K. Barnard, V. Cardei, and B. Funt, "Estimating the Scene Illumination Chromaticity using a Neural Network," *J. Opt. Soc. Amer. A*, vol. 19 no. 12, pp. 2374–2386, 2002.

[4] G. D. Finlayson, S. D. Hordley, and P. M. Hubel, "Color by correlation: a simple, unifying framework for color constancy," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 23, no. 11, pp. 1209–1221, 2001.

[5] G. Sapiro, "Color and illuminant voting," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 21, no. 11, pp. 1210–1215, 1999.

[6] Sheng–Fuu Lin, Huang–Tsun Chen, and Tsung–Han Lin, "Color casts detection and adjustment," *International Journal of Computer Science Issues*, vol. 8, no 2, Jul. 2011.

[7] F. Li and H. Jin, "An approach of detecting image color cast based on image semantic," in *Proc. of IEEE Conf.: Machine Learning and Cybernetics*, Shanghai, 2004, pp. 3932–3936.

[8] F. Gasparini and R. Schettini, "Color correction for digital photographs," in *Proc. of IEEE Conference on Image Analysis and Processing*, Mantova, 2003, pp. 646 – 651.

[9] F. Gasparini, R. Schettini, and P. Gallina, "An Innovative Algorithm for Case Detection," in *Proc. of SPIE*, San Jose, CA, 2002, vol. 4672, pp. 280–286.

[10] F. Gasparini, R. Schettini, and P. Gallina, "Tunable Cast Remover for Digital Photographs," in *Proc. of SPIE*, Santa Clara, CA, 2003, vol. 5008, pp. 92–100.

[11] F. Gasparini and R. Schettini, "Color balancing of digital photos using simple image statistics," *Pattern Recognition*, vol. 37, no. 6, pp. 1201–1217, 2004.

Manuscript received: January 21, 2013

Biographies

Ping Lu

Ping Lu (lu.ping@zte.com.cn) graduated from South East University, China, majoring in automatic control theory and applications. He is the chief executive officer of the Service Institute of ZTE Corporation. For more than a decade, he has steered the institute towards innovative research and development of value added services, cloud computing, Internet services, ICT services, and home network services.

Xia Jia

Xia Jia recieved her MS degree in automatic control theory from Dalian University of Technology, China, in 2001. She joined ZTE Corporation and is now the project manager of multimedia technology research. Her main fields of research are IPTV, OTT, video conferencing, and augmented reality.

Tirui Wu

Tirui Wu received his MS degree in computer science and technology from Jiangsu University of Science and Technology, China, in 2009. In 2010, he joined ZTE Corporation R&D. His main research interests include image enhancement, pattern recognition, and human computer interaction.

Battery Voltage Discharge Rate Prediction and Video Content Adaptation in Mobile Devices on 3G Access Networks

Is-Haka Mkwawa and Lingfen Sun
(School of Computing and Mathematics, University of Plymouth, Plymouth, PL4 8AA, UK)

Abstract

According to Cisco, mobile multimedia services now account for more than half the total amount of Internet traffic. This trend is burdening mobile devices in terms of power consumption, and as a result, more effort is needed to devise a range of power-saving techniques. While most power-saving techniques are based on sleep scheduling of network interfaces, little has been done to devise multimedia content adaptation techniques. In this paper, we propose a multiple linear regression model that predicts the battery voltage discharge rate for several video send bit rates in a VoIP application. The battery voltage discharge rate needs to be accurately estimated in order to estimate battery life in critical VoIP contexts, such as emergency communication. In our proposed model, the range of video send bitrates is carefully chosen in order to maintain an acceptable VoIP quality of experience. From extensive profiling, the empirical results show that the model effectively saves power and prolongs real-time VoIP sessions when deployed in power-driven adaptation schemes.

Keywords

QoE; power; mobile devices; quality adaptation; discharge rate

1 Introduction

According to the Cisco Visual Networking Index of the Global Mobile Data Traffic Forecast, mobile video traffic accounted for 52% of total Internet traffic at the end of 2011 [1]. Mobile video traffic will continue to increase so that by 2016, two-thirds of mobile data traffic will be video. YouTube and Facebook account for more than 30% of all Internet traffic, YouTube accounting for a bigger portion of this traffic than Facebook [1]. Most video traffic goes to mobile devices; however, mobile applications use a considerable amount of power, especially when video is involved [2], [3]. Processing power and storage capacity has grown exponentially over the past few years, but battery life has not. Power conservation in mobile devices has been extensively researched. Power-saving techniques, such as mobile resource management, power-aware operating systems, and Wi-Fi/3G sleep scheduling have been exploited [4].

CPU, LCD, GPS, Wi-Fi, and 3G components consume a lot of power, as do multimedia applications. Any effort to reduce power consumption in any one of these components and applications is of paramount importance. In the case of video VoIP, GPS can be switched off because it is not needed, and one of the network interfaces can also be switched off depending on the access network being used. However, the LCD must not be switched off in order to allow the video watching.

Sleep scheduling techniques cannot be applied in this situation because the communication is in real time. Sleeping schedules conflict with highly delay-sensitive VoIP services and degrade VoIP (QoE) [5]. Therefore, it is more appropriate to adapt multimedia VoIP applications than network interfaces in order to save energy. This approach has led to a content-adaptation technique that can be used to save energy and maintain VoIP QoE at an acceptable level [6].

This paper describes an extension of the power-driven adaptation scheme proposed in [6]. In this scheme, video send bitrates (SBRs) were mapped into battery charge levels. In the proposed scheme, battery life during VoIP communication was not predicted; hence, selection of SBRs did not determine the length of the VoIP session.

In this paper, we use a multiple linear regression analysis to predict the battery voltage discharge rate. By predicting this rate, we can accurately estimate the battery life, and appropriate SBRs can be used to save power or prolong the VoIP session. The secondary purpose of this paper is to revise the power-driven VoIP adaptation scheme proposed in [6] so that it includes the battery voltage discharge rate for SBR selection.

2 Related Work

The authors of [7] proposed a framework that reduces power consumption in Wi-Fi video streaming. The proposed framework uses various video SBRs to adjust sleep intervals of the Wi-Fi. This reduces power consumption and, at the same time, maintains video quality. In [8], the same technique was

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 284863 (FP7 SEC GERYON).

also proposed, and the Power Save Mode (PSM) standard was used [9]. These frameworks do not address key issues in real–time VoIP communication, where the delay or loss of important signalling traffic, such as SIP, can affect the real–time communication.

The authors in [10] proposed a system context–aware approach to predict the life of a smartphone battery. They showed how changing mobile system components affects the life of a battery. A video player application with LCD brightness was used as a dependent variable; however, video content adaptation was not considered. This can also lead to another dependent variable, and more power can be saved. In this paper, we go beyond their approach and consider video content adaptation in terms of video SBRs.

In [6], we proposed a power–driven adaptation scheme in which SBRs are switched in order to save power over Wi–Fi networks. Wi–Fi power is consumed in either the listening state or transmission state. Power consumption levels in these two states were constant and could be broken down into low and high power levels. However, we did not predict the battery voltage discharge rate, which is useful for estimating the battery life for each proposed SBR during the VoIP session.

Here, we propose a multiple linear regression model that can be used to predict the battery voltage discharge rate for several video SBRs. The proposed model is then used to estimate the battery life during a VoIP session. The range of video SBRs is carefully chosen in order to maintain acceptable QoE.

3 Experimental Testbed

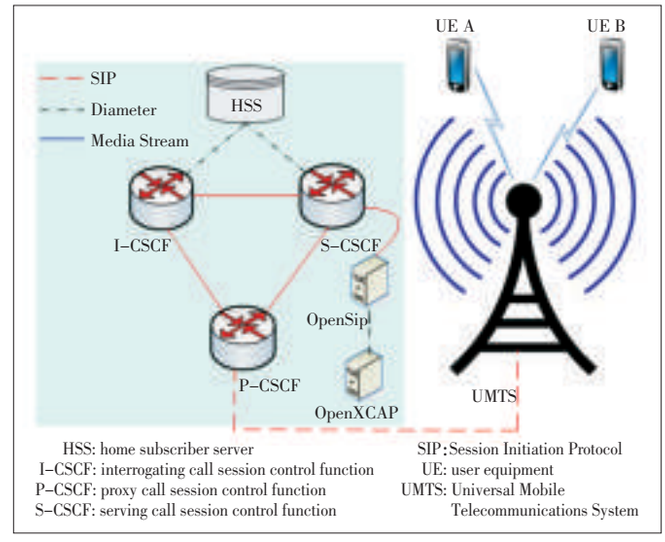
We developed a testbed based on Open IMS Core to evaluate the proposed model (Fig. 1). Session Initiation Protocol (SIP) [11], the de facto signaling protocol in IMS, was used to establish and terminate a VoIP session.

Two HTC Dream G1 mobile phones with Android 2.3 were ported with IMSDroid [12] and used as clients for VoIP communication. The Universal Mobile Telecommunications System (UMTS) access network was provided by O2 UK. For voice sessions, the AMR–NB base profile codec was used, and for video sessions, the H264 base profile codec was used. An Android–based power monitoring tool called Powertutor [13] was installed on the mobile phones, and statistics were collected every second. The power sources for both mobile phones were rechargeable 1100 mAh, 3.7 V lithium–ion batteries.

OpenSips [14] and OpenXCAP [15] were deployed in the testbed for presence capabilities and to store monitored battery charge and power levels in a centralized way.

4 Evaluation of Power Consumption

In this section, we evaluate the power consumption of mobile components that significantly contribute to overall power consumption in the system. Power consumption differs between mobile devices because of software and hardware.

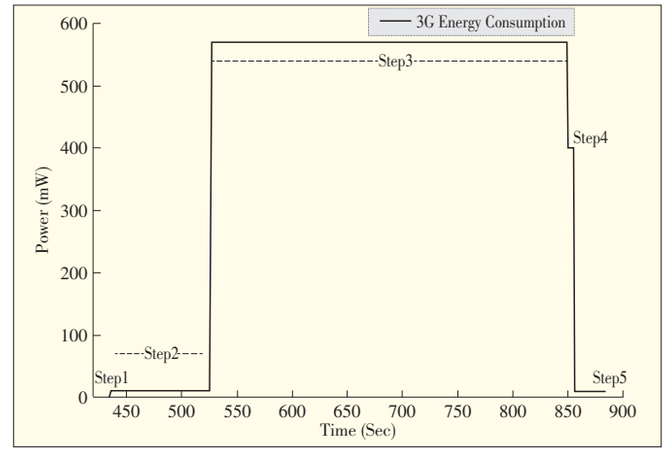


▲Figure 1. Experimental testbed.

4.1 3G Interface

Fig. 2 shows the power consumed by the 3G interface in the HTC Dream G1. The chipset is a Qualcomm RTR6285. According to 3GPP, 3G has three states: IDLE, dedicated channel (DCH), and forward–access channel (FATCH). IDLE is the low–power–consumption state: the radio resource control is in IDLE when no data is being transmitted. DCH and FATCH are high–power–consumption states. DCH guarantees low delay and high throughput by reserving dedicated channels to mobile devices. The FATCH state occurs when there is less traffic, and the channel is shared between mobile devices.

Timers control the transitions between states. The transition from low–power to high–power state is immediate, but the transition from high–power to low–power state occurs only when the network has been inactive for a certain period of time. The transitional power from low–power to high–power state is



▲Figure 2. Power consumed by the 3G interface.

called *ramp* power, and the transitional power from high–power to low–power state is called *tail* power [16]. The following steps describe how power is used by the 3G interface (Fig. 2):

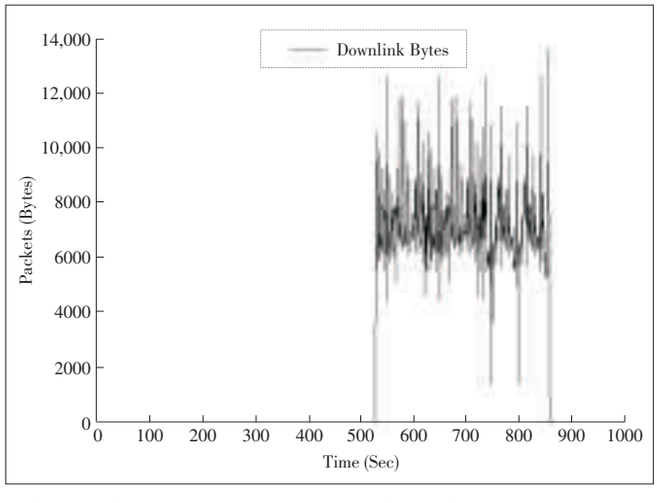
1) The 3G interface is off and no power is consumed.
2) The 3G interface is on, but no data is transmitted. An average of 10 mW is recorded, and the interface is in the low–power state.
3) VoIP multimedia transmission is started, and the interface is in a high–power state. An average of 570 mW of power is consumed. The power consumed during the transition from low– to high–power states is called *ramp* power.
4) VoIP multimedia transmission is stopped. The 3G interface waits for a fixed time before reverting to the low–power phase. An average of 401 mW of power is consumed, and the average wait time is six seconds. The power consumed during the transition from high– to low–power states is called *tail* power.
5) Low–power state.

The empirical results show that the high–power state of the 3G interface does not affect the transmission rate (Fig. 3).

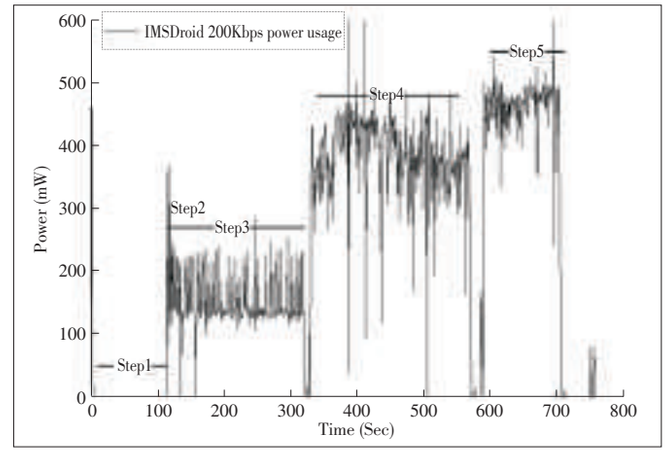
4.2 VoIP Application

Fig. 4 shows the power consumed by the IMSDroid AV application during the VoIP communication, which can be described by the following steps:

1) IMSDroid is off, and no power is consumed.
2) IMSDroid is switched on. The SIP and RTP stacks are initialized, and VoIP registration occurs at the IMS server. An average of 230 mW of power is consumed.
3) SIP signaling traffic for session negotiation is communicated which is then followed by the establishment of the RTP voice media communication. An average of 160 mW of power is consumed.
4) The incoming video transmission is initiated. An average of 370 mW of power is consumed.



▲Figure 3. Data transmission rate when 3G interface is in a high–power state.



▲Figure 4. Power consumed by IMSDroid.

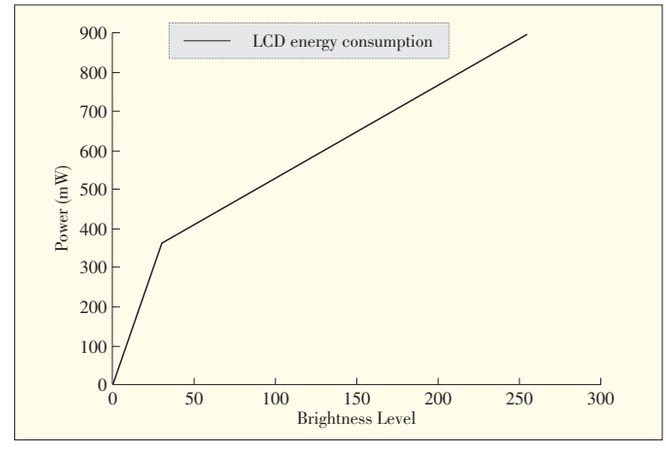
5) The outgoing video is triggered, and power consumption increases to an average of 470 mW.

4.3 LCD

Fig. 5 shows the power consumed by the glass TFT–LCD touch–sensitive HVGA screen. At an average of 360 mW, power consumption is the lowest when the brightness level lowest (i.e. when the brightness level is 30). At an average of 900 mW, power consumption is the highest when the brightness level is highest (i.e. when the brightness level is 255). Lowering the brightness significantly saves power, but the brightness level must be carefully selected so that it does not degrade the VoIP QoE. We set the brightness level at 30, 127 and 255 in the proposed multiple–regression model.

4.4 CPU

Fig. 6 shows the power consumed by the MSM7201A chipset, which includes an ARM11 application processor, ARM9 modem, and high–performance DSP. At an average of 45 mW, power consumption is lowest when the VoIP application is not running. At an average of 335 mW, power consumption is high-



▲Figure 5. LCD power consumption.

est when the VoIP application is running at 50 Kbps.

4.5 Total Power Consumption

Fig. 7 shows the total power consumed by the mobile phone during the video VoIP session, which can be described by the following steps:

1) The 3G interface is off.
2) The 3G interface is on and in listening mode.
3) The VoIP application is started.
4) The VoIP application is registered to the IMS.
5) The VoIP session is established.
6) RTP audio traffic flows.
7) The incoming video is played.
8) The outgoing video is transmitted.

5 Battery Voltage Discharge Rate Prediction Model

The battery life can be estimated by predicting the battery voltage discharge rate during a VoIP session. It is given as

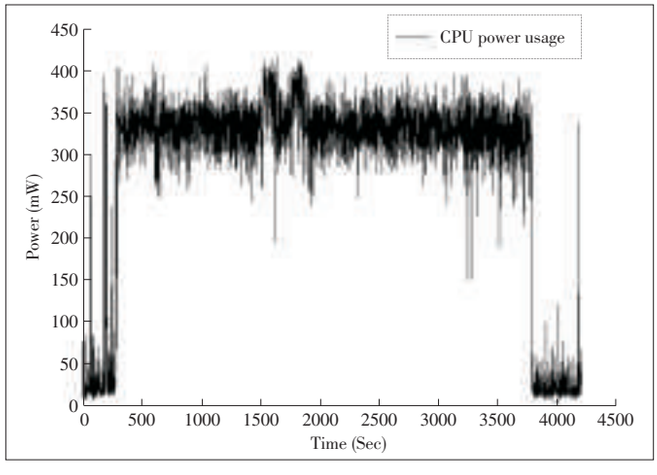
$$T(v_i, v_j) = (v_i - v_j)/\beta \quad (1)$$

where β is the battery voltage discharge, and v_i and v_j (for $v_i > v_j$) are the battery voltages at time i and j.

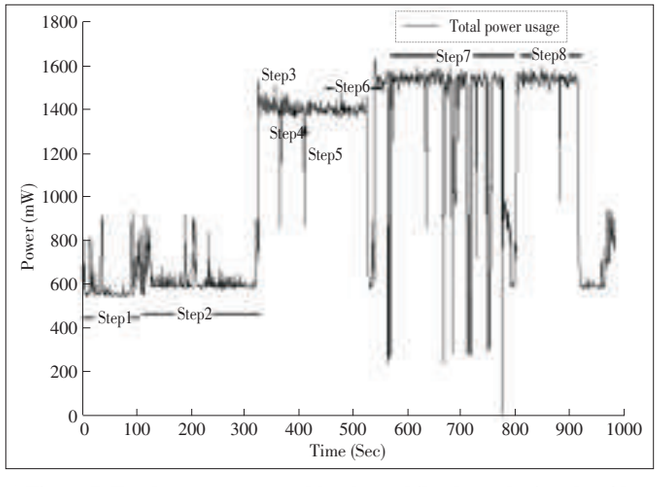
Fig. 8 shows the battery voltage discharge rate for different SBRs during the VoIP session. In the steady state, power consumption and, consequently, battery voltage discharge increases as the video SBR increases [6]. An appropriate video SBR can be chosen to save power and/or prolong VoIP communication without compromising the video QoE. The range of H264 codec SBRs for which there is an acceptable QoE has been tabled in [17]. These SBRs were deployed in [6].

Through extensive profiling, we found that during the VoIP session, there is a linear relationship between battery voltage discharge (a dependent variable) and independent variables such as CPU, LCD, GPS, Wi–Fi, and 3G interfaces. The video SBRs can therefore be varied and multiple linear regression analysis used to predict the battery voltage discharge. The battery voltage discharge rate is the dependent variable, and the main power–consuming components (i.e. CPU, LCD, GPS, Wi–Fi, 3G and IMSDroid) are independent variables. During the VoIP session, the CPU, GPS, Wi–Fi and 3G are kept constant. The 3G and Wi–Fi interfaces are either on or off, and only one interface is used at a time during the VoIP session. In this paper, a 3G interface is used and is switched on, from the start to the end of the session. The brightness level of the LCD backlight in the HTC phones is allowed to vary between 30 and 255, which improves the quality of the video session. The GPS interface is not needed in this scenario and is therefore switched off. Table 1 lists independent variables and their range of values considered in this paper.

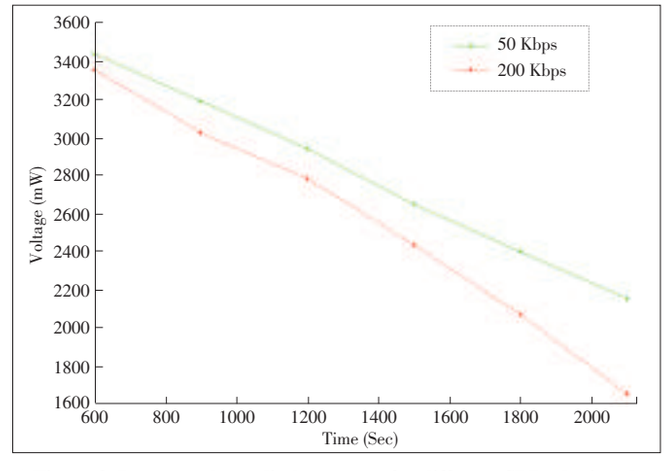
Table 2 shows the sample data for the battery voltage discharge, LCD, and VoIP application. The CPU frequency remained constant at 245 MHz during the course of the experiment. The 3G interface was switched on from the beginning to



▲Figure 6. CPU power consumption.



▲Figure 7. Total power consumed by the mobile phone during the video VoIP session.



▲Figure 8. Battery voltage discharge rate for different SBRs during the VoIP session.

▼Table 1. Independent variables

Variable	Range of Values
3G interface	On or off [0,1]
VoIP application	SBR range [50, 100, 200, 380]
CPU	Frequency (MHz) [245, 528]
GPS interface	On or off [0,1]
LCD backlight	Level [30, 127, 255]

▼Table 2. Sample Data

LCD Backlight Level	VoIP SDR Range	Battery Voltage Discharge Rate (mV/s)
255	50	0.6663
127	100	0.5734
255	200	0.7294
30	380	0.5211

the end of the VoIP session and did not vary in its power consumption. The GPS interface was off and did not affect power consumption in the experiment.

Therefore, the remaining independent variables were the LCD and VoIP application via its video SBRs.

The regression model is then expressed as

$$V = \alpha + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon \quad (2)$$

where V is the battery voltage, X_1, ... , X_k are independent variables (Table 2), and α and β_1, ... , β_k are the regression coefficients to be estimated. The independent random error is given by ε. In regression analysis, α and β_1 , ... , β_k are estimated by assuming ε is normally distributed; that is, the mean $\mu = 0$ and standard deviation $\delta = 1$.

The method of least squares is used to calculate the coefficients of (2) in order to yield the best–fitting equation. In this paper, $k = 2$, X_1 = VoIP, and X_2 = LCD; therefore (2) becomes

$$V = \alpha + \beta_1 VoIP + \beta_2 LCD + \varepsilon \quad (3)$$

If only VoIP is considered and the rest of the independent variables are kept constant, (3) can be reduced to

$$V = \alpha + \beta_1 VoIP + \varepsilon \quad (4)$$

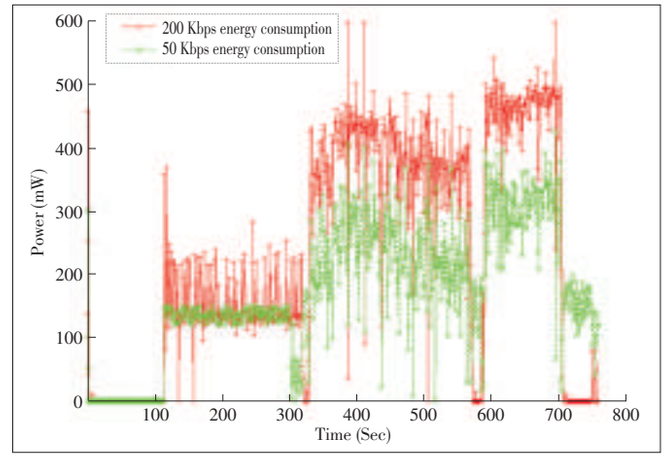
where β_1 is the slope of the regression line that represents the battery voltage discharge rate contributed by the VoIP application, and α is the line intercept, which represents the voltage when the VoIP application is off. When the VoIP application is on, α is the combined voltage contributed by other independent variables.

6 Experimental Results and Evaluation

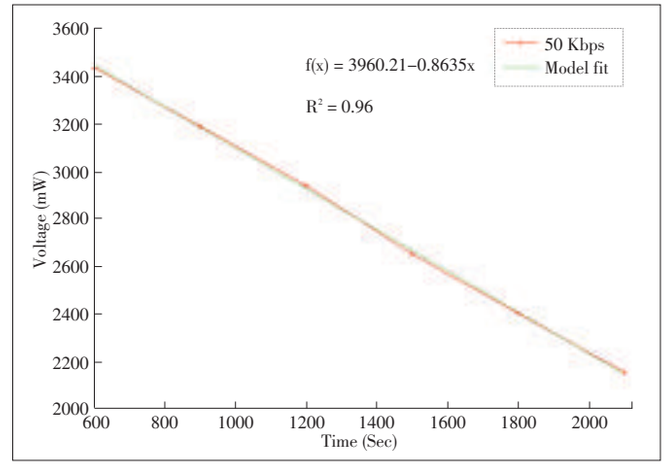
Experimental results in [6] showed that 10–30% power was saved when SBRs were changed from 200 Kbps to 50 Kbps (Fig. 9). In this paper, the results [6] are extended and used to predict the battery voltage discharge in order to estimate the battery life. The estimated battery life is used as one of the inputs into the proposed power–driven adaptation scheme [6].

Fig. 10 shows the battery voltage discharge rate when the video VoIP session is operating at 50 Kbps. Fig. 11 shows the battery voltage discharge rate when the video VoIP session is operating at 200 Kbps. When the video VoIP application is running at 50 Kbps, the battery voltage discharge rate is 0.8635 mV/s. When the same video VoIP application is running at 200 Kbps, the battery voltage discharge rate is 1.104 mV/s. When the video VoIP application is running at 50 Kbps, the voltage drops to 2400 mV around 1800 s after the battery is fully charged. When the video VoIP application is running at 200 Kbps, the voltage drops to 2400 mV approximately 1500 s after the battery is fully charged.

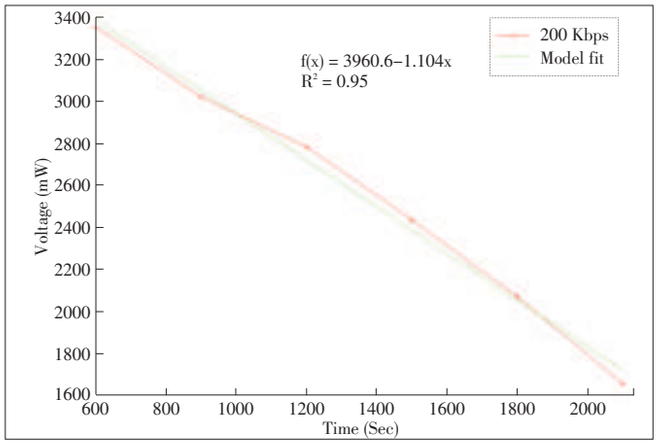
With extensive profiling, we found that the CPU frequency was constant at 245 MHz during the VoIP session. When the 3G interface was always on, CPU frequency was constant. When the GPS was switched off, the remaining independent variables were LCD and VoIP. The multiple regression coefficients are estimated by using the least square method, and the following equation is derived:



▲Figure 9. Power saving as the result of switching SBRs.



▲Figure 10. Battery voltage discharge rate at 50 Kbps.



▲Figure 11. Battery voltage discharge rate at 200 Kbps.

$$V = 3780.4 - 0.721\,VoIP - 1.563\,LCD \quad (5)$$

From (5), the intercept has not significantly changed (Figs. 10 and 11). This behavior is expected because if the LCD backlight is switched off and the VoIP application is not running, the intercept is the expected maximum value when both LCD and VoIP application are switched off. The VoIP coefficient, which is the battery voltage discharge rate due to VoIP application, changes very little, and this means that power consumption of the VoIP application and LCD backlight is not related. The battery voltage discharge caused by the LCD backlight was 1.563 mV, and the battery voltage discharge caused by the VoIP application was 0.721 mV.

The proposed model is evaluated using the mean residual error and mean prediction error. The mean residual error was 0.7%, and the mean prediction error was 2.21%.

7 Power–Driven VoIP Adaptation Scheme

In [6], a simple nonlinear regression analysis was done to estimate power consumption in the SBR range 50–500 Kbps (Fig. 12):

$$Power = 95.211\ In(SBR) + 311.84 \quad (6)$$

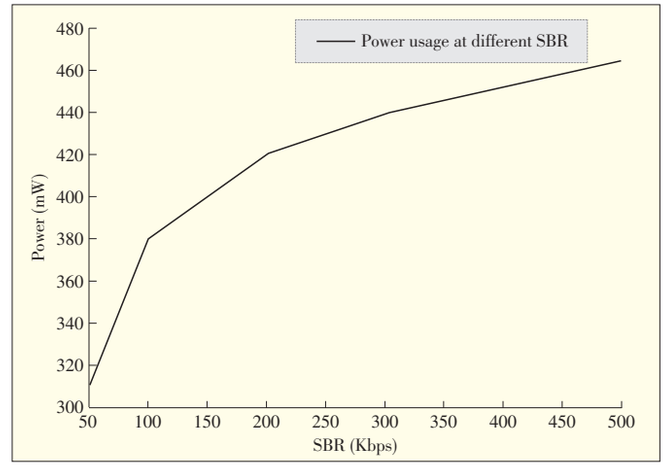
where SBR (50 Kbps $\leqslant SBR \leqslant$ 500 Kbps) is the video send bit rates of the H264 codec.

The battery charge levels (BCL) were mapped to the corresponding SBR values for VoIP quality adaptation. This mapping was then used in the power–driven adaptation scheme (Table 3, Fig. 12).

The BCL was not associated with the battery life; therefore, it was not possible to estimate how long it would take to stay in each BCL. The proposed battery voltage discharge rate model allows us to accurately estimate the battery life in each BCL. For example, when the VoIP application is running at 50 Kbps, the battery power drops to 2400 mV 1800 s after the battery is fully charged (Fig. 10). The power–driven VoIP adaptation scheme in [6] is revised to include the battery voltage discharge rate model (Algorithm 1):

Algorithm 1. VoIP Quality Adaptation Scheme

```
MBCC = Max_Batt_Charge_Capacity()
BCL = Batt_Charge_Level()
while Phone is Still Registered to the IMS do
  for Time Interval of 5 Seconds do
    PUT_Max_Batt_Charge_Cap(MBCC,XDMS)
    PUT_Batt_Charge_Level(BCL,XDMS)
    PUT_Batt_Vol_Discharge_Rate(BVDR,XDMS)
    while VoIP Session is Ongoing do
      BCL = GET_Batt_Charge_Level(Callee,XDMS)
      MBCC = GET_Max_Batt_Charge_Cap(Callee,XDMS)
      BVDR = GET_Batt_Vol_Discharge_Rate(Callee,XDMS)
      Comput_Batt_Lifetime(Callee,BVDR)
      Adapt_SBR_If_Needed()
      if BCL Treshold is reached then
        Switch off video transmission
        Switch off LCD
      end if
    end while
  end for
end while
```



▲Figure 12. Power consumption for different SBRs.

▼Table 3. Mapping of BCL to SBR

Level	BCL (%)	SBR (Kbps)	MOS
0	100–75	⩾500	4.2
1	75–50	300	4.2
2	50–40	100	3.5
3	40–15	50	3.5
5	15–0	Only voice	

1) The mobile phone uploads its maximum battery capacity, current battery charge level, and battery voltage discharge rate to the XDM server at the time of registration and then

at five second intervals provided the mobile phone is still registered. Five seconds is specified in the RTCP communication standard.

2) The mobile phone uses the presence server and XDM server to retrieve power capabilities when initiating a VoIP session.
3) When initiating the VoIP session, the mobile phone chooses the video SBR for acceptable QoE and battery life.
4) The mobile phone monitors power capabilities and the battery life through the published data in the XDM server at five second intervals.
5) Using the battery charge level, the mobile phone computes the battery voltage discharge rate and calculates the battery life.
6) The remaining battery life determines how SBRs should be adapted while maintaining acceptable QoE.
7) If the battery is low, the video and LCD are switched off. Only voice communication is left running. In this paper, low battery charge is 1600 mV.

8 Conclusion and Future Work

In this paper, we have proposed a model for predicting battery voltage discharge rate in mobile devices on 3G networks. The model is based on a multiple linear regression analysis. We use a video VoIP application, in which different video send bit rates and LCD backlight levels are independent variables, and battery voltage is a dependent variable. The proposed model can be used to accurately estimate battery life in real time during critical VoIP communications, such as emergencies. The battery voltage discharge rate model has been used in the proposed power-driven adaptation scheme [6], and it was found that 10-30% of the total power could be saved when video send bit rates were changed.

The proposed model can be extended to include several other video codecs, mobile devices, and VoIP applications.

References

[1] CISCO. (2012). *Cisco visual networking index: Global mobile data traffic forecast update, 2011-2016* [Online]. Available: http://www.cisco.com/en/US/ solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.pdf
[2] W. Yuan, K. Nahrstedt, S. V. Adve, D. L. Jones, and R. H. Kravets, "Grace-1: Cross-layer adaptation for multimedia quality and battery energy," *IEEE Transactions on Mobile Computing*, vol. 5, no. 7, pp. 799-815, 2006.
[3] J. Flinn and M. Satyanarayanan, "Energy-aware adaptation for mobile applications," in *Proc. of the seventeenth ACM symposium on Operating systems principles*, ser. *SOSP '99*, New York, NY, USA: ACM, 1999, pp. 48-63. [Online]. Available: http://doi.acm.org/10.1145/319151.319155
[4] N. Vallina-Rodriguez and J. Crowcroft, "Energy management techniques in modern mobile handsets," *Communications Surveys Tutorials*, IEEE, vol. 15, no. 1, pp. 179-198, 2013.
[5] C. Zhu, H. Yu, X. Wang, and H.-H. Chen, "Improvement of capacity and energy saving of voip over ieee 802.11 wlans by a dynamic sleep strategy," in *IEEE GLOBECOM 2009*, Honolulu, HI, Nov. 30-Dec. 4 2009, pp. 1-5.
[6] I.-H. Mkwawa and L. Sun, "Power-driven voip quality adaptation over wlan in mobile devices," in *IEEE GLOBECOM 2012 Workshop - QoEMC*, Anaheim, CA, Dec. 2012, pp. 1-5.
[7] M. Csernai and A. Gulyas, "Wireless adapter sleep scheduling based on video qoe: How to improve battery life when watching streaming video?" in *2011 Proc. of 20th International Conf. on Computer Communications and Networks (ICCCN)*, Maui, HI, Jul. 31-Aug. 4, 2011, pp. 1-6.
[8] V. Namboodiri and L. Gao, "Energy-efficient voip over wireless lans," *IEEE Transactions on Mobile Computing*, vol. 9, no. 4, pp. 566 -581, April 2010.
[9] *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications*, IEEE Std 802.ll-1999.
[10] X. Zhao, Y. Guo, Q. Feng, and X. Chen, "A system context-aware approach for battery lifetime prediction in smart phones," in *Proceedings of the 2011 ACM Symposium on Applied Computing, ser. SAC'11*. New York, NY, USA: ACM, 2011, pp. 641-646. [Online]. Available: http://doi.acm.org/10.1145/1982185.1982327
[11] J. Rosenberg, H. Schulzrinne, G. Camarillo, A. Johnston, J. Peterson, R. Sparks, M. Handley, and E. Schooler, "SIP: Session Initiation Protocol," Internet Engineering Task Force, RFC 3261, Jun. 2002. [Online]. Available: http://www.rfc-editor.org/rfc/rfc3261.txt
[12] D. Telecom. (2011). *Sip/ims client for android Website* [Online]. Available: http://code.google.com/p/imsdroid/
[13] L. Zhang, B. Tiwana, R. Dick, Z. Qian, Z. Mao, Z. Wang, and L. Yang, "Accurate online power estimation and automatic battery behavior based power model generation for smartphones," in *2010 IEEE/ACM/IFIP Int. Conf. on Hardware/Software Codesign and System Synthesis (CODES+ISSS)*, Scottsdale, AZ, Oct. 2010, pp. 105 -114.
[14] V. System. (2009). *Open sip server* [Online]. Available: http://www.opensips.org/
[15] M. Amarascu, R. Klaver, L. Stanescu, D. Bilenko, and S. Ibarra. (2006). *Open xcap server* [Online]. Available: http://www.openxcap.org/
[16] N. Balasubramanian, A. Balasubramanian, and A. Venkataramani, "Energy consumption in mobile phones: a measurement study and implications for network applications," in *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*, ser. IMC '09. New York, NY, USA: ACM, 2009, pp. 280-293. [Online]. Available: http://doi.acm.org/10.1145/1644893.1644927
[17] A. Khan, L. Sun, E. Jammeh, and E. Ifeachor, "Quality of experience-driven adaptation scheme for video applications over wireless networks," *IET Communications*, vol. 4, no. 11, pp. 1337 -1347, 2010.

Manuscript received: January 31, 2013

Biographies

Is-Haka Mkwawa

Is-Haka Mkwawa (Is-Haka.Mkwawa@plymouth.ac.uk) received his PhD in computing from the University of Bradford. He is currently working as a research fellow on the EU FP7 GERYON project at Plymouth University. Since 2002, he has also worked in various capacities on the EU FP6 and FP7 projects at Plymouth University, the University of Bradford, and University College Dublin. He has previously worked on other projects, including ADAMANTIUM, VITAL, NoE Euro FGi, SFI, NoE Euro NGi, and IASON. He is the author of several published works on parallel computing and communication, distributed systems, next-generation networks, grid computing, VoIP quality adaptations, energy conservation techniques and mobility management in mobile and wireless networks, and performance analysis and evaluation of computer networks. He is the co-author of the textbook Guide to Voice and Video over IP: For Fixed and Mobile Networks.

Lingfen Sun

Lingfen Sun (L.Sun@plymouth.ac.uk) received her PhD degree in computing and communications from the University of Plymouth in 2004. She received her MSc in communications and electronics systems and BEng in telecommunications engineering from the Institute of Communications Engineering, China, in 1988 and 1985. She is currently a reader in multimedia communications and networks in the School of Computing and Mathematics, University of Plymouth. She has been involved in several funded projects, including FP7 GERYON (as scientific manager and principal investigator), COST Action QUALINET, FP7 ADAMANTIUM, and FP6 BIOPATTERN. She also led an industry funded project on multimedia over 3G networks. She has published one textbook, four book chapters, and more than 60 peer-refereed technical papers. She was the chair of QoE Interest Group of IEEE MMTC during from 2010 to 2012. Her current research interests include multimedia quality assessment, QoE control and management, VoIP/IPTV, and multimedia services for emergency communications and eHealthcare.

FBAR–Based Radio Frequency Bandpass Filter for 3G TD–SCDMA

Mingke Qi, Liangzhen Du, and Hao Zhang
(College of Precision Instrument and Opto–electronic Engineering, Tianjin University, Tianjin 300072, China)

Abstract

In this paper, we describe a high–performance TD–SCDMA bandpass filter based on film bulk acoustic resonator (FBAR) technology. The filter comprises a group of FBARs connected in a ladder configuration. Excellent quality factor greater than 1000 has been achieved at resonant frequency near 2 GHz for the FBAR. The TD–SCDMA FBAR filter has been fabricated and tested. The filter has low passband insertion loss of 1.7 dB and high stop–band rejection greater than 35 dB.

Keywords

FBAR; filter; TD–SCDMA; 3G

1 Introduction

Time–division synchronous code–division multiple access (TD–SCDMA) is a 3G wireless communication standard developed in China and adopted by ITU. It has high spectrum efficiency; it provides good system stability; and it lowers network construction costs. After many years of development, it is now at the stage of large–scale application.

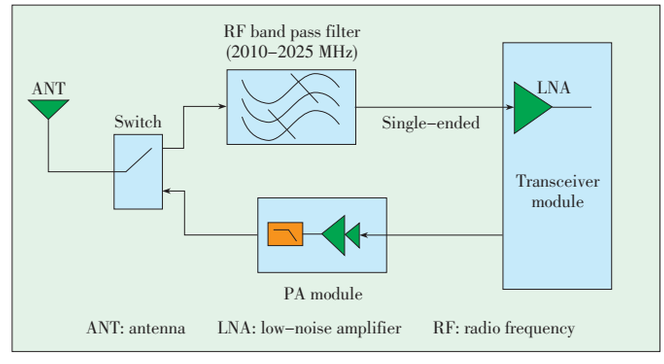
Fig. 1 shows the RF front–end of a TD–SCDMA system. A small, highly selective RF bandpass filter with low insertion loss is required before the low–noise amplifier (LNA) to select the frequencies of interest coming from the antenna.

One of the allocated operating frequency bands of TD–CDMA is 2010–2025 MHz with a bandwidth of 15 MHz. The RF bandpass filter is designed to allow the signal to propagate through the passband with low loss while blocking the out–of–band signal. The transition from passband to nearby stop–bands in the filter has to be sharp enough to minimize the interference from near–band emissions and harmonics. LC filters are not suitable here because of their slow roll–offs. Surface acoustic wave (SAW) filters use interdigitated (IDT) electrodes to produce acoustic resonances. For filters operating above 1 GHz, advanced photolithography and complicated etching techniques are required for SAW device fabrication. Film bulk acoustic wave resonator (FBAR) filters are preferred in gigahertz and higher–frequency applications. A typical FBAR is a three layer structure with a piezoelectric film sandwiched between two metal electrodes, and the sandwich structure is fabricated on a silicon substrate (Fig. 2). Longitudinal bulk acoustic waves are excited in the piezoelectric film by applying an RF electrical signal to the two metal electrodes. The resonant frequency of the device depends on the layer thicknesses, and for filters working at higher frequencies, thinner films are deposited for piezoelectric and metal layers. FBARs have higher quality (Q) factors than SAW resonators, and Q over 2000 in an FBAR at around 2 GHz has been reported [1]. Lower insertion loss could be achieved in a bandpass filter made of high–Q resonators. The bandwidth of the filter depends on the electromechanical coupling coefficient K_t^2 of the resonators. The narrower the bandwidth of the filter, the smaller the K_t^2 needed.

In this paper, we describe the design, fabrication, and testing of a TD–SCDMA bandpass FBAR filter. The filter has excellent passband insertion loss levels, fast transitions from passband to stop–bands, and deep stop–band rejections.

2 High–Performance FBAR

An FBAR can be described by the Mason model, which comprises one electrical port and two acoustic ports [2]. The electrical port is coupled to the acoustic ports through a transformer, which represents the electromechanical coupling between the electrical energy and acoustic energy in the piezoelectric material. The metal electrode is merely a mechanical material;



▲Figure 1. RF front–end of a TD–SCDMA system.

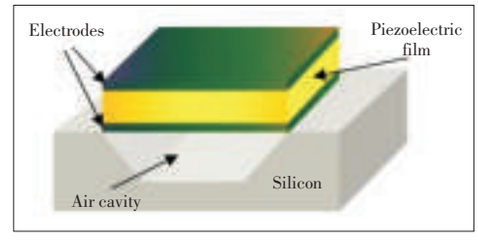


Figure 2.▶ An FBAR.

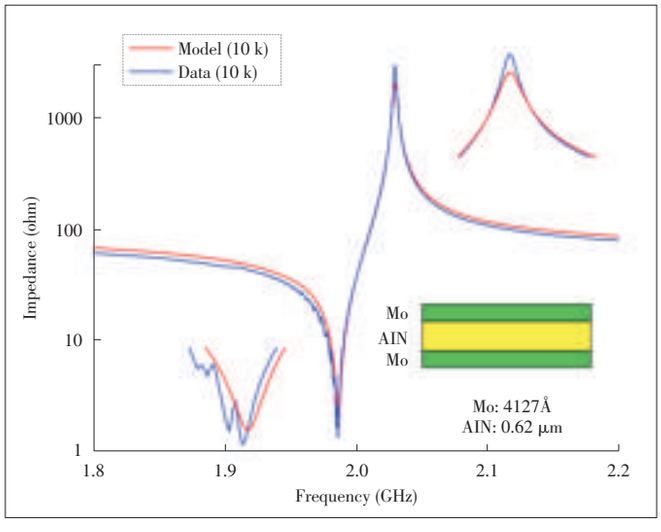
thus, in the model, it contains only two acoustic ports.

The filters and duplexers for UMTS and PCS bands have a center frequency around 2 GHz and a bandwidth of 60 MHz. However, the passband width of the TD–SCDMA filter is merely 15 MHz with a center frequency of around 2 GHz. The narrower bandwidth of the filter requires FBARs with smaller K_t^2. One approach to reducing K_t^2 is to reduce the thickness of the piezoelectric film and increase the thicknesses of both electrodes [3]. With increased electrode loading, the resonator becomes less efficient converting between the acoustic energy and electrical energy, and this leads to reduced K_t^2. Fig. 3 shows a cross section of an FBAR resonator for TD–SCDMA filter. It also shows the simulated and measured frequency responses of the FBAR with an area of 10,000 μm^2. The measured Qs, Qp, and K_t^2 of the resonator are 1878, 1034, and 5.2%, respectively. Table 1 shows the simulation and experimental data of the resonator. The simulated K_t^2 in the model matches the experimental results well. The experimental Qs and Qp are higher than those of the model because in the model, expected Q is conservatively defined.

3 Filter Design and Layout

A common topology for an FBAR filter is a group of series resonators and a group of shunt resonators connected in a ladder form (Fig. 4). The shunt resonators have a lower resonant frequency than the series resonators to enable bandpass transmitting characteristics. At much lower or higher frequencies than the resonant frequencies of the FBARs, the piezoelectric film in the resonators does not act as the energy–conversion body but purely play as a dielectric layer. The out–of–band attenuations are then determined by the voltage dividends in the capacitor network. Fig. 4 shows a selected TD–SCDMA filter topology. A group of series resonators and a group of shunt resonators are used, and the bonding wires from the filter chip to the in/out ports and ground are included as well. The layer thickness and resonator areas have been optimized to attain the desired passband insertion loss, fast transitions from passband to near–stop bands, and sufficient stopband attenuations. During optimization, the thicknesses of the top and bottom electrodes are set the same to keep the resonators symmetric. A symmetric FBAR structure does not allow a second mode of the resonator to be excited. The areas of the series resonators are much smaller than those of the shunt resonators after optimization because of the deep out–of–band attenuations (more than 40 dB) initially set to be reached.

Fig. 5 shows the layout of the filter chip. With all the bonding pads, the chip size is 1.0 × 0.9 mm. The bonding pads have been well–arranged to reduce coupling between the bonding wires.



▲Figure 3. Simulated and measured frequency responses of the FBAR with an area of 10,000 μm^2.

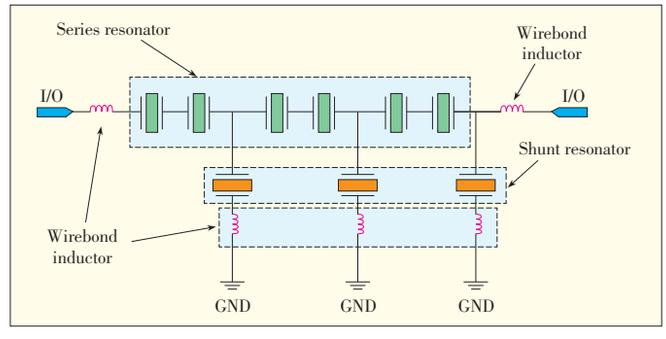
▼Table 1. Simulated and measured resonator parameters

Parameter	Simulated (Model)	Measured
Area (μm^2)	10,000	10,000
Rs (Ω)	1.6	1.3
Rp (Ω)	2066	3102
Qs	1099	1878
Qp	638	1034
K_t^2(%)	5.1	5.2

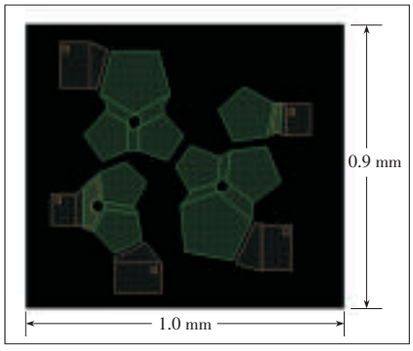
4 Measurement and Discussion

As the wafer fabrication and on–wafer test are completed, the wafer is singulated to individual dies. The good dies are identified and picked for laminate or board assembly.

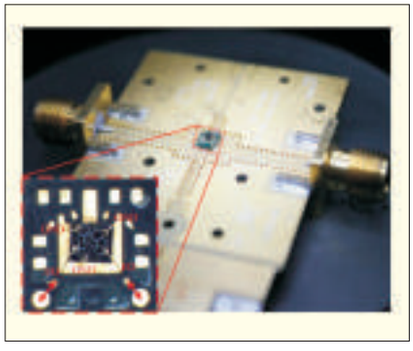
A filter die is assembled on a laminate using five bonding wires, two for in/out signals and three for ground pads (Fig. 6). The adjacent bonding wires should be perpendicular to each other to minimize mutual coupling. The laminate is then soldered on an evaluation board to make the filter measurement (Fig. 6). The measurement data and simulation results for the filter are plotted in Fig. 7. The simulation and measurement data match each other reasonably well; however, some transmission notches do not fit, which is caused by the inevitable mutu-



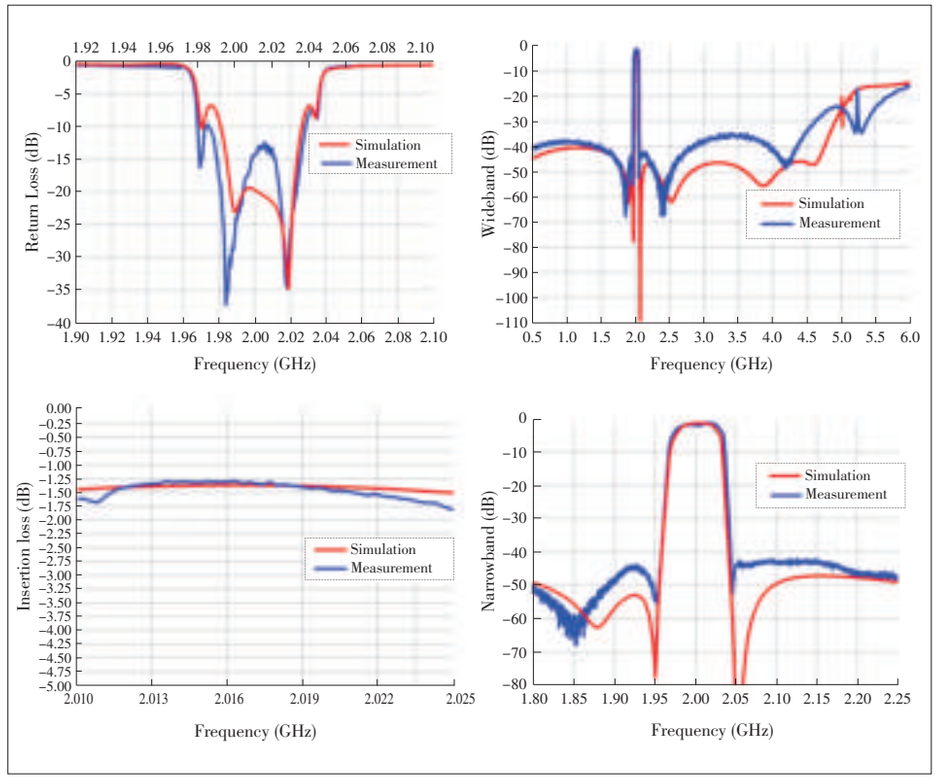
▲Figure 4. Topology of the ladder–type TD–SCDMA filter.



▲Figure 5. Layout of the ladder–type TD–SCDMA filter.



▲Figure 6. Assembled TD–SCDMA filter.



▲Figure 7. Measured and simulated TD–SCDMA filter data.

al couplings between bonding wires.

A low passband insertion loss of 1.7 dB has been measured. The return loss in the passband is better than –12 dB (VSWR is better than 1.7). The filter provides signal attenuation greater than 35 dB from 0.5 to 4.0 GHz. Very fast transitions from passband to near–stop bands have been achieved because of the high–Q FBAR resonator.

5 Conclusion

A high–performance FBAR bandpass filter for TD–SCDMA wireless communication system has been presented. The electrode thickness of the FBAR is increased to reduce K_t^2, which is crucial for constructing the TD–SCDMA filter with relatively narrow bandwidth. The FBAR has been measured and has very high Q (greater than 1000) at a high frequency of 2 GHz. The fabricated TD–SCDMA filter based on high–performance FBARs has low passband insertion loss of 1.7 dB and high stop–band rejection greater than 35 dB. The filter was fabricated in the MEMS Lab of Tianjin University.

References

[1] R. Ruby, R. Parker, and D. Feld, "Method of Extracting Unloaded Q Applied Across Different Resonator Technologies," *IEEE Ultrasonics Symposium*, 2008, pp.1815–1818.
[2] J. Rosenbaum, *Bulk Acoustic Wave Theory and Devices*, Artech House, Boston, 1988.
[3] H. Zhang, W. Pang, W. Chen, and C. Zhou, "Design of unbalanced and Balanced Radio Frequency Bulk Acoustic Wave Filters for TD–SCDMA," *Microwave and Millimeter Wave Technology (ICMMT)*, 2010, pp.878–881.

Manuscript received: July 10, 2012

Biographies

Mingke Qi

Mingke Qi (mingkeqi@gmail.com) received his BS degree from Tianjin University in 2011. He is currently pursuing his MS degree in instrument science and technology Tianjin University. His research interests include bulk acoustic wave device modeling and characterization, FBAR filter and duplexer design, and RF circuit design.

Liangzhen Du

Liangzhen Du (du.liangzhen@zte.com.cn) received his BS degree from Tianjin University in 1967. Professor Du's research interests are fiber optic communications, timing and frequency control devices, and MEMS technology applications. He completed three National 863 Plan projects and was the first in China to industrialize and mass produce EDFA and fiber optic isolators. He was deemed to be an "advanced individual who made important contributions to the National 863 Plan" by the Ministry of Science and Technology, China. He also won six Science and Technology Progress Awards from Shenzhen government and other provincial governments. He has published approximately 40 papers.

Hao Zhang

Hao Zhang (haozhang@tju.edu.cn) received his MS degree in physics and Ph.D. degree in electrical engineering from the University of Southern California, Los Angeles, in 2002 and 2006. He is currently a professor in the College of Precision Instrument and Optoelectronics Engineering, Tianjin University. His research interests are fields of RF/microwave MEMS, MEMS sensors, microfluidics, and biochips.

Data Center Network Architecture

Yantao Sun, Jing Cheng, Konggui Shi, and Qiang Liu
(School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China)

Abstract

The rapid development of cloud computing has created significant challenges in data center architecture. In this paper, we discuss these challenges. We introduce the latest research on data center network architecture, especially in terms of structure and virtual machine migration. We also introduce research in areas related to network architecture. Finally, we suggest future research areas in data center networks.

Keywords

data center network; network architecture; network topology; virtual machine migration

1 Introduction

The history of data centers can be traced back to the 1960s. Early data centers were deployed on mainframes that were time–shared by users via remote terminals. The boom in data centers came during the internet era. Many companies started building large internet–connected facilities, and these were called internet data centers (IDCs) [1]. In 2006, Google first proposed cloud computing; and later, Amazon, Microsoft, Yahoo, IBM, and other IT companies put great effort into promoting it. Cloud computing requires data center networks (DCNs) to be scalable, flexible, powerful, and energy–efficient.

A large–scale network and virtual machine (VM) migration are the main features of today's data centers, and cloud computing is the most important service in data centers. There are many problems in data centers that researchers have been trying to solve. Research on DCNs has become very important in the field of computer networks. Every year since 2008, SIGCOMM and INFOCOM have both included special sessions to discuss research on data DCNs.

Some papers have been written on the problems of current data centers. In 2009, Krishna Kant introduced state–of–the–art DCN technologies and discussed storage, networking, management, power, and cooling in data centers [2]. In the same year, Albert Greenberg et al. described costs in data centers and methods to reduce these costs [3]. In particular, they pointed out that conventional network architecture lacks agility, and they discussed principles that should be followed to design an agile new architecture.

Especially after 2008, DCN technologies have developed rapidly, and much innovative research has been done on network architecture and protocols, QoS, VM migration, and configuration and management. Kant and Greenberg's works do not cover the latest research.

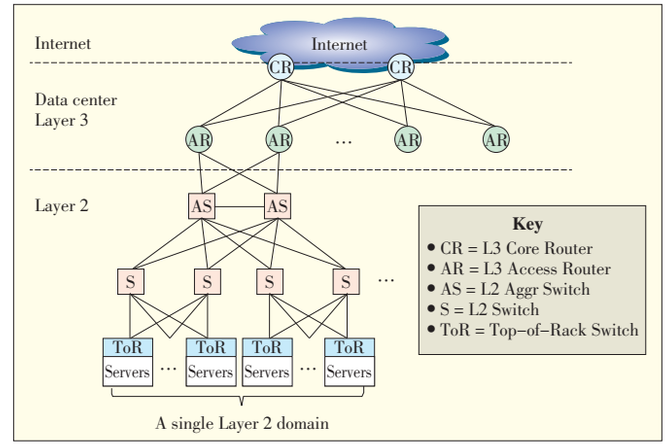
In this paper, we introduce the latest research on DCN architecture, including research on network structure and VM migration solutions. In section 2, we discuss existing problems in current DCNs. In section 3, we compare network architectures proposed in recent years. In section 4, we review the latest solutions for migrating VMs over the entire data center. In section 5, we introduce research related to DCN architecture. Section 6 concludes the paper.

2 Issues in Existing DCNs

A multiroot tree structure is commonly used in today's data centers. In such a structure, many layer–2 domains are connected via layer–3 networks. Fig. 1 shows a conventional DCN architecture [4], [5]. Many server nodes connected via switches constitute a layer–2 domain. In practice, a single layer–2 domain is limited in size to about 4000 servers because of the need for rapid convergence when there is a failure. Furthermore, a layer–2 domain is divided into subnets by using virtual local area networks (VLANs). Each VLAN has no more than a few hundred servers, because the overhead of the broadcast traffic, for example, address resolution protocol, limits the size of an IP subnet.

There are two salient problems that prevent conventional architecture from supporting a large–scale data center with up to tens of thousands of servers at one site.

The first of these problems is a shortage of bandwidth in the



▲Figure 1. Conventional DCN architecture.

higher layers. In practice, the typical oversubscription ratio between neighboring layers is 1:5 or more. In the top layer, this ratio may reach 1:80 to 1:240. Even if the fastest, most advanced switches and routers are used, only 50% of the aggregation bandwidth of edge networks can be supported in the top layer [3]. The top layer is therefore becoming the bottleneck of the entire network, especially in today's cloud computing environment where the requirement for intra-network traffic is increasing rapidly.

The second problem is that VM migration is limited in a single layer-2 domain; that is, a VM cannot move from one layer-2 domain to others. VM migration is a very important feature of cloud computing data centers. By leveraging VM migration, a data center can save energy [6], [7], improve scalability and reliability [8], and rapidly deploy services [9]. In conventional networks, different layer-2 domains have different IP ranges, so a VM has to change its IP address when it migrates to other layer-2 domains. In many applications, service cannot be interrupted during VM migration, and this requires the VM's IP address to remain unchanged, even when the VM migrates to another domain. This is an urgent dilemma that has to be solved in new data center architectures.

With the rapid development of cloud computing, the demand for large, centralized data centers has become urgent worldwide. New network architectures are required because existing architectures don't work well. There are many other issues related to DCNs, including QoS, routing protocols, and network configuration, but we do not broach them in this paper.

3 Network Structure

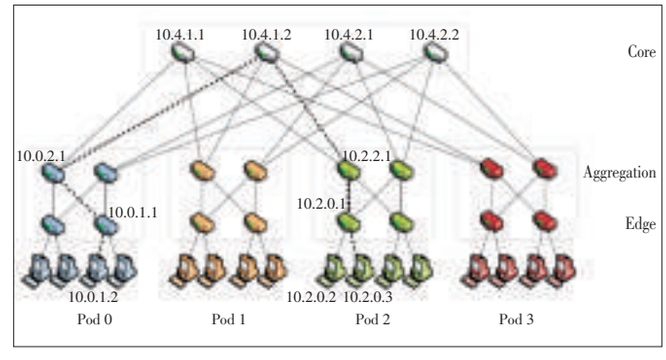
To solve bandwidth and scalability problems, researchers have proposed many novel DCN structures over the past several years. These structures can support up to tens of thousands of servers without any bandwidth bottleneck. Generally, these network structures can be categorized as switch-centric, server-centric, or irregular. In a switch-centric network, switches are the fundamental components of the network fabric; servers are attached to access switches and are the leaves of network. In a server-centric network, servers provide both computing and routing and are the main network components. Unlike irregular networks, switch-centric networks and server-centric networks both have regular, symmetrical topology. An irregular network has an arbitrary topology.

3.1 Switch-Centric Networks

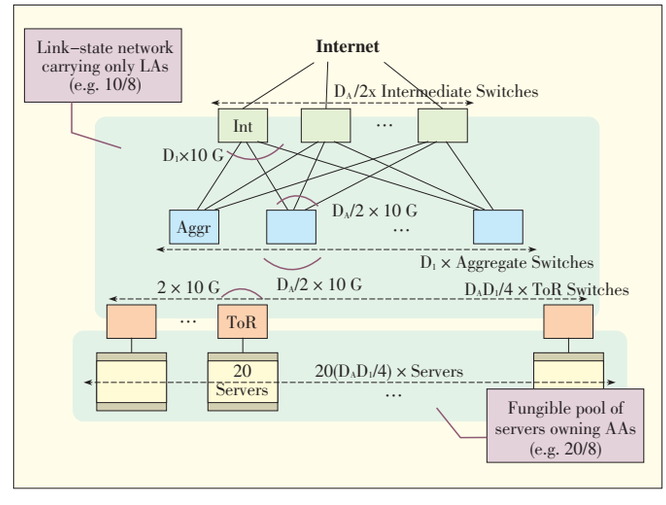
In 2010, a fat-tree network was proposed at SIGCOMM [10] (Fig. 2). A fat-tree network is divided into core layer, aggregation layer, and edge layer, and all the servers are connected to the switches in the edge layer. A fat-tree network is a multipath network in which there are many equal-cost paths between adjacent layers. It is also non-blocking and can have an oversubscription ratio of up to 1:1. Therefore, it eliminates the bandwidth bottleneck in the core layer. Furthermore, a fat-tree topology can support large-scale networks with tens of thousands of physical servers. Using 48-port switches, it can contain up to 27,648 servers with 2280 switches. Because a fat-tree network can be constructed using cheap Ethernet switches, it costs less than a conventional network. To leverage the vast bandwidth of multiple paths, a novel routing method is used in a fat-tree network.

To support non-blocking communication, the fat-tree architecture requires a large number of switches in the core and aggregation layers as well as wires interconnecting these switches. These make the fat-tree network very expensive, energy intensive, and complicated to manage. In VL2 [11], Helios [12], and c-Through [13] topologies, cheap core and aggregation switches are replaced with expensive, high-speed switches.

VL2 was proposed by Microsoft in 2009. In VL2, a Clos network is used to build the DCN (Fig. 3). Other than 1 Gbit/s switches in fat-tree, VL2 leverages 10 Gbit/s switches in the core and aggregation layers, so the link speed between core and aggregation layers is 10 times faster than in fat-tree. Also, the number of links required between the core and aggregation layers is only 10% that required in the fat-tree. One weakness



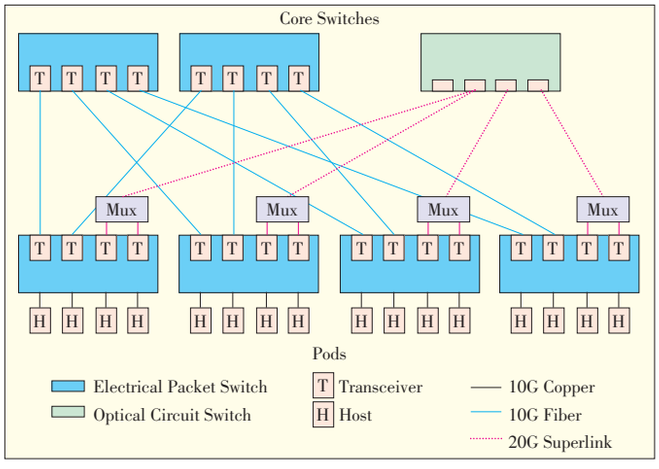
▲Figure 2. Fat-tree network topology.



▲Figure 3. VL2 topology.

of VL2 is that the maximum supported network size is only half that supported by fat-tree.

Helios is a hybrid electrical/optical switch architecture [12] (Fig. 4). In Helios, core switches are replaced with optical circuit switches, and copper cables are replaced with optical fibers to connect the pod switches to core switches. Circuit switches are used to deliver baseline, slowly changing inter-pod communication. Packet switches are used to deliver bursty inter-pod communication. Another hybrid electrical/optical DCN is called c-Through [13]. In c-Through, the entire network comprises an electrical packet-switched network and an optical circuit-switched network. The packet-switched network uses a traditional hierarchy of Ethernet switches arranged in a tree, and the circuit-switched network connects the top-of-rack switches. The optical circuit switch automatically reconfigures circuits between top-of-rack switches to achieve the maximum throughput. To make the best use of high-capacity circuits, servers buffer traffic in order to collect sufficient volumes for high-speed transmission.



▲Figure 4. Helios topology.

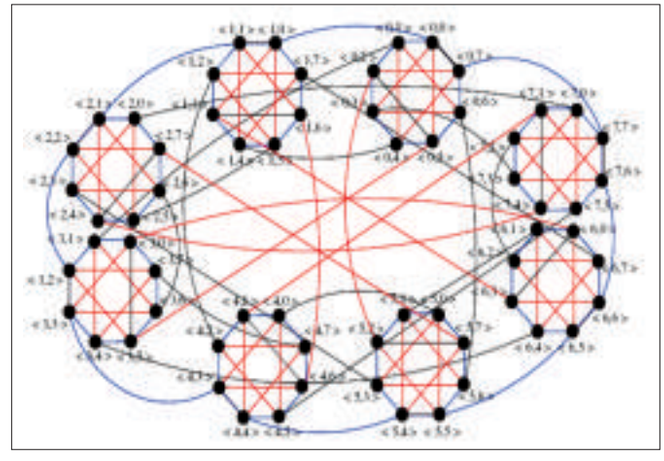
The several previously-mentioned architectures are all based on the multi-root tree structure. Conversely, HyScale [14] is a non-tree structure that has high scalability and that uses hybrid optical networks (Fig. 5). It uses optical burst switching for transmitting low volumes of data and optical circuit switching for transmitting high volumes of data in a data center. HyScale is a recursively defined topology denoted $\Psi(k, \Phi, T)$, where k is the number of levels in the topology, T is an integer, and Φ is the address space of all nodes in Ψ [14]. A HyScale is constructed by connecting T of $k-1$ HyScales, that is, $\Psi(k-1, \Phi, T)$ [14].

In 2011, we proposed MatrixDCN, another non-tree structure [15]. In MatrixDCN, a switch may be a row switch, a column switch, or an access switch. Access switches are deployed as a matrix with rows and columns. An 8×8 matrix has 8 rows, 8 columns, and 64 access switches. A row switch is deployed at the head of one row and links all the access switches in the row. A column switch it is deployed at the head of a column and links all the access switches in the column. Fig. 6 shows a 2×2 MatrixDCN. With 48-port switches, MatrixDCN can support up to 100,000 servers without bandwidth bottleneck. This fabric is simple and extendable, and its routing is very effective. Furthermore, the fabric supports one-to-many and many-to-many traffic in cloud computing.

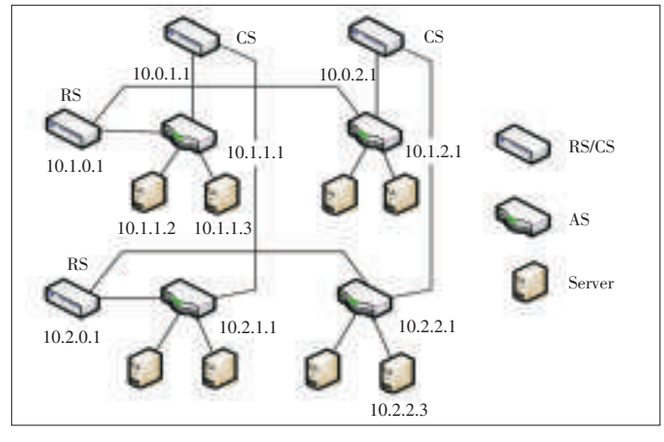
3.2 Server-Centric Architectures

DCell is a recursively defined architecture that uses servers with multiple network ports (Fig. 7)[16]. A high-level DCell is constructed from low-level DCells, and low-level DCells are connected together via links between servers. A DCell can scale exponentially with the server node degree. Therefore, a DCell with a small server node degree can support up to several million servers. However, DCell has a low bisection bandwidth that may lead to traffic jam in the network.

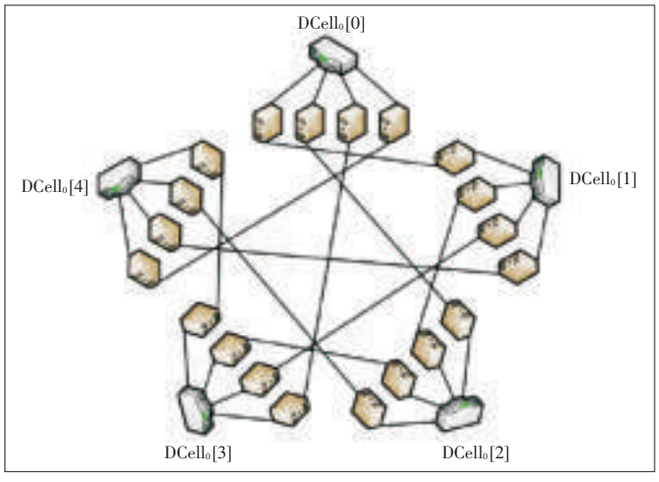
FiConn shares the same design principle with DCell. The network is constructed by giving the interconnection ability to servers [17]. Unlike in a DCell, the server node degree in a k-level DCell is $k+1$; however, that of FiConn is always two.



▲Figure 5. HyScale Topolgy [14].



▲Figure 6. MatrixDCN topology.



▲Figure 7. DCell topology.

Today's commodity servers usually have two Ethernet ports—one for network connection and one for backup. FiConn only uses the existing backup port for interconnection; no other hardware cost needs to be incurred by adding to the servers.

BCube provides more bandwidth in the top layer than DCell [18] (Fig. 8). BCube comprises multiport servers and switches that only connect with servers. A BCube$_k$ has $N = n^{k+1}$ servers and $k + 1$ levels of switches. Each level has n^k n–port switches. BCube can also support very large networks. With 3–port servers and 48–port switches, a data center can be constructed that contains more than 100,000 servers.

DPillar comprises n–port switches and dual–port servers [19] (Fig. 9). The servers are arranged into k columns and so are the switches. Visually, the topology looks like the $2k$ columns of servers and switches that are attached to the cylindrical surface of a pillar. A server in each server column is connected to two switches in the two neighboring switch columns.

An expansible DCN structure using hierarchical compound graphs has also been proposed [20]. The structure is called bi–dimensional compound network (BCN), and compared with the previously mentioned structures, it is more complicated. Like DCell, it does not eliminate traffic bottleneck.

3.3 Irregular Networks

Most DCN architectures have a regular symmetric topology. However, an asymmetric data center topology called Scafida has been proposed [21]. It is inspired by the scale–free Barabási and Albert topologies [21]. In Scafida, the network structure is generated iteratively according an algorithm. The nodes are added one by one to the network. A new node is attached probabilistically to an existing node proportional to the existing node's degree. New nodes have more than one link, so they are attached to several existing nodes. In Scafida, a node's degree (links) is limited to the number of its ports.

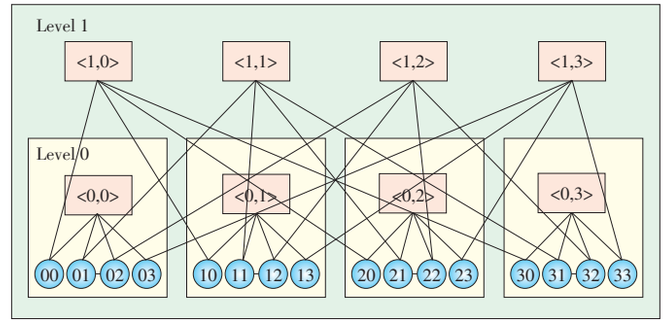
Existing solutions to data center scalability require the network architecture to be changed. However, a scheme called traffic–aware virtual machine placement has been proposed to improve network scalability [8]. It works by optimizing the placement of virtual machines without changing the network structure.

REWIRE is a framework for designing, upgrading, and expanding DCNs [22]. In REWIRE, unstructured networks are built instead of topology–constrained networks, which are found in most existing data centers. It uses local search to find a network that maximizes bisection bandwidth while minimizing latency and satisfying a large number of user–defined constraints. Demonstrations have shown that arbitrary topologies can boost DCN performance and reduce expenditure on network equipment.

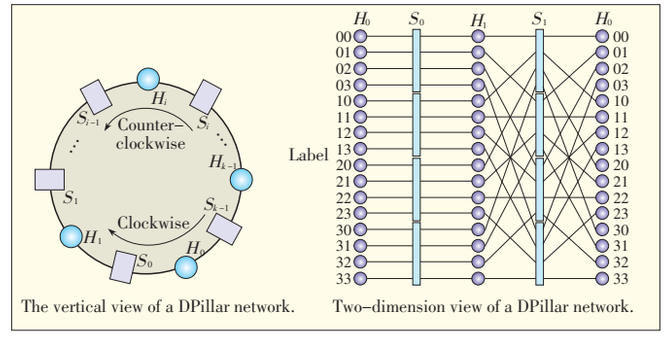
3.4 Other Structures

Containerizing is an important trend in data centers. In 2007, Microsoft proposed using standard shipping container to modularize data centers [23]. A data center module comprising more than 1000 pieces of equipment can be built in a shipping container with full networking support and cooling. Each module includes networking gear, compute nodes, and persistent storage. The modules are self–contained with enough redundancy so that individual failed systems do not need to be replaced. A large data center container is packed with 1k to approximately 4k servers.

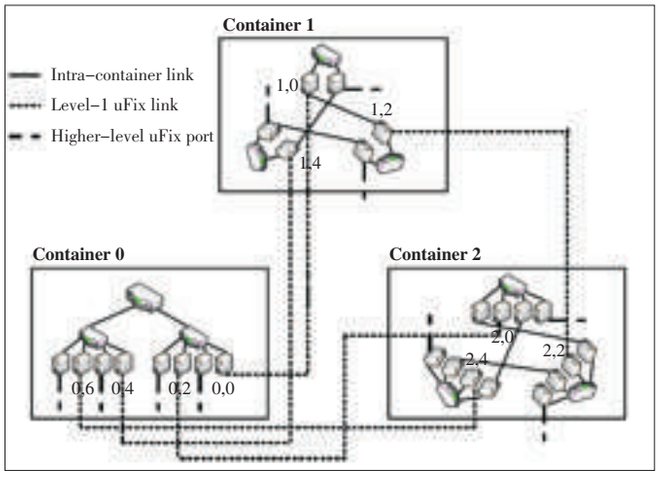
The uFix proposed in [24] is a scalable and modularized architecture that interconnects heterogeneous data center containers (Fig. 10). Every container can have a different struc-



▲Figure 8. BCube topology.



▲Figure 9. DPillar topology.



▲Figure 10. uFix topology.

ture, such as a fat-tree or BCube structure. In uFix, each server in a container reserves an NIC port for intercontainer connection. The uFix is defined iteratively; that is, a level l uFix domain comprises a number of level / – 1 uFix domains. A server with a level / uFix link is called a level / uFix proxy. It routes between containers.

Wireless technology is also used in DCNs. In a typical data center, a data center rack comprises 40 servers connected to a top-of-rack (ToR) switch with 1G links. The ToR switch is connected to the aggregation switch via a 10G link. Thus, the links from ToRs to aggregation switches are oversubscribed by a ratio of 1:4. These up-links are the potential hotspots that hinder network performance. Rather than adding wired links to the network, multigigabit wireless links have been proposed to provide additional bandwidth [25]. Each ToR switch has one or more 60 GHz wireless device with electronically steerable directional antennas. A central controller monitor switches the beams of the wireless devices to set up flyways between ToR switches. These flyways provide added bandwidth as needed.

3.5 Comparisons

Most of today's data centers are based on switch-centric architectures. Although their scalability and flexibility is not good enough, switch-centric architectures have inherent advantages: They are similar to traditional network architectures, so it is easier to upgrade traditional switches to support these new architectures. Most network components and protocols can be directly used in new architectures or can be used with slight modification. Server-centric architectures eliminate switch restrictions so that routing and scaling up is easier. New features and functions can be flexibly added on the servers. However, the intricate network topology, bandwidth bottleneck, lower packet-routing speed, and occupation of the server resource are all drawbacks.

An arbitrary, irregular network structure is flexible and has good scalability, but it is unlikely to be applied in a large data center because it is very difficult to manage and maintain.

4 VM Migration

In section 3, we introduced state-of-the-art solutions to network scalability. With these new network structures, large networks with tens to hundreds of thousands of servers can be built out. To provide the huge bandwidth needed in data centers, multiple paths are deployed between any pair of servers. To fully use these paths, layer-3 routing is used in these solutions.

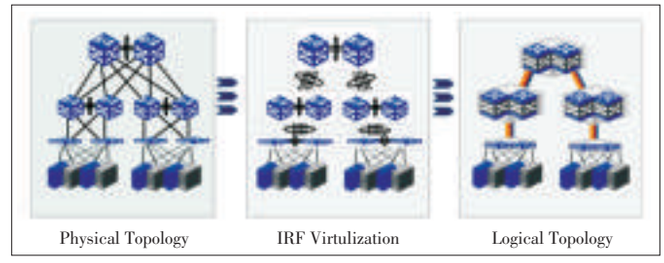
Layer-3 routing limits the VM migration within a single layer-2 domain. However, to take full advantage of virtualization, it is desirable that VM can migrate anywhere in the data center. That means that the entire DCN looks like a single layer-2 subnet, at least from the perspective of end users, that is, VMs. To solve this problem, big layer-2 network solutions have been proposed.

4.1 Device Virtualization

One such solution is device virtualization technology. Multiple switches are virtualized into one logical switch, and multiple links are aggregated into one link. H3C's IRF technology [26] (Fig. 11) and the Cisco's VSS technology are examples of device virtualization technology. With such technology, a multiroot tree with loops is re-formed into a simple one-root tree without loops, and all the links are fully utilized. Otherwise, STP blocks the redundant links to eliminate loops in network. Aggregation technologies are relatively mature and have been used in practice. However, they use private enterprise protocols that have poor interoperability; they are difficult to automatically configure; and their supported networks are not very large.

4.2 Layer-2 Routing

Device virtualization technology can be used only in small or medium-sized data centers. To support large networks, layer-2 routing has been proposed. In the layer-2 switching, layer-3 routing technology, such as TRILL [27] and Cisco's FabricPath [28], is applied. TRILL is an IETF standard that is used in devices called RBridges (routing bridges) or in TRILL Switches, which provide multipath forwarding for Ethernet frames. TRILL applies an IP routing mechanism in the Ether-



▲Figure 11. IRF-based network without STP.

net frames' forwarding. In TRILL, RBridges compute the shortest path and equal-cost paths in layer-2 by using TRILL IS-IS, which is a link-state routing protocol similar to IS-IS routing protocol. MAC-in-MAC packets are forwarded to the destination host via the switched network comprising RBridges. FabricPath is a similar (but private) technology provided by Cisco.

Another layer-2 routing solution is SEATTLE, which uses link-state routing protocol to establish a routing path between switches [29]. Unlike TRILL, SEATTLE uses the global switch-level view provided by a link-state routing protocol to form a one-hop DHT. This DHT stores the IP to MAC mapping and MAC to host location mapping of each host in switches. SEATTLE converts an ARP request into a unicast-based message to obtain the destination host's MAC address. It then determines its location according to the MAC address.

4.3 Offline Routing

Layer-2 routing technologies put no requirements on the network structure and can be used in large DCNs. However, they import a routing protocol into the layer-2 network. Such a protocol is difficult to implement on switches and increases the complexity of the control plane. New switches should be developed for the new layer-2 routing protocol. SPAIN [30] and NetLord [31] demonstrate new thinking about layer-2 interconnection based on existing switch devices in an arbitrary topology. With these methods, a set of paths is pre-computed offline for each pair of source-destination hosts by exploiting the redundancy in a given network topology. Then, these paths are merged into a set of trees, and each tree is mapped onto a separate VLAN. In this way, a proxy application is installed on the hosts, and the proxy chooses several VLAN paths to transmit packets to the destination host. The advantage of this is that multipath is implemented, and routing load is balanced on multiple paths in an arbitrary topology. Its drawbacks are inflexibility to changes in topology and required modifications to hosts.

4.4 Topology-Aware Routing

The previously mentioned technologies are versatile and are applicable to any network structure; however, specialized routing protocols are required so that they can learn the network topology. In contrast, PortLand is a big layer-2 network solution specifically for the fat-tree network [32]. It uses a lightweight location discovery protocol (LDP) that allows switches to discover their location in the topology. In PortLand, every end host is assigned an internal pseudo MAC (PMAC) that encodes the location of the end host. Compared with the other layer-2 technologies, PortLand leverages the information of network structure within layer-2 routing.

4.5 Discussion

We have introduced some big layer-2 network solutions for VM migration and discussed their features. Each solution has its own advantages and disadvantages and is suitable for certain environments. More research has to be done to find more general,

better-performing solutions for VM migration.

Ideally, a big layer-2 solution for VM migration should be simple and efficient. It should be easy to implement and should involve less importation of new technologies and less device modification. Efficiency implies rapid forwarding with less overhead. Because regular topologies are used in most data centers, leveraging the topology's regularity simplifies the VM migration solution and makes it more efficient. It is better if such a topology-centered solution can be applied to any topology with some regularity. The previously-mentioned solutions blend packet routing with VM migration, which makes them more complicated. We suggest VM migration should be separated from network routing, and overlay on the network routing, like NVO3 [33]. NVO3, VXLAN [34], and NVGRE [35] are some multitenancy solutions for data centers. They can be used to solve VM migration problems as well. However, because these solutions are not very mature, we will not discuss them here.

5 Related Works

As well as research on the network architecture itself, other research on architecture-related areas such as network performance, energy-saving, and configuration has been done.

In [36], the performance of FiConn [17] and fat-tree network architectures are compared through experimentation. In these experiments, a three-tier transaction system is deployed on the two types of networks. The results show that FiConn performs better than fat-tree in terms of throughput because the traffic between two virtual machines must pass through the upper switches in fat-tree. Fat-tree results in better network reliability and stability. When routing nodes break down in a FiConn network, network performance declines significantly.

Some works leverage the network structure to improve overall network performance. In [37], a VM migration scheme is proposed to avoid network overload caused by VM migration. Inter-VM dependencies and underlying network topology are incorporated into VM migration decisions. In [38], a source-to-receiver expansion approach based on regular topology is used to build efficient multicast trees and routing.

In some works, the network is slightly modified to save energy. ElasticTree continually monitors data center traffic and calculates a subnet that covers all the traffic and meets network performance and fault tolerance targets [39]. Then, it powers down the other unneeded links and switches to save energy. Honeyguide saves energy by means of VM migration [7]. It moves the VMs together in order to increase the number of unused servers and then powers down these servers and related unused switches and links. To improve network fault tolerance,

bypass links are added between upper-tier switches and physical servers.

In [40], a generic and automatic address configuration system for data centers is proposed. In [41], a new layer-2 for data centers is proposed. This layer comprises interconnected policy-aware switches, and middleboxes, such as firewalls and load balances, into those switches that are off the network path.

Much research has been done on all aspects of DCNs; however, we do not introduce it all here.

6 Conclusion

Cloud computing services have created new challenges in data centers. Next-generation data centers will require large networks with more internal bandwidth. Moreover, data centers will need to support free VM migration across the entire DCN. These features will require DCNs to have new architectures.

In this paper, we have described the latest research on DCN architecture. We have classified these architectures and determined their features and differences. This paper is intended to inform readers about the newest research in DCN architecture so that breakthroughs may be made in this field.

Many kinds of architectures have been proposed to solve problems in existing data centers. These architectures have their own advantages, and different data centers use different architectures depending on their supported applications. We suggest further research on the general routing method compatible with different network architectures. This method should fully capitalize on the regular network topology. We also suggest further research on separating VM migration from network routing. Finally, we suggest further research into areas such as QoS and VM migration policies that are related to topology.

Acknowledgements

We thank Xiaoli Song and Bin Liu of ZTE Inc. for their great support and help to this paper. This work is supported by the ZTE-BJTU Collaborative Research Program under Grant No. K11L00190 and the Fundamental Research Funds for the Central Universities under Grant No. K12JB00060.

References

[1] Wang Qingbo, Virtualization and Cloud Computing (in Chinese), Publishing House of Electronics Industry, Oct. 2009.
[2] Krishna Kant, "Data center evolution: A tutorial on state of the art, issues, and challenges," *Computer Networks*, vol.53 no. 17, pp. 2939-2965, Dec. 2009.
[3] Albert Greenberg, James Hamilton, David A. Maltz, Parveen Patel, "The cost of a cloud: research problems in data center networks," *ACM SIGCOMM Computer Communication Review*, vol. 39, no.1, pp. 68-73, 2009.
[4] Cisco Inc., "Data center: Load balancing data center services," *Solutions Reference Network Design*, Mar. 2004.
[5] Cisco Technical Report: New Trends Affect the Architecture of Data Center Networks (in Chinese), 2008.
[6] Kim Khoa Nguyen, Mohamed Cheriet, Mathieu Lemay, Victor Reijs, Andrew Mackarel, Alin Pastrama, "Environmental-aware virtual data center network," *Computer Networks*, vol.56, no. 10, 2538-2550, July 2012.
[7] Hiroki Shirayanagi, Hiroshi Yamada, Kenji Kono, "Honeyguide: A VM Migration-Aware Network Topology for Saving Energy Consumption in Data Center Networks," in *IEEE Symposium on Computers and Communications (ISCC 2012)*, pp.460-467, July 2012.
[8] Xiaoqiao Meng, Vasileios Pappas, Li Zhang, "Improving the Scalability of Data Center Networks with Traffic-aware Virtual Machine Placement," San Diego, in *INFOCOM'10*, pp.1-9, Mar. 2010.
[9] Mohammad Hajjat, Xin Sun, Yu-Wei Eric Sung, David Maltz, Sanjay Rao, Kunwadee Sripanidkulchai, Mohit Tawarmalani, "Cloudward Bound: Planning for Beneficial Migration of Enterprise Applications to the Cloud," in *the ACM Special Interest Group on Data Communication (SIGCOMM'10)*,New Delhi, pp.243-254, Aug. 2010.
[10] M. Al-Fares, A. Loukissas, and A. Vahdat, A Scalable, "Commodity Data Center Network Architecture," in *the ACM Special Interest Group on Data Communication (SIGCOMM'08)*, Seattle, pp.63-74, Aug. 2008.
[11] A. Greenberg et al., "VL2: A Scalable and Flexible Data Center Network," in *the ACM Special Interest Group on Data Communication (SIGCOMM'09)*, Barcelona ,PP.51-62, Aug. 2009.
[12] Nathan Farrington, George Porter, Sivasankar Radhakrishnan, Hamid Hajabdolali Bazzaz, Vikram Subramanya, Yeshaiahu Fainman, George Papen, Amin Vahdat, "Helios: A Hybrid Electrical/Optical Switch Architecture for Modular Data Center," in *the ACM Special Interest Group on Data Communication (SIGCOMM'10)*,New Delhi, pp.339-350, Aug. 2010.
[13] Guohui Wang, David G. Andersen, Michael Kaminsky, Konstantina Papagiannaki, T. S. Eugene Ng, Michael Kozuch, Michael Ryan, "c-Through: Part-time Optics in Data Centers," in *the ACM Special Interest Group on Data Communication (SIGCOMM'10)*, New Delhi, pp.327-338, Aug. 2010.
[14] Shivashis Saha, Jitender S. Deogun, Lisong Xu, "HyScale: A Hybrid Optical Network based Scalable, Switch-centric Architecture for Data Centers," in *IEEE ICC'12*, Ottawa, pp.2967-2971, June 2012.
[15] Yantao Sun, Xiaoli Song, Bin Liu, Qiang Liu, Jing Cheng, MatrixDCN: A New Network Fabric for Data Centers[online], Available:http://tools.ietf.org/html/draft-sun-matrix-dcn-00
[16] Chuanxiong Guo, Haitao Wu, Kun Tan, Lei Shiy, Yongguang Zhang, Songwu Lu, "DCell: A Scalable and Fault-Tolerant Network Structure for Data Centers," in *the ACM Special Interest Group on Data Communication (SIGCOMM'08)*, Seattle, pp.75-86, Aug. 2008.
[17] Dan Li, Chuanxiong Guo, Haitao Wu, Kun Tan, Yongguang Zhang, Songwu Luy, "FiConn: Using Backup Port for Server Interconnection in Data Centers," in *the IEEE Conference on Computer Communications (INFOCOM 2009)*, pp. 2276-2285, Apr. 2009.
[18] Chuanxiong Guo et al., "BCube: A High Performance, Server-Centric Network Architecture for Modular Data Centers," in *the ACM Special Interest Group on Data Communication (SIGCOMM'09)*, Barcelona,pp.63-74, Aug. 2009.
[19] Yong Liao, Dong Yin, Lixin Gao, "DPillar: Scalable Dual=port server interconnection for data center networks," in *Proc. 19th Int. Conf. Comp. Commun. Netw. (ICCCCN'10)*, ZTH Zurich, Aug. 2010.
[20] Deke Guo et al., "BCN: Expansible Network Structures for Data Centers Using Hierarchical Compound Graphs," *INFOCOM'11*, Shanghai, PP.61-65, Apr. 2011.
[21] László Gyarmati, Tuan Anh Trinh, "Scafida: A Scale-Free Network Inspired Data Center Architecture," in *Conf.ACM SIGCOMM Computer Communication Review*, vol.40, no. 5, pp.5-12, Oct. 2012.
[22] Andrew R. Curtis, Tommy Carpenter, Mustafa Elsheikh, Alejandro L'opez-Ortiz, S. Keshav, "REWIRE: An Optimization-based Framework for Unstructured Data Center Network Design," in *31st Annual International Conf. on Computer Communications IEEE INFOCOM*, Orlando, pp.1116-1124, Mar. 2012.
[23] J. R. Hamilton, "An Architecture for Modular Data Centers," *CIDR 2007*, Asilomar, pp.306-313, Jan. 2007.
[24] Dan Li, Mingwei Xu, Hongze Zhao, Xiaoming Fu, "Building Mega Data Center from Heterogeneous Containers," in *the nineteenth IEEE International Conference on Network Protocols(ICNP)2011*, Vancouver, pp.256-265, Oct. 2011.
[25] Daniel Halperin, Srikanth Kandula, Jitendra Padhye, Paramvir Bahl, David Wetherall, "Augmenting Data Center Networks with Multi-Gigabit Wireless Links," in the *Special Interest Group on Data Communication (SIGCOMM'11)*, pp 38-49, Aug. 2011.
[26] H3C IRF Technical Architecture White Paper [Online]. Available: http://www.h3c.com.cn/download.do?id=1155459
[27] RFC 5556, "Transparent Interconnection of Lots of Links (TRILL): Problem and Applicability Statement"
[28] Cisco, FabricPath [Online]. Available: http://www.cisco.com/en/US/netsol/ns1151/index.html
[29] M. C. Changhoon Kim and J. Rexford, "Floodless in SEATTLE: A Scalable Eth-

ernet Architecture for Large Enterprises," in *the ACM Special Interest Group on Data Communication (SIGCOMM'08)*, Seattle, pp. 3–14, Aug. 2008.
[30] Jayaram Mudigonda, Jayaram Mudigonda, "SPAIN: COTS Data-Center Ethernet for Multipathing over Arbitrary Topologies," in *7th USENIX Symposium on Networked Systems Design and Implementation*, Vancouver, pp. 18–19, Mar. 2010.
[31] Jayaram Mudigonda, Praveen Yalagandula, "NetLord: A Scalable Multi-Tenant Network Architecture for Virtualized Datacenters," in *the ACM Special Interest Group on Data Communication (SIGCOMM'11)*, Toronto,pp.62–73, Aug. 2011.
[32] Radhika Niranjan et al., "PortLand: A Scalable Fault-Tolerant Layer 2 Data Center Network Fabric," in *the ACM Special Interest Group on Data Communication (SIGCOMM'09)*, Barcelona , PP.39–50, Aug. 2009.
[33] NVO3: Network Virtualization Overlays [Online]. Available: http://datatracker.ietf.org/wg/nvo3/charter/
[34] VXLAN: A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks [Online], Available: http://tools.ietf.org/html/draft-mahalingam-dutt-dcops-vxlan-00
[35] NVGRE: Network Virtualization using Generic Routing Encapsulation[online], Available:http://tools.ietf.org/html/draft-sridharan-virtualization-nvgre-00
[36] Yueping Zhang, Ao-Jan Su, Guofei Jiang, "Understanding data center network architectures in virtualized environments: A view from multi-tier applications," *Computer Networks*, vol.55, no. 9, pp. 2196–2208, June 2011.
[37] Vivek Shrivastava et al., "Application-aware Virtual Machine Migration in Data Centers," in *the IEEE Conference on Computer Communications (INFOCOM '11)*, Shanghai,pp.66–70, Apr. 2011.
[38] Dan Li et al., "Exploring Efficient and Scalable Multicast Routing in Future Data Center Networks," in *the IEEE Conference on Computer Communications (INFOCOM '11)*, Shanghai , pp.1368–1376, Apr. 2011.
[39] Brandon Heller, Srini Seetharaman, Priya Mahadevan, Yiannis Yiakoumis, Puneet Sharma, Sujata Banerjee, Nick McKeown, "ElasticTree: Saving Energy in Data Center Networks," in *7th USENIX Symposium on Networked Systems Design and Implementation(NSDI'10)*, pp.1–17, March 2010.
[40] Kai Chen et al., "Generic and Automatic Address Configuration for Data Center Networks," in *the ACM Special Interest Group on Data Communication (SIGCOMM'10)*, New Delhi, pp.39–50, Aug. 2010.
[41] Dilip A. Joseph, Arsalan Tavakoli, Ion Stoica, "A Policy-aware Switching Layer for Data Centers," the *ACM Special Interest Group on Data Communication (SIGCOMM'08)*, Seattle, pp.1–62, Aug. 2008

Manuscript received: Junuary 13, 2012

Biographies

Yantao Sun

Yantao Sun (ytsun@bjtu.edu.cn) received his PhD degree from the Institute of Software, Chinese Academy of Sciences, in 2006. He is currently a lecturer at the School of Computer and Information Technology, Beijing Jiaotong University. His research interests include cloud computing, datacenter networks, and network management. He has participated in several national 863 Plans and 973 Plans and has cooperated with Intel, ZTE, and Shenghua Group. He has been published in international journals and conferences proceedings. He has authored three national and enterprise standards and two textbooks. He has applied for four national invention patents.

Jing Cheng

Jing Cheng (yourney3@gmail.com) is completing her MS degree in computer science at Beijing Jiaotong University. She received her BS degree in information security from Beijing Information Technology University. She was awarded the honor of outstanding graduate. Her research interests are data center networks and network simulation.

Konggui Shi

Konggui Shi (shikonggui@gmail.com) is completing his MS degree in computer science at Beijing Jiaotong University. He received his BS degree in software engineering from Beijing Jiaotong University. His research interests include distributed systems and network management. He is a skilled developer and has participated in several key corporate and organizational projects.

Qiang Liu

Qiang Liu (liuq@bjtu.edu.cn) received his MS and PhD degrees in communication and information systems from Beijing Institute of Technology in 2004 and 2007. He is currently a lecturer at the School of Computer and Information Technology, Beijing Jiaotong University. His research interests include mobile communication networks, mobile ad hoc networks, and network simulation. During the tenth Five-Year Plan and Twelfth Five-Year Plan, he participated in several national defense pre-research projects and international cooperative research projects as a member of Intel. He has published more than 20 papers as the first author. At present, he is leading several research projects.

ZTE Converged FDD/TDD Solution Wins GTI Innovation Award

27 February 2013, Shenzhen-ZTE Corporation today announced that its converged FDD/TDD solution has won the Global TD-LTE Initiative (GTI) Innovation Award at the Mobile World Congress in Barcelona.

GTI Night is held each year to recognize outstanding achievements in telecommunications. This year's winner was selected by a GTI steering committee comprising multiple operators. GTI Night 2013 was attended by more than 30 global operators and several industry partners.

In 2011, ZTE worked with H3G to build the world's first commercial FDD LTE/TD-LTE dual-mode network. The network was built in Sweden and is the model for dual-mode networks. H3G's network was constructed exclusively by ZTE using an integrated FDD LTE/TD-LTE solution. It can be seamlessly evolved and upgraded.

In December 2011, a dual-mode network was jointly built by ZTE and China Mobile Hong Kong. To date, ZTE has built TD-LTE networks for 42 operators in 30 countries, and 12 of these networks have been put into commercial operation.

GTI is dedicated to developing and promoting the TDD ecosystem. It was founded in 2011 by a consortium of operators, including China Mobile, SoftBank, Clearwire, Bharti Airtel, and Vodafone. It currently has 51 operator members and 44 industrial partners. (ZTE Corporation)

Android Apps: Static Analysis Based on Permission Classification
Zhenjiang Dong, Hui Ye, Yan Wu, Shaoyin Cheng, and Fan Jiang

Android Apps: Static Analysis Based on Permission Classification

Zhenjiang Dong[1], Hui Ye[2], Yan Wu[1], Shaoyin Cheng[2], and Fan Jiang[2]
(1.ZTE Corporation, Nanjing 210012, China;
2. Information Technology Security Evaluation Center, University of Science and Technology of China, Hefei 230027, China)

Abstract

Android has a strict permission management mechanism. Any applications that try to run on the Android system need to obtain permission. In this paper, we propose an efficient method of detecting malicious applications in the Android system. First, hundreds of permissions are classified into different groups. The application programming interfaces (APIs) associated with permissions that can interact with the outside environment are called sink functions. The APIs associated with other permissions are called taint functions. e construct association tables for block variables and function variables of each application. Malicious applications can then be detected by using the static taint-propagation method to analyze these tables.

Keywords

malware; software analysis; static analysis; Android

1 Introduction

Smartphones have become more complex in terms of functions and third-party applications, and this makes them a living space for malware. People store private information such as accounts and passwords on their smartphones, the loss of which could have serious consequences.

Malware that runs on smartphones has the same characteristics as malware that runs on desktops. Thus, traditional malware analysis methods for desktops can also be used for Android. Software analysis techniques develop very fast; static-analysis methods are particularly efficient and easy to use because an application can be analyzed without having to be run [1]. General static-analysis methods include data-flow analysis, control-flow analysis, and type analysis. Symbolic execution is another commonly used static-analysis method [2], [3]. It is used for intelligent path scheduling and constraint resolution.

Android uses a strict permission management mechanism to restrict the behavior of applications. If a program needs to write files to the storage card, the WRITE_EXTERNAL_STORAGE permission must be granted to the program. In other words, all the required permissions need to be granted to the application before it can run on the system.

Because Android is an open-source system, researchers are always interested in investigating its security mechanism. Android is studied using various kinds of methods that can be classified as either dynamic monitoring or static analysis.

Dynamic monitoring often requires system modification so that an application can be monitored as it runs in the Dalvik virtual machine (DVM) or native environment. Some methods require stricter permission management for applications. Kirin checks whether the installed application violates the permission requirement strategy, which comprises an unsafe permission combination [4]. Any application that requests an unsafe combination of permissions should be barred from running. Saint uses a similar method to Kirin but goes further. Developers can design permission assignment on installation and permission use at runtime. TaintDroid add middleware to the system and uses a dynamic, lightweight taint-propagation engine to detect privacy leakage in applications [5]. Apex modifies the Android frameworks to restrict permission-granting [6]. With this method, the system can grant parts of requested permissions and can even withdraw some permissions at runtime. MockDroid modifies the Android system to hide user resources from running processes [7]. TISSA is a privacy protection model for preventing unauthentic applications from accessing private information [8].

Static analysis methods are somewhat different from each other. ScanDroid analyzes the source code and AndroidManifest.xml file of an application and generates a certificate that describes the use of permissions [9]. PiOS uses program profiling to detect privacy leakage from applications on an iOS system [10]. In [11], a decompiling tool called ded is used to decompile Dalvik bytecode to java source code so that the application can be analyzed using current java source code analysis tools. In our static method, permissions are first divided into groups. Then, the association table of block variables (ATBVs) and function variables (ATFVs) is constructed according to the Dalvik bytecode. Finally, we test our static taint-propagation method on these two types of association tables.

2 Android Security Mechanism

The Android smartphone operating system can be divided into application (Fig. 1), application framework, libraries and Android Runtime, and Linux kernel layers. The bottom layers of

the software stack are based on Linux kernel 2.6. Basic device driving, memory management, process management, and network management are implemented in this kernel. Above the kernel layer is the libraries and Android Runtime layer. Android Runtime is a VM for applications, and each application runs on a separate VM. Libraries such as Surface Manager, Media Framework, WebKit, and SQLite are indirectly provided to developers. The second layer contains the application framework, which comprises all kinds of APIs that allow developers to reuse system components and servers. In the top (applications) layer, Google has pre-installed basic applications such as contacts. Users can install any third-party applications in this layer. The Android security mechanism comprises user ID (UID), permission, and signature.

2.1 User ID

In Android, each application has its own UID that is assigned by the system when the application is installed on a device. The UID is not changed. Security limitation is implemented at the process level. By default, applications cannot execute operations that hurt other applications or the system. For an application to run, the system allocates a separate DVM according to the application's UID. The DVM, working as a sandbox, separates applications from each other so that they do not interfere with each other. Directly accessing data of another DVM is forbidden by the system. However, if one application obtains another application's shared UID, it can access data of the other application.

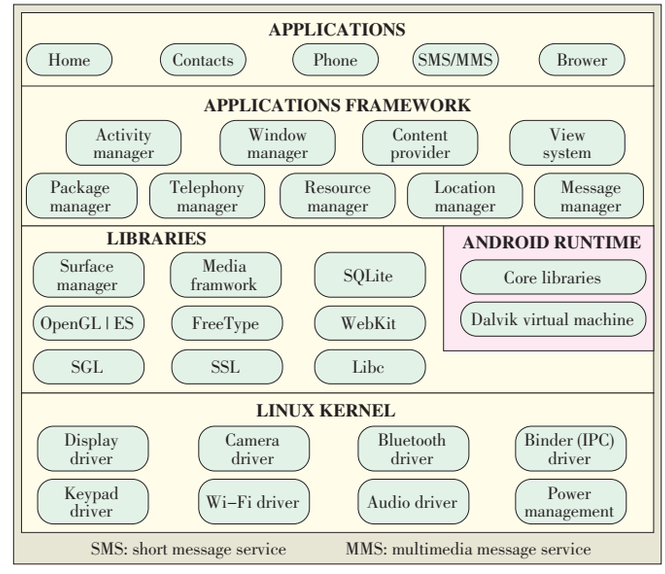
2.2 Permission

Permissions describe the rights of an application to execute some operations. Permissions are complex; there are 115 items in Android 2.3.3. An application must register all required permissions at installation rather than at runtime. If it does not, reinstallation is necessary.

There are three pieces of information associated with a permission: the permission name, group that the permission belongs to, and protection level. Permission groups are classified according to function. For example, the permission group PHONE_CALLS includes the permissions READ_PHONE_STATE, PROCESS_OUTGOING_CALLS and other permissions related to phone calls. The protection level identifies how the permission is protected. There are four protection levels: normal, dangerous, signature, and signature/system. Normal and dangerous permissions are only granted when they are requested; however, unless the application has the same digital certificate as the system, it cannot get permissions at the signature or signature/system level.

2.3 Signature

Each application needs a signature in order to establish a trusted relationship between developers and the application. A signature-level permission can only be granted to applications



▲Figure 1. Android system structure.

that have the same signature as each other. The digital signatures in Android can be designed by application developers and do not need to be authenticated by a digital certificate agency. Each signature has an expiration date that is checked during installation. The system will not check the expiration date after the application has been installed. Even if an application expires after it has been installed, it can still run normally. The signature is also used to update the application. However, in this case, if the signature has expired, the application cannot be updated.

3 Malware

Malware has evolved with the development of the software industry; however, the purpose of malware has never changed. It is software installed on a computer or other device without the user's authorization. It collects sensitive information from the system or does other harmful things. In general, malware can be classified according to whether it

- tries to get remote control of the target system. This category includes bug-exploiting programs, Trojan horses, worms, bots, and viruses.
- tries to maintain remote control. This category includes backdoors and rootkits.
- tries to accomplish specific tasks. This category includes spyware, spamming, adware, phishing, and other similar software.

These classifications were initially designed for malware targeting PCs. Even though smartphone malware might be slightly different, we can still learn a lot from PC malware. Currently, smartphones malware is mainly classified according to malicious behavior, that is, malicious charging, expenses consuming, backdoor operation, privacy violation, and other malicious

behavior [12].

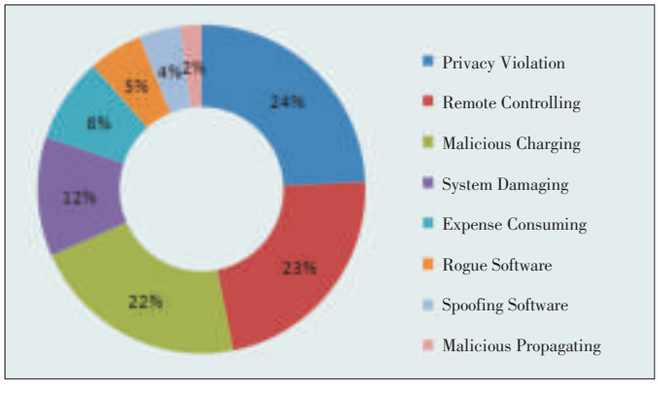
A total of 3523 types of malware were detected in the first quarter of 2012, and nearly 4.12 million phones were infected [12]. There is a clear increase in the types of malware on Android systems. Malicious behavior, including privacy violation, remote controlling and malicious charging, accounted for 60% of malware behavior (Fig. 2).

4 Static Analysis Method based on Permission Classification

4.1 Permission Classification

There are 130 permissions in the latest 4.1 version of Android, and it is difficult to classify them appropriately [13]. Motivated by the permission group design in Android development document [14] and the malware classifications in section 3, we classify these permissions as interacting, controlling and system resource, privacy, and fee (Table 1). Each category is assigned a risk level. Malware in the interacting category poses the highest risk; malware in the controlling and system resource category poses a high risk; malware in the privacy category poses a medium risk; and malware in the fee category poses a low risk. Permissions in the interacting category interact with websites and other outside devices. The reason we assign the interacting category the highest risk level is that without this kind of permission, the phone would not be able to interact with outside devices. Thus, there are no threats to the phone in other permission categories. We assign the controlling and system resource category a high risk level because with the power of control, permissions in the privacy and fee categories would be easy to obtain. Because private information is usually more valuable than money, the privacy category is assigned a medium risk level, and the fee category is assigned a low risk level.

All permissions and application requests are declared in the application's manifest file and are determined on installation. Permissions are used to restrict the operations of a program, so in the program's source code, there should be functions that use the corresponding permissions. The foundation of static



▲Figure 2. Android malware classification in the first quarter of 2012.

▼Table 1. Permission classifications

Permission Group	Included Permission Groups	Risk Level
Interacting	MESSAGES NETWORK	Highest
Controlling and system resource	ACCOUNTS DEVELOPMENT_TOOLS HARDWARE_CONTROLS	High
Privacy	PERSONAL_INFO LOCATION	Medium
Fee	CONST_MONEY	Low

analysis based on a permission–classification method is the construction of a map from functions to their corresponding permissions. Functions that request permissions from the controlling and system resource, privacy, and fee categories are called taint functions.

4.2 Static Analysis Algorithm

In section 4.1, we divided permissions into four groups and assigned risk levels to each of these groups. Functions in the interacting group are the preconditions that allow malware in the other three groups to hurt the system. Therefore, functions belonging to the interacting group are called sink functions. They are the terminating functions of static analysis.

The static analysis algorithm comprises the ATBV&ATFV engine as well as the static taint–propagation engine.

The ATBV is the association table of block variables. Generally, one program may contain hundreds of functions, and one function may contain several basic blocks. In these blocks, there are some variables that are associated through assignment or other operations. For example, *int* $v_1 = v_2$ means variable v_1 is associated with v_2 by assignment. Therefore, we analyze all the variables in one block and construct an association table from them. Similarly, the ATFV is the association table of function variables. It is constructed from the scale of functions. Because one function often contains several blocks, the ATFV is constructed from the ATBV.

4.2.1 ATBV&ATFV Engine

The ATBV&ATVF engine scans the Dalvik bytecode of an application and applies the following steps to each function:

1) Divide the function into several basic blocks using the basic block algorithm mentioned in [15].
2) Calculate the ATBV. In Dalvik VM, operands of an instruction are stored in registers which are reused in a program, so registers should not occur in the items of association table.

When calculating the variables, the engine reaches instructions such as aput and aput–object that write to a register. The engine first clears the register association variable then associates it with the new variable. Similarly, the engine reaches instructions such as aget and aget–object opcode that read a register. The engine adds a register association variable to the association table that has association variables of destination registers. When calculating the association table of variables, the variables are initially untainted. Meanwhile, the engine adds

all the functions that are called by the current function to the function-call list:

1) Calculate ATFV by merging ATBV with ATVF. For block-crossing variables, redundant table items should be deleted during merging.
2) Calculate the entry function list based on all the function-calling lists. This is accomplished by calculating the number of calls for each function and adding functions with a zero call number to the entry function list.

4.2.2 Static Taint-Propagation Engine

There are fifteen different taint states: NONE, MESSAGE, CONTACT, MAIL, CALLS, CALL_RECORD, LOCATION, LOCAL_DATABASE, LOCAL_LIB, FILE, CAMERA, MICROPHONE, OTHER_DEVICE, OTHER_CONTENT and WEB_DATA. The taint state of most variables is NONE. However, when a variable is related to a taint function, its state may change. For example, if the value of a variable comes from message-sending APIs, the taint state will be MESSAGE. When handling a variable, we check whether it is tainted or not rather than determine its specific taint state. The engine takes the output entry function list of the first engine as input and applies the following steps to each function:

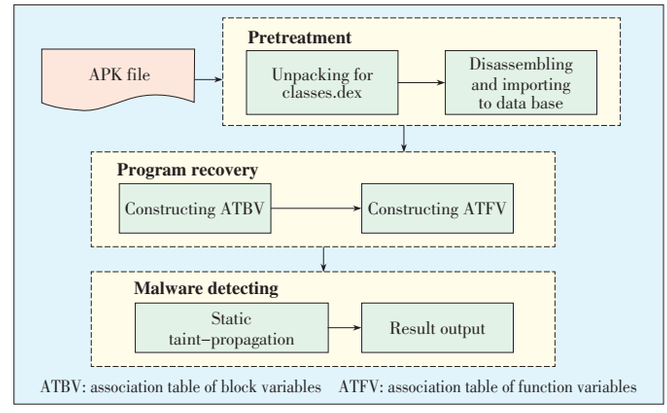
1) Start deep traversing from the entry function. When a taint function is encountered, the taint state of its return value is set to tainted. Then, all the associated variables in ATFV are set to tainted. The taint states of function-crossing variables are propagated from the formal parameters. Therefore, all the variables associated with the formal parameters in ATFV are set to tainted.
2) In the deep traversing process, only when a taint function and sink functions appear in the same path can this path be recorded. This strategy can reduce false positives because no taint function in the path means that there is no resource or privacy in the variables, so the path should be safe.

This method reduces false positives, which is the main shortcoming of other static analysis methods. This method concentrates on variables and simulative execution of the target program on these variables. Current static methods often contain both data flows and control flows, which means they are time-consuming and memory-consuming.

5 Experimental Evaluation

We have developed a prototype system based on the previously mentioned method. The system framework is shown in Fig. 3. The system analyzes the bytecode of an application without accessing the source code.

The system comprises pretreatment, program-recovery, and malware-detection modules. The pretreatment module unpacks the classes.dex file from an Android apk. Then, this file is disassembled to bytecode using disassembling tools, and the output bytecode is imported into a database. After pretreatment, the program recovery module reads bytecode from database and starts constructing the ATBV and ATFV. Finally, the malware- detection module analyzes each execution path in the taint-propagation algorithm and outputs the results.



▲Figure 3. Prototype system framework.

We use this system to detect 7806 Android applications from an online application market. The results are shown in Table 2. A total of 2629 (33.68%) of the applications were potentially malicious, and the remaining 5177 (66.32%) of the applications were normal.

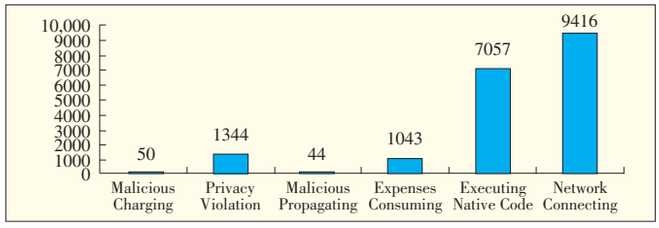
Of the 2629 malicious apps, 609 demonstrated high-risk malicious behavior (Table 3). A total of 18,811 malicious behaviors are detected in all the malicious apps (Fig. 4). We detected 50 types of malicious charging behavior in 31 apps, 1344 types of privacy violation behavior in 614 apps, 44 types of malicious propagation behavior in 40 apps, 1043 types of expense-consuming behavior in 350 apps, 7057 types of native code executing behaviors in 1729 apps, and 9416 types of unauthorized network connection behavior in 1612 apps. One executing path or one application can exhibit multiple types of malicious behavior. Because apps developed in Java are easy to disassemble, many developers use native codes to enhance copyright protection. This method is also used by hackers to hide malicious code. The main profit model of Android apps is to deliver advertising. Michael C. Grace analyzed some advertising packages and found many problems [16]. Privacy violation and expense-consumption behavior are also common. Some malicious apps set out to obtain a user's private informa-

▼Table 2. Results of detecting applications in an online application market.

Type	Number of Apps	Percentage
Potentially malicious	2629	33.68%
Normal	5177	66.32%

▼Table 3. The risk level of potentially malicious applications

Risk Level	Number of Apps	Percentage
High risk	609	23.16%
Other	2020	76.84%



▲**Figure 4. Malicious behavior distribution.**

tion, such as bank accounts and passwords. This results in monetary loss. Connecting to the network or sending messages in the background are expense-consuming, and these types of behavior are common in today's apps. However, malicious charging and malicious propagation are relatively uncommon.

6 Conclusion

In this paper, we have proposed a static analysis method based on permission classification. This analysis system comprises the ATBV&ATFV engine, which is used to construct variable tables, and the static taint-propagation engine, which is used to analyze the program. We used this system to detect 7806 apps from an online market. The experimental results show that our method is not only feasible but also effective in detecting malicious behavior in Android apps.

Acknowledgment

This research was supported in part by the Fundamental Research Funds for the Central Universities of China (Grant No. WK0110000007), the Specialized Research Fund for the Doctoral Program of Higher Education of China (Grant No. 20113402120026), the Natural Science Foundation of Anhui Province, China (Grant No. 1208085QF112), the Foundation for Young Talents in College of Anhui Province, China (Grant No. 2012SQRL001ZD) and the Research Fund of ZTE Corporation.

References

[1] MEI Hong, WANG Qianxiang, ZHANG Lu, WANG Ji. "Software Analysis: A Road Map," *Chinese Journal of Computers*, vol.32, no. 9, pp.1697-1710, 2009.
[2] James C K, "Symbolic execution and program testing," *Communication of the Association for Computing Machinery*, vol.19, no.7, pp. 385-394,1976.
[3] Jinbin Lin, Xiaofei Zhang, Hui Liu, Research of Symbolic Execution[Online], Available: http://www.itsec.gov.cn/zxzz/jsyyy/10511.htm
[4] W. Enck, M. Ongtang, and P. McDaniel. "On Lightweight Mobile Phone Application Certification,"in *Proceedings of the 16th ACM Conference on Computer and Communications Security*, CCS '09, Chicago, October 2009,pp. 235-245.
[5] W. Enck, P. Gilbert, B.-G. Chun, L. P. Cox, J. Jung, P. McDaniel, and A. N. Sheth, "TaintDroid: An Information-Flow Tracking System for Realtime Privacy Monitoring on Smartphones," in *Proceedings of the 9th USENIX Symposium on Operating Systems Design and Implementation(OSDI '10)*, Vancouver, February 2010, pp. 1-6.
[6] M. Nauman, S. Khan, and X. Zhang,"Apex: Extending Android Permission Model and Enforcement with User-Defined Runtime Constraints," in *Proceedings of the 5th ACM Symposium on Information, Computer and Communications Security*, NewYork, April 2010, pp. 328-332.
[7] A. R. Beresford, A. Rice, N. Skehin, and R. Sohan, "MockDroid: Trading Privacy for Application Functionality on Smartphones," in *Proceedings of the Twelfth Workshop on Mobile Computing Systems & Applications(HotMobile'11)*, Phoenix, May 2011.
[8] Y. Zhou, X. Zhang, X. Jiang, and V. Freeh, "Taming Information-Stealing Smartphone Applications (on Android)," in *Proceedings of the 4th International Conference on Trust and Trustworthy Computing(TRUST'11)*, Pittsburgh ,June 2011.
[9] A. P. Fuchs, A. Chaudhuri, and J. S. Foster, SCanDroid: Automated Security Certification of Android Applications[Online], Availabel: http://www.cs.umd.edu/~avik/papers/scandroidascaa.pdf
[10] M. Egele, C. Kruegel, E. Kirda, and G. Vigna, "PiOS: Detecting Privacy Leaks in iOS Applications," in *Proceedings of the 18th Annual Network and Distributed System Security Symposium(NDSS '11)*, Sandiego, February 2011.
[11] W. Enck, D. Octeau, P. McDaniel, and S. Chaudhuri, "A Study of Android Application Security," in *Proceedings of the 20th USENIX Security Symposium*, August 2011.
[12] NQ Mobile Inc. Android smartphone security on the first quarter of 2012[Z/OL] 2011-04-14. http://cn.nq.com/neirong/2012Q1.pdf
[13] Manifest.permission [Online], Available: http://developer.android.com/reference/android/Manifest.permission.html
[14] Manifest.permission_group [Online], Available: http://www.ideasandroid.com/android/sdk/docs/reference/android/Manifest.permission_group.html
[15] Alfred V.Aho, Monica S.Lam, Ravi Sethi, et al, Compilers: Principles, Techniques, and Tools, 2008.
[16] Michael C. Grace, Wu Zhou, Xuxian Jiang, Ahmad-Reza Sadeghi, "Unsafe exposure analysis of mobile in-app advertisements," in *Proceedings of the fifth ACM conference on Security and Privacy in Wireless and Mobile Networks (April 2012)*, Tucson, pp. 101-112.

Manuscript received: October 16, 2012

Biographies

Zhenjiang Dong

Zhenjiang Dong (dong.zhenjiang@zte.com.cn) received his Master's degree from Harbin Institute of Technology in 1996. His research interests include switches, intelligent networks, business platform development, data-service platform development, and architecture design. He is currently the assistant dean of ZTE Communication Business Institute and leads the Business Technology Group of the ZTE Committee of Experts. He is a member of CCF and a committee member of CCF TCSC. He is a senior research scientist in the fields of business network architecture, communication technology and protocols, mobile internet technology, and cloud computing. He has authored several articles and patents.

Hui Ye

Hui Ye (yehui11@mail.ustc.edu.cn) is pursuing his MS degree at the University of Science and Technology of China. His research interests include software security and mobile phone terminal security.

Yan Wu

Yan Wu (wu.yan2@zte.com.cn) received her BS degree in computer science from Southeast University in 2002. She is currently a pre-research engineer in ZTE Communication Business Institute. Her research interests include intelligence network, architecture design, and business standards. She has authored several patents.

Shaoyin Cheng

Shaoyin Cheng (sycheng@ustc.edu.cn) is a lecturer in the Department of Information Security, University of Science and Technology of China. His research interests include network and system security, and protocol analysis and testing.

Fan Jiang

Fan Jiang (fjiang@ustc.edu.cn) is a professor in the School of Computer Science and Technology, University of Science and Technology of China. His research interests include computer network, protocol and software testing, and information security.