MULTIMEDIA COMMUNICATIONS TECHNICAL COMMITTEE
IEEE COMMUNICATIONS SOCIETY
http://www.comsoc.org/~mmc

E-LETTER

Vol. 6, No. 8, August 2011

IEEE COMMUNICATIONS SOCIETY

CONTENTS

SPECIAL ISSUE ON QUALITY OF EXPERIENCE ISSUES IN MEDIA DELIVERY

Quality of Experience Issues in Media Delivery

Zhibo Chen, Technicolor Research & Innovation, China
Zhibo.chen@technicolor.com

For Media delivery industry, guarantee of user experience is always a key factor for many media service to consumers. Therefore contrary to QoS, concept of QoE (Quality of Experience) has been highlighted concerns for media delivery industry referring to "the overall acceptability of an application or service, as perceived subjectively by the end user".

Media delivery industry takes end-user QoE monitoring as either "critical" or "very important" to their video initiatives, and meanwhile the top issue reported from industry is that current QoE assessment solutions deployed today are not accurate enough and too costly to measure end user experience.

Therefore this E-Letter tries to provide a compact and insightful understanding on current research status on QoE, standard activities and applications related to accurate and light weight QoE assessment solutions for media delivery.

The first paper, titled 'Perceptual-based visual quality metrics: Methodologies, directions and challenges', provides an overview on QoE concept, difference from QoS, existed classification, research challenges and opportunities. Visual Quality Metrics are classified according to the availability of the reference and employed methodologies. Challenges from aspect of multi-dimensional measurement, multi-disciplinary processing, inadequate understanding of the human perception system, 3D content are also highlighted.

Standardization is always the most efficient way to push new technologies to be adopted by industry. The second paper, titled 'Recent Multimedia QoE standardization activities in ITU-T SG12', briefly outlines recent and ongoing standardization activities of SG12 on multimedia QoE assessment. A structured figure on ITU-related standardization bodies working on multimedia QoE assessment was given in the paper. Standard activities related to speech QoE, Audio, Video and Audiovisual QoE are introduced with their latest status in the ITU-T standard group. Relation between different standard group and future outlook are also concluded in the paper.

We need to clearly clarify and classify different applications scenarios when designing applicable QoE assessment schemes. The third paper, titled 'QoE Models for Quality Planning and Monitoring for Video Delivery', tries to clearly differentiate the requirements for QoE models from two typical scenarios in video delivery, namely, video system quality planning scenario and the quality monitoring scenario. Then QoE modeling approaches for the two scenarios and different criteria to construct the evaluation datasets are also illustrated in the paper.

3D video becomes more and more important in media delivery applications. The fourth paper, titled 'Objective quality assessment for 3-D video delivery', presents a brief review on recent research on objective 3-D video quality assessment based on subjective observations, i.e., based on the study of the characteristics of the human visual system. The importance of no-reference and reduced-reference metrics for the evaluation and adaptation of media delivery systems is also highlighted in the paper.

For media delivery applications, accurate QoE monitoring is never end of game, it is more practical significance to optimize service quality based on results from QoE monitoring. The fifth paper, titled 'QoE Aware Mobile Video Delivery', discusses some key research challenges for the QoE maintenance in mobile video delivery across heterogeneous wireless networks. In particular, it presents the need for QoE-aware mobility and the need for the tight integration between handover, network selection and rate control.

In conclusion, I hope readers can get a comprehensive impression on the status quo, potential challenges and opportunities to solve QoE issues in media delivery from this E-letter.

Zhibo Chen is principal scientist in Technicolor Research & Innovation Department, Distinguished Fellow of Technicolor Fellowship Program, manager of media processing lab in Technicolor. He received his B.Sc. and Ph.D. from EE Tsinghua University separately in 1998 and 2003. He has been with Technicolor since 2004 and worked in Sony Research before that. His areas of expertise and interests include: media processing and coding, media Quality of Experience analysis and management for content delivery, perceptual based rendering, etc. He was the Media QoE research work lead for Technicolor Research & Innovation and contributed to the design of media QoE assessment platforms for CE applications. He has more than 50 granted and filed EU and US patent applications, more than 50 publications and standard proposals, his contribution of UMH Fast ME algorithm has been adopted by H.264 standard, widely used in standard reference software and largely cited. He is member of IEEE Visual Signal Processing and Communications Committee, and member of IEEE Multimedia Communication Committee. He is RC member of ISCAS meetings, TPC member of PCS 2006 and VCIP 2010/2011, Chair of ICME 2011 Multimedia Standard, Services and QoE track, Co-Editor of IEEE Journal on Selected Areas in Communications QoE-Aware Wireless Multimedia Systems 2011.

Perceptual-based Visual Quality Metrics: Methodologies, Directions and Challenges

Junyong You[1], Liyuan Xing[1], Touradj Ebrahimi[1,2], and Andrew Perkis[1]
[1] Norwegian University of Science and Technology, Norway
(junyong.you@ieee.org, andrew@iet.ntnu.no, liyuan.xing@q2s.ntnu.no)
[2] École Polytechnique Fédérale de Lausanne, Switzerland
(touradj.ebrahimi@epfl.ch)

1. Introduction

With the advent of multimedia applications nowadays, perceived quality assessment is becoming more and more important in the visual signal processing chain. In order to offer the best quality experience to the end-users of multimedia services, a new notion, Quality of Experience (QoE) driven by user experience, such as preference and expectation, has been proposed to replace traditionally used Quality of Service (QoS) criteria [1].

Subjective quality assessment performed by human subjects is considered to be the most reliable and accurate way to evaluate the perceived visual quality. A lot of subjective assessment methodologies have been standardized by the International Telecommunication Union (ITU) for different application scenarios [2]. However, subjective quality assessment is always time-consuming and cannot be performed in real-time applications. Therefore, a lot of research efforts have been made in development of objective visual quality metrics that can evaluate quality degradation automatically by machine [3]. The task of objective metrics is to predict the perceived quality as much as possible correlatively with the subjective quality judgment given by human viewers.

Objective quality metrics can in principle be classified into three categories: full reference (FR), reduced reference (RR) and no reference (NR) or blind, according to the availability of the reference, undistorted visual signals. Two widely used FR metrics, namely MSE and PSNR, have been found not to be well correlative with the subjective quality assessment, as they do not take into account the attributes of the human perception system in their design [4]. Consequently, many quality metrics based on simulating the visual perception mechanism have been proposed, showing much better performance than MSE and PSNR [3]. However, as the human perception process of visual stimuli is a quite complicated mechanism and it is not adequately understood by the research community, existing visual quality metrics are still long away from being widely applicable and universally recognized.

This paper will provide an overview of perceptual-based visual quality metrics by introducing some widely employed methodologies, potential directions, and challenges. The remainder of this paper is organized as follows. Section 2 presents state-of-the-art methodologies in the development of visual quality metrics. Some promising research directions and challenges are then summarized in Section 3.

2. Methodologies of Visual Quality Metrics

As introduced earlier, visual quality metrics can be classified into FR, RR and NR categories, in which FR metrics have complete access to the reference signals with perfect quality, NR metrics are based on distorted presentations alone, and between FR and NR metrics applies RR quality models which assume that reference signal is only partially accessible. Typically, the amount of information from the reference signal is significantly less than the signal itself. FR metrics compare the perceived difference between the reference signal and its distorted version, in which the characteristics of the human visual system (HVS) can be integrated. So far FR metrics have attracted the highest research interests and a lot of advanced models have been proposed, or even standardized by VQEG and ITU [5]. Statistically speaking, most FR metrics can achieve high correlation with the subjective quality judgment in certain application scenarios. However, RR and NR metrics have much wider applications in practical scenarios, since the reference signal is always unavailable due to current network conditions and

application limitations. RR metrics usually extract some features that can represent quality characteristics as side information and then transmit it to the receiver through ancillary channel associated with the distorted visual signals. The receiver usually extract the same quality features from the received, distorted signal and compare them with the original quality features, which are assumed to be not degraded through the transmission. Quality measurement is conducted by comparing two feature sets. In some RR metrics, the quality features extracted from the reference signal may also be degraded through transmission, e.g. embedding pseudo watermark into the signal, and the quality evaluation is finally accomplished by measuring the degradation on the embedded watermark, under the assumption that the degradation on the watermark can approximate the signal distortion. NR quality assessment is relatively complicated compared to FR and RR metrics, as the reference signal is completely inaccessible. Most existing NR metrics aim to evaluate the quality degradation caused by certain compression errors, such as blockiness widely appearing in block-based compression schemes and transmission in error-prone networks, and blurring artifacts caused by lossy quantization [6]. Additionally, another important issue in NR quality assessment is the training of the quality models. This is because the quality scores computed by NR metrics might be hard to interpret as there is no reference signal as a benchmark of the best quality. Machine learning algorithms have been widely used in training and tuning NR visual quality metrics.

In addition, visual quality metrics can also be classified into two approaches according to the employed methodologies, namely psychophysical and engineering. As the quality perception is generated in the human brain through the psychological process of the received visual stimuli, the processing mechanisms in the human perception system should be taken into account in accurately estimating the perceived quality. A widely used perceptive characteristic is contrast sensitivity modeling the visual sensitivities to different spatial and temporal frequencies, foveal, as well as attentive information [7]. Other important attributes that have significant impact on visual quality perception include masking effect, color perception, etc. Statistically speaking, the quality metrics in the psychological approach show promising performance when evaluating the perceived visual quality, while their computation is also accordingly complex. Thus, simplified metrics based on the extraction and analysis of certain features or artifacts have been proposed. These metrics, categorized as the engineering approach, assess how pronounced the detected artifacts are. These metrics do not necessarily disregard the attributes of the human perception system, but visual content and distortion analysis rather than fundamental vision modeling is the conceptual basis for the their design [8].

3. Future directions and challenges

Although a lot of research efforts have been made in the development of accurate visual quality metrics, we are still a long way from visual quality assessment methodologies that are widely applicable and universally recognized. For example, a visual quality metric developed for standard definition visual signals might show quite poor performance in evaluating the quality of high definition presentations. Most importantly, with the evolution of multimedia applications, their kernel goal has been changing from providing the content itself to satisfying various requirements of different users. Therefore, the measurement and improvement of a revolutionary concept, QoE, have attracted more and more research interests. According to the ITU definition, QoE refers to “the overall acceptability of an application or service, as perceived subjectively by the end user.” [1]. Contrary to QoS solutions, QoE is mainly driven by user experience, such as preference, expectation and knowledge background. Due to its variety of QoE influence factors, it is extremely difficult to define and assess QoE in a simple framework. Traditional QoS metrics usually use single dimensional measure to evaluate the quality of a visual signal. However, QoE can be influenced by a lot of factors, and thus, it is believed multi-dimensional measurement should be an appropriate solution to assess QoE. A practical difficult in development of QoE metrics is that many external factors, such as environment, context and individual preference, are not easily modeled in the metrics. This research issue requires multi-

disciplinary collaboration covering signal processing, cognition, and psychology, etc.

Another challenge might come from the inadequate understanding of the human perception system. Processing visual stimuli and generating visual perception are a quite complicated process in our brain. Only a small parts of this process have been well understood and modeled. Nonetheless, visual quality metrics only make use of some simple psychological mechanisms that can be easily modeled in a computable approach. For example, the widely used contrast sensitivity function (CSF) is mainly determined by spatiotemporal frequency. However, the contrast sensitivity of human vision can be influenced by many more factors, such as the visual attention mechanism, eye movement. Additionally, visual attention is another integral mechanism of the perception system, and it is always guiding the vision behavior in perceiving visual stimuli. Therefore, integration of visual attention in quality metrics should not be ignored [9]. However, existing work on this topic usually uses a simple combination between attention map derived from computable attention models and quality map generated in quality metrics to compute the attention-based quality measure. However, is such a simple combination really appropriate for integrating the visual attention mechanism into quality assessment? To answer this question, deep understanding of the working progress of the generation of attention is required.

Furthermore, even though some FR visual quality metrics, especially for static image presentations, have achieved good performance, other two metric categories, namely RR and NR metrics, are not widely accepted by content and network services. Some subjective quality experiments conducted by the VQEG have demonstrated that the quality judgments given by human subjects between with and without reference signals have high correlation. Whereas, almost all the NR metrics perform quite worse than FR metrics. Thus, the development of RR and NR quality metrics still requires a lot of research efforts.

Finally, three-dimensional (3D) content has attracted more and more interests from both academic research communities and industrial market, as 3D presentations can provide more immersive experience to viewers compared to the 2D counterparts. Standardization efforts are also in progress to establish standard algorithms for 3D contents. However, there is still lack of accurate 3D quality assessment methodologies for either subjective methods or objective metrics. Existing works in the literature usually adopt traditional 2D quality methods in 3D quality assessment [10]. Though some studies have demonstrated this is one of the feasible solutions, 3D quality experience can in principle be influenced by some specific characteristics of 3D presentations, such as depth perception. Therefore, more research efforts on investigating the fundamental mechanism of 3D visual perception and quality assessment are required.

References

[1] ITU-T P.10/G.100, Vocabulary for performance and quality of service: Amendment 1New Appendix I – Definition of Quality of Experience (QoE), International Telecommunication Union, Jan. 2007.

[2] ITU-T P.910, Subjective video quality assessment methods for multimedia applications, International Telecommunication Union, 1999.

[3] J. You, U. Reiter, M. M. Hannuksela, M. Gabbouj, and A. Perkis, “Perceptual-based objective quality metrics for audio-visual services - A survey,” Signal Processing: Image Communication, vol. 25, no. 7, pp. 482-501, 2010.

[4] B. Girod, “What’s wrong with mean-square error,” in Digital Images and Human Vision, A. B. Watson, Cambridge, MA: MIT Press, pp. 207-220, 1993.

[5] ITU-T J.247, “Objective perceptual multimedia video quality measurement in the presence of a full reference,” International Telecommunication Union, Aug. 2008.

[6] S. S. Hemami, and A. R. Reibman, “No-reference image and video quality estimation: Applications and human-motivated design,” Signal Processing: Image Communication, vol. 25, no. 7, pp. 469-481, 2010.

[7] S. Daly, “The visible differences predictor: An algorithm for the Assessment of Image Fidelity”, in Digital Images and Human Vision, A. B. Watson, Cambridge, MA: MIT Press, pp. 179-206, 1993.

[8] 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, Sep. 2004.

[9] J. You, A. Perkis, M. Gabbouj, and M. M. Hannuksela, "Perceptual quality assessment based on visual attention analysis", in Proc. ACM Int. Conf. Multimedia (MM), Beijing, China, Oct. 2009, pp. 561-564.

[10] J. You, G. Jiang, L. Xing, and A. Perkis, "Quality of visual experience for 3D presentation – stereoscopic image," in High-Quality Visual Experience: Creation, Processing and Interactivity of High-resolution and High-dimensional Video Signals, Springer, pp. 51-77, 2010.

Junyong You is a research scientist in the Department of Electronics and Telecommunications at the Norwegian university of Science and Technology (NTNU) in Trondheim, Norway. He received his M.S. degree in Computational Mathematics in 2001 and Ph.D. degree in Electronics and Information Engineering in 2007, respectively, both from Xi'an Jiaotong University, China. Dr. You worked as a senior researcher at the Tampere University of Technology (TUT), Finland, from 2007 to 2009. He is serving as Editorial Board member for two international journals, and has served as TPC member for several international conferences and workshops. He was the local arrangement co-chair of the second International Workshop on Quality of Multimedia Experinece (QoMEX'10) and is the general co-chair of International Workshop on Multimedia Quality of Experience: Modeling, Evaluation and Directions (MQoE'11). His research interests include semantic multimedia content analysis, vision technology, QoS and QoE mechanisms and models. Dr. You is the author or co-author of more than 40 articles for scientific books, journals and conferences. He is a member of the IEEE.

Liyuan Xing is a Ph.D. candidate with the Centre for Quantifiable Quality of Service (Q2S) in Communication Systems at the Norwegian university of Science and Technology (NTNU) since August 2008. She received her M.S. degree in computer science from Institute of Computing Technology, Chinese Academy of Sciences, in 2006. During her master study, she worked on content based multimedia analysis. She won the Chinese Academy of Sciences Liu Yongling Scholarship Excellence Award in July 2006. She worked on medical image processing and healthcare enterprise integrating at the R&D Center, Toshiba (China) Co., Ltd., as a senior software engineer. Currently her research interests are quality assessment and modeling within new application scenarios, namely in the realm of 3D, virtual environments, gaming and 3DTV like scenarios, especially focus on stereoscopic media.

Touradj Ebrahimi received his M.Sc. and Ph.D., both in Electrical Engineering, from the Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland. He is currently a full professor at EPFL heading its Multimedia Signal Processing Group. He is also adjunct Professor with the Centre of Quantifiable Quality of Service at Norwegian University of Science and Technology (NTNU). Prof. Ebrahimi has initiated more than two dozen National, European and International cooperation projects with leading companies and research institutes around the world, and is currently the Chair of the COST Action IC1003 QUALINET, a European network on Quality of Experience in multimedia products and services. Prof. Ebrahimi has served as Scientific Expert and Evaluator for the European Commission, The Greek Ministry of Development and other research funding agencies. He is also the head of the Swiss delegation to MPEG, JPEG and SC29, and acts as the Chairman of Advisory Group on Management in SC29. In 1993, he was a research engineer at the Corporate Research Laboratories of Sony Corporation in Tokyo, where he conducted research on advanced video compression techniques for storage applications. In 1994, he served as a research consultant at AT&T Bell Laboratories working on very low bitrate video coding. Prof. Ebrahimi founded two start-ups and is a member of Scientific Advisory Board of various start-up and established companies.

Andrew Perkis is a full professor of Digital Image Processing at the Department of Electronics and Telecommunications at the Norwegian university of Science

and Technology (NTNU) in Trondheim, Norway. He received his Siv.Ing and Dr.Techn. degrees in 1985 and 1994, respectively. He is member of the management team of the National Centre of Excellence - Q2S - Quantifiable Quality of Service in Communication Systems, where he is responsible for "Networked Media Handling". He is Vice Chair of COST action IC1003 QUALINET European Network on Quality of Experience in Multimedia Systems and Services (End date: November 2014). Currently he is focusing on Multimedia Signal Processing, specifically within methods and functionality of content representation, quality assessment and its use within the media value chain in a variety of applications. He is a senior member of the IEEE. He has more than 150 publications at international conferences and workshops and more than 50 contributions to International standards bodies.

.

Recent Multimedia QoE standardization activities in ITU-T SG12

Alexander Raake and Sebastian Möller
Deutsche Telekom Laboratories, Germany
Alexander.Raake@telekom.de, Sebastian.Moeller@telekom.de

1. Introduction

Study Group 12 (SG12) of the Telecommunications Sector of the International Telecommunication Union (ITU-T) was mainly concerned with speech transmission networks and the related quality of speech services in the past. In the meantime, it has increasingly broadened its focus towards multimedia service quality assessment. This development started with SG12 becoming ITU-T's lead Study Group on Performance and Quality of Service (QoS), and by the evolution of telephony towards VoIP. As a logical step forward, multimedia telephony and IP-based video became topics of work covered by ITU-T SG12, and as part of the transition from QoS to Quality of Experience (QoE), SG12 became Lead Study Group on Quality of Service and Quality of Experience [1]. In this short-paper, we will briefly outline recent and ongoing standardization activities of SG12 on multimedia QoE assessment. Here, we will include both non-interactive audiovisual media as well as interactive, voice-based media such as VoIP, videotelephony and conferencing. Assessment methods will be distinguished according to whether they are "subjective", that is, involve test subjects, or whether they are "objective", that is involve signals or technical parameters as the basis for quality predictions. Since ITU-T contributions are not publicly available, we will make reference only to recommendations, and if not yet available to published papers where possible.

2. Multimedia QoE-related Workgroups

There are a number of standardization bodies that deal with assessment methods for multimedia (MM) QoE. At the level of ITU and associated groups, several respective work groups can be differentiated, which all deal with different aspects of multimedia QoE, see Figure 1 for an overview. Apart from ITU-T, the picture shows ITU-R, the radiocommunication sector of ITU [2], and VQEG, the Video Quality Expert Group [3][2]. ITU-R traditionally deals with radio and television broadcast quality assessment. VQEG is a joint group founded in 1997 by experts from ITU-R and ITU-T, that does not formally depend on ITU; with a few exceptions, VQEG's main emphasis has been on developing standards on subjective and objective assessment methods, which often are adopted as standards by ITU-T SG9. In the following, we will focus on SG12's activities, and refer to recommendations under the responsibility of other standardization bodies where appropriate.

3. Study Group 12 and MM QoE activities

Study Group 12 is, as all of ITU-T, organized in

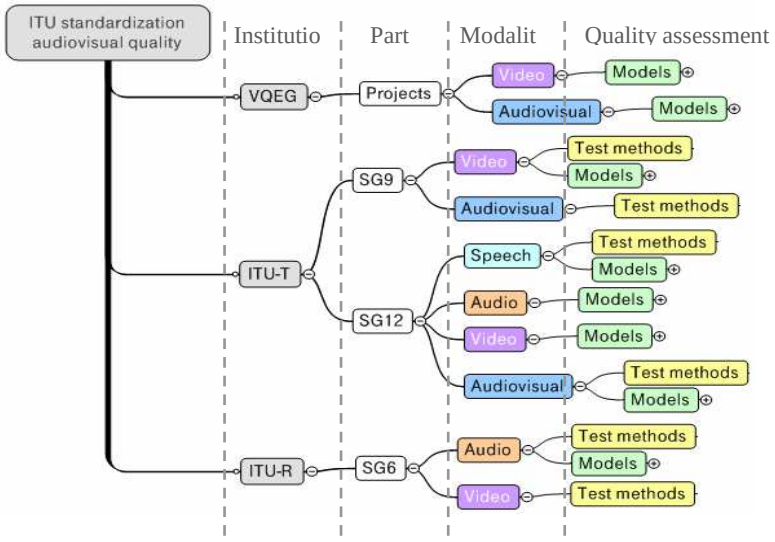


Figure 1: ITU-related standardization bodies working on multimedia QoE assessment.

terms of "Questions", which are the workgroups developing and maintaining different sets of standards. Table 1 lists the Questions that deal with multimedia quality assessment. SG12's central question for all topics related to subjective testing is Q.7/12, which also develops the test plans for speech codec evaluation campaigns. Apart from these tasks, in early 2011 Q.7/12 has finalized the "Handbook of subjective testing practical procedures", which so far mainly addresses speech services, and will later be extended towards audiovisual quality issues. Another running activity is the development of the upcoming standard P.MULTI, a multi-dimensional method for assessing speech quality, targeting diagnostic information on why a given service is considered to have bad quality.

#	Modality	Type	Title (shortened)
Q.7	Speech, audio, video, audiovisual	Subj. methods	Methods, tools & test plans for subjective assessment of speech, audio & audiovisual quality
Q.8	Speech	Obj. methods	E-Model extension towards WB transmission & future telecommunication & application scenarios
Q.9	Speech, (audiovisual)	Obj. methods	Perceptual-based objective methods for voice, audio & visual quality measurements in telecommunication services
Q.13	Audio, video, audiovisual	Requirements	QoE, QoS & performance requirements and assessment methods for multimedia including IPTV
Q.14	Audio, video, audiovisual	Obj. methods	Development of parametric models and tools for audiovisual & multimedia quality measurement
Q.15	Speech	Obj. methods	Objective assessment of speech & sound transmission performance quality in networks
Q.18	Speech, audiovisual	Studies, Requirements	Conferencing and telemeeting assessment

Table 1: Overview of ITU-T SG12 Questions directly dealing with Multimedia QoE

3.1 Speech QoE

Q.8/12 is responsible for the "E-model" [4], which is ITU-T's recommended parametric tool for planning speech transmission networks. Recent advances here include a first model draft for listening quality of wideband speech services (50 – 7000 Hz instead of 300 – 3400 Hz as for narrowband) [5], and first steps towards extending the model to include user interfaces in terms of the quality-impact of echo-cancellers and noise-suppression algorithms [6].

An important achievement was made by Q.9/12 with their finalization of the POLQA development [7], a signal-based full reference (FR) model for speech quality assessment that will succeed PESQ [8]. It can be applied both to narrowband and superwideband speech, and shows considerably better performance than PESQ. In parallel to Q.7/12's work on a subjective method P.MULTI, Q.9/12 currently develops models for the instrumental prediction of individual quality dimensions (P.AMD). In a related work carried out together with Q.16/12, new objective methods for the instrumental prediction of the technical cause for quality problems are being developed (P.TCA). Another ongoing activity is concerned with the objective evaluation of noise reduction systems (project P.ONRA). Just recently, Q.9/12's scope has been extended towards the analysis and recommendation of "methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models".

Q.15/12 conducts work in the domain of speech quality monitoring, and here is responsible for standards such as Rec. P.564. A currently ongoing activity is the work on P.CQO, a model

targeting quality predictions that apply for a conversational situation, reflecting quality ratings obtained for listening quality, talking quality and interactivity (see e.g. [13]).

As a newly created work group, Q.18/12 deals with all aspects related with conferencing and telemeeting assessment. First contributions received at SG12's January 2011 meeting were concerned with multiparty conversation analysis, assessing conversational quality under delay, and investigating the quality of spatial-audio and video conferencing.

3.2 Audio, Video and Audiovisual QoE

In 2010, Q.13/12 has issued Rec. G.1011, which is a recommendation to serve as an introductory document for quality assessment in general, and the objective methods developed by SG12 in particular [9]. Along these lines, Q.13/12 deals with all multimedia QoS and QoE requirement-related issues that are not directly handled elsewhere in SG12. Among these topics are the maintenance and extensions of the parametric model for videotelephony quality assessment, Rec. G.1070 [12], and the ongoing development of the audiovisual service planning tool G.OMVAS.

Q.14/12 currently has two active competitions running:

(1) P.NAMS is an upcoming standard for audiovisual quality monitoring based on packet-header information, as it is available, for example, when the payload information is encrypted. The models to be described in the standard will provide an audio, a video and an integral, audiovisual quality score. The input to the model is a PCAP-file as it can be captured by typical network monitoring devices. Two parallel workstreams are ongoing, one targeting the high-bitrate application area (HBR), with IPTV-typical video resolutions (Standard Definition – SD, High Definition, HD, including 1080i, 1080p and 720p), the other targeting the low-bitrate case (LBR), with mobile video typical resolutions (QCIF, QVGA, HVGA). The respective packetization schemes are RTP/UDP/IP and MPEG2-TS/RTP/UDP/IP for HBR, and RTP/UDP/IP for LBR. The considered codecs are H.264 and MPEG2 for HBR-video, H.264 and MPEG4 for LBR-video, AAC-LC, HE-AAC, MPEG-1 L2, MPEG-2 audio and AC3 for audio. In the current phase, the seven proponent parties are finalizing a set of training databases available to all participants. The trained models will be submitted for the validation phase in September 2011.

(2) A closely entangled activity is referred to as P.NBAMS. The respective competition targets models for single-ended video-only quality monitoring based on bitstream information for IP-based transmission. Like for P.NAMS, the model input data are PCAP-files as obtained, for example, from network measurement probes. For P.NBAMS, two modes can be distinguished: One with access to information in the encoded video bitstream and packet-header domains (mode 1), and one with access also to fully decoded pixel information (mode 2). Similarly to P.NAMS, models for the HBR and for the LBR application areas are developed separately. The packetization schemes considered in this phase are the same ones as for P.NAMS, with the difference of evaluating only the video-related information. For P.NBAMS, the target video codec is the H.264. The development is conducted in parallel with P.NAMS. Due to the strong synergies, the training databases for the video-only case for H.264 are identical, which reduces the training data creation workload for the participants in both the P.NAMS and P.NBAMS competitions (seven of the eight P.NBAMS participating institutions are the P.NAMS participants). According to the current planning, the P.NBAMS models will be submitted in September 2011, and the final standard is expected for mid 2012.

Parts of the work are conducted jointly with Q.17/12: Depending on where in the network a P.NAMS or P.NBAMS model is applied, the acquired information may be different from what is available at the client, that is, the actual media player at the user site. For example, when error-resilience methods such as Automatic Repeat reQuest (ARQ) or Forward Error Corection (FEC) are being employed, or when jitter buffers are active, loss rates at the client may substantially differ from those observed in the network. Here, Q.14/12 and Q.17/12 work on ways to convert network-observed traces into traces as they are present at the client, or to convert the parameters used by a P.NAMS or P.NBAMS model from values valid at probe location to values valid for the client location.

There is a related activity currently being conducted by VQEG. In their running hybrid video quality modelling activity, signal-based models are combined with bitstream- or packet-header-based approaches. Apart from the

additional signal-information obtained from the actual decoder employed in the video player, a further difference to P.NBAMS and P.NAMS is the fact that VQEG targets no-reference, reduced-reference and full-refernce models, while P.NAMS and P.NBAMS are no-reference.

4. Outlook

Work in SG12 is ongoing, and needs to target new challenges. For example, a fully working, conversational WB E-model will be standardized. Existing work on VoIP in Next Generation Mobile Networks must be extended to the pecularities of access technologies such as LTE (Long Term Evolution). Further, models valid only for the two-paty conversation situation need to be extended towards multiparty conferencing. For the domain of audiovisual media distribution, new distribution techniques such as HTTP-streaming, and new media formats such as 3D video must be considered in future QoE monitoring and planning models.

References

[1] ITU-T Study Group 12 Website, http://www.itu.int/ITU-T/studygroups/com12/index.asp

[2] ITU-R Website, http://www.itu.int/ITU-R/

[3] Video Quality Experts Group (VQEG) Website, http://www.its.bldrdoc.gov/vqeg/

[4] ITU–T Rec. G.107, "The E-model: A Computational Model for Use in Transmission Planning", Int. Telecomm. Union, Geneva, 2009.

[5] A. Raake, S. Möller, M. Wältermann, N. Côté, J.-P. Ramirez, "Parameter-based prediction of speech quality in listening context – Towards a WB E-model", in Proc. Second Int. Workshop on Quality of Multimedia Experience (QoMEX'10), June 21-23, Trondheim, 2010.

[6] S. Möller, F. Kettler, H.-W. Gierlich, N. Côté, A. Raake, M. Wältermann, "Extending the E-model to better capture terminal effects", in Proc. 3rd Int. Workshop on Perceptual Quality of Systems (PQS 2010), Bautzen, 2010.

[7] ITU–T Rec. P.863, "Perceptual Objective Listening Quality Assessment (POLQA)", Int. Telecomm. Union, Geneva, 2011.

[8] ITU–T Rec. P.862, "Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-end Speech Quality Assessment of Narrow-band Telephone Networks and Speech Codecs", Int. Telecomm. Union, Geneva, 2001.

[9] ITU–T Rec. P.563, "Single-ended Method for Objective Speech Quality Assessment in Narrow-band Telephony Applications", Int. Telecomm. Union, Geneva, 2004.

[10] ITU–T Rec. P.564, "Conformance Testing for Voice over IP Transmission Quality Assessment Models", Int. Telecomm. Union, Geneva, 2007.

[11] ITU–T Rec. G.1011, "Reference guide to quality of experience assessment methodologies", Int. Telecomm. Union, Geneva, 2010.

[12] ITU–T Rec. G.1070, "Opinion model for video-telephony applications", Int. Telecomm. Union, Geneva, 2007.

[13] M. Guéguin, R. Le Bouquin-Jeannès, V. Gautier-Turbin, G. Faucon, V. Barriac, "On the evaluation of the conversational speech quality in telecommunications," EURASIP Journal on Advances in Signal Processing, Vol. 2008, Article ID 18524

Alexander Alexander Raake is an Assistant Professor and heads the group for Assessment of IP-based Applications at Deutsche Telekom Labs, TU Berlin. From 2005 to 2009, he was a senior scientist at the Quality & Usability Lab of Deutsche TelekomLabs, TU Berlin. From 2004 to 2005, he was a Postdoctoral Researcher at LIMSI-CNRS in Orsay, France. From the Electrical Engineering and Information Technology Faculty of the Ruhr-Universität Bochum, he obtained his doctoral degree (Dr.-Ing.) in January 2005, with a book on the speech quality of VoIP (extended version appeared as Speech Quality of VoIP, Wiley, 2006). After his graduation in 1997, he took up research at the Technical University in Lausanne (EPFL) on ferroelectric thin films, before he joined Ruhr-Universität Bochum in 1999. Before, he studied Electrical Engineering in Aachen (RWTH) and Paris (ENST/Tlcom). His research interests are in speech, audio and video transmission, Quality of Experience assessment, audiovisual and multimedia services and user perception. Since 1999, he has been involved in the standardization activities of the International Telecommunication Union (ITU-T) on transmission performance of telephone networks and terminals, where he currently acts as a Co-Rapporteur for question Q.14/12 on monitoring models for audiovisual services.

Sebastian Möller was born in 1968 and studied electrical engineering at the universities of Bochum (Germany), Orléans (France) and Bologna (Italy). From 1994 to 2005, he

held the position of a scientific researcher at the Institute of Communication Acoustics (IKA), Ruhr-University Bochum, and worked on speech signal processing, speech technology, communication acoustics, as well as on speech communication quality aspects. Since June 2005, he works at Deutsche Telekom Laboratories, TU Berlin. He was appointed Professor at TU Berlin for the subject "Quality and Usability" in April 2007, and heads the "Quality and Usability Lab" at Deutsche Telekom Laboratories. He received a Doctor-of-Engineering degree at Ruhr-University Bochum in 1999 for his work on the assessment and prediction of speech quality in telecommunications. In 2000, he was a guest scientist at the Institut dalle Molle d'Intélligence Artificielle Perceptive (IDIAP) in Martigny (Switzerland) where he worked on the quality of speech recognition systems. He gained the qualification needed to be a professor (venia legendi) at the Faculty of Electrical Engineering and Information Technology at Ruhr-University Bochum in 2004, with a book on the quality of telephone-based spoken dialogue systems. In September 2008, we worked as a Visiting Fellow at MARCS Auditory Laboratories, University of Western Sydney (Australia) on the evaluation of avatars.
Since 1997, he has taken part in the standardisation activities of the International Telecommunication Union (ITU-T) on transmission performance of telephone networks and terminals. He is currently acting as a Rapporteur for question Q.8/12.

QoE Models for Quality Planning and Monitoring for Video Delivery

Ning Liao, Technicolor Research & Innovation, China
ning.liao@technicolor.com

1. Introduction

With the development of video service delivery over IP networks, there is a growing interest in low-complexity no-reference (NR) video quality assessment (VQA) models to measure the joint impact of transmission losses and coding configurations on the perceived video quality. In ITU-T standardization organization, there are some active work [1] on the packet-layer NR VQA model (e.g. P.NAMS [2], G.OMVS) and the bitstream-level NR VQA model (e.g. P.NBAMS [3]) for the application scenarios like IPTV and mobile streaming. The packet-level model and the bitstream-level model are of low complexity because only packet headers, coding configurations, and/or video bitstream without full decoding are used as model inputs.

Two use cases of the low-complexity NR VQA models have been identified in ITU-T SG12/Q14: video system quality planning and in-service video quality monitoring.

As a video system quality planning tool, a VQA model for system planning can help system designers to avoid over-engineering the applications, terminals, and networks while guaranteeing user's satisfactory QoE. Early in the literature [4], Verscheure et. al. pointed out that the QoE resulting both from video coding quality and network transmission impairment, should be optimized by considering the entire system, not by optimization of individual system components in isolation. For an example, they demonstrated that the video quality doesn't increase monotonically with the MPEG-2 coding bitrate (BR). Instead, for a given packet loss rate (PLR), the video quality reaches a maximum value at an optimal BR, and then goes bad even with the increase of BR. Obviously, the VQA models for video system planning should provide a relationship between the system parameters of interest (e.g. coding bitrate, frame rate, spatial resolution, PLR), and the perceptual video quality.

As a video quality monitoring tool for video delivery, a VQA model should be light-weight so as to be deployed in large scale along the video distribution chain. Operators may ensure video quality Service Level Agreement (SLA) by monitoring and diagnosing video quality degradation caused by network issues. Therefore, a further requirement for the video quality monitoring model is to predict the perceptual video quality as accurately as possible to avoid false alarm or miss alarm.

From the two different perspectives of the video system quality planning scenario and the quality monitoring scenario, in this letter, we provide our viewpoints on VQA modeling approaches for the two scenarios and on the criteria for evaluating the respective VQA model's performance.

2. Modeling approaches for quality planning and monitoring

Generally in literatures, two approaches are followed in packet-layer and bitstream-level NR VQA modeling. One is the parameter-based modeling approach [5][6][7][8]; another is the loss-distortion chain based modeling approach [9]. The parameter-based approach estimates perceptual quality by extracting the parameters of a specific application (e.g. coding bitrate, frame rate, spatial resolution) and the transmission packet loss, then building a relationship between the parameters and the overall video quality. Obviously, the parameter-based VQA model is in nature consistent with the requirement of system planning. Notice that it is video content-blind and predicts the average video quality over different video contents. The coefficient table of this model needs to change with the codec type and configuration, the error concealment strategy of a decoder, the display resolution, and the video content types.

There are some variants [10][11] of the simple parametric model, overcoming the shortage of the content-blind parametric model. Two content descriptors, namely, the scene complexity and the level of motion, are used as input parameters in model of [10]. The content descriptors are calculated based on bitstream information, e.g., the bits of the frames of different coding types and the average quantization parameter (QP). But packet loss impairment is not considered in [10][11].

Loss-distortion chain based approach [9] has the merit of accounting in error propagation, content features, and error concealment effectiveness. The frame type and the initial artifacts of the frame having packet loss are identified as the most important factors to the visibility of losses in [12][13]. Different frame type essentially leads to different error propagation effects. The initial visible artifacts are very different due to the different concealment strategies [14] and content features.

Because iteration process is generally involved in, loss-distortion chain based approach is suitable for quality monitoring, not for system planning model. Keeping low computational complexity, which is very important to in-service monitoring, is one challenge of this approach. Another challenge, in case of packet-layer VQA model, is to estimate the video content and compression information at packet layer.

In [15], we proposed a VQA model using the loss-distortion chain based approach and dealt with the above challenges. The initial visible artifacts (IVA) of a frame suffering from packet loss and error concealment, is estimated based on the error concealment effectiveness. Unlike in [16], the error concealment effectiveness is determined based on the spatiotemporal complexity estimation with packet-layer information; and the different error concealment effects are considered. Then, the IVA is incorporated into an error propagation model to predict the overall video quality. The estimate of spatiotemporal complexity is used to modulate the propagation of the IVA in the error propagation model. Experiment results showed that the performance gain comes primarily from the accurate assessment of initial visible artifacts and then from incorporating an error propagation model.

3. Criteria and methods of building evaluation datasets for the two scenarios

The video system planning model is for network QoS parameter planning and video codec parameter planning, given a target video quality. It may predict average perceptual quality degradation, ignoring the impact of different distortion and content types on the perceived quality. Therefore, it should predict well the quality of the loss-affected sequences with large occurrence probability. Whereas, the VQA model for monitoring purpose is expected to give quality degradation alarm with high accuracy and should be able to estimate as accurate as possible the quality of each specific video sequence distorted by packet losses. Correspondingly, the respective subjective datasets for evaluating the planning model and the monitoring model should be built differently.

In [15], we chose the processed video sequences (PVSs) for subjective test according to two different criteria:

Criterion 1): for each video content, select the PVSs that are representatives of the dominant MOS-PLR distribution as done in [17];

Criterion 2): for each video content, select the PVSs that cover the MOS-PLR distribution widely by including the PVSs of the best and the poorest quality at a given PLR level, in addition to those representing the dominant MOS-PLR distribution.

The PVSs in dataset-2 (built by criterion 2) present much more diverse relationship between PLR and subjective video quality than those in dataset-1 (by criterion 1). In dataset-2, the PVSs present very different perceptual quality even under the same PLR. It is shown that the parameters PLR/ BLF (Burst Loss Frequency) [7] /IFR (Invalid Frame Ratio) [18] used in existing models may be effective for video system quality planning modeling, but are not effective for video quality monitoring applications. Our proposed distortion-chain based VQA model demonstrated robust performance improvement compared with the existing parameter-based metrics on both dataset-1 and dataset-2.

4. Conclusions

In this letter, we differentiate the requirements for VQA models from two typical scenarios in video delivery, namely, video system quality planning scenario and the quality monitoring scenario. Further, we discussed our viewpoints on the VQA modeling approaches for the two scenarios and proposed using different criteria to construct the evaluation datasets for the respective scenario's VQA model. We state the argument that the dataset for evaluating the video quality monitoring model should be more challenging than that for video system planning model.

References

[1] A. Takahashi, D. Hands, V. Barriac, "Standardization activities in the ITU for a QoE assessment of IPTV", IEEE Communications Magazine, 46(2), 78-84, (2008)
[2] ITU-T document, "Draft terms of reference (ToR) for P.NAMS," Oct. 2009, available at http://www.itu.int/md/meetingdoc.asp?lang=en&parent=T09-SG12-091103-TD-GEN-0146.
[3] ITU-T document, "Draft Terms of Reference (ToR) for P.NBAMS", March 2009, available at http://www.itu.int/md/T09-SG12-110118-TD-GEN-0521
[4] O. Verscheure, P. Frossard, and M. Hamdi, "MPEG-2 video services over packet networks: Joint effect of encoding rate and data loss on user-oriented QoS," in NOSSDAV, 257–264, (1998).
[5] S. Mohamed and G. Rubino, "A study of real-time packet video quality using random neural networks," IEEE Transactions on Circuits and Systems for Video Technology, 12(12), 1071-1083 (2002).
[6] K. Yamagishi, T. Hayashi, "Video-quality planning model for videophone services," Information and Media Technologies 4(1): 1-9 (2009)
[7] K. Yamagishi and T. Hayashi, "Parametric Packet-Layer Model for Monitoring Video Quality of IPTV Services," IEEE International Conference on Communications, 110-114, (2008).
[8] A. Raake, M.-N. Garcia, S. Moller J. Berger, F. Kling, P. List, J. Johann, C. Heidemann, "T-V-Model: Parameter-Based Prediction of IPTV Quality," Proc. ICASSP, 1149-1152, (2008).
[9] A. R. Reibman, V. A. Vaishampayan, and Y. Sermadevi, "Quality monitoring of video over a packet network", IEEE Trans. on Multimedia, 6(2), 327-334, (2004)
[10] J. Hu and H. wildfeuer, "Use of content complexity factors in video over IP quality monitoring", International workshop QoMEX, 216-221, (2009)
[11] M. N. Garcia, R. Schleicher, A. Raake, "Towards a content-based parametric video quality model for IPTV", in VPQM, 2010.
[12] S. Kanumuri, P. C. Cosman, A. R. Reibman, V. A. Vaishampayan, "Modeling packet loss visibility in MPEG-2 video," IEEE Transactions on Multimedia, 8(2), 341-355, (2006).
[13] S. Kanumuri, S. G. Subramanian, P. C. Cosman, A. R. Reibman, "Predicting H.264 packet loss visibility using a generalized linear model", Proc. ICIP, 2245-2248, (2006).
[14] A. R. Reibman, D. Poole, "Predicting packet-loss visibility using scene characteristics", Packet Video, 308-317, (2007)
[15] N. Liao, Z.B. Chen, "A packet-layer video quality assessment model with spatiotemporal complexity estimation", VCIP, (2010)
[16] T. Yamada, Y. Miyamoto, and M. Serizawa, "No-reference video quality estimation based on error-concealment effectiveness", Packet Video, 288-293, (2007)
[17] F. D. Simone, M. Naccari, M. Tagliasacchi, F. Dufaux, S. Tubaro, T. Ebrahimi, "Subjective assessment of H.264/AVC video sequences transmitted over a noisy channel," Proc. International Workshop on Quality of Multimedia Experience (QoMEx), 204-209, (2009), available at http://mmspl.epfl.ch/
[18] T. Hayashi†, M. Masuda, T. Tominaga, and K. Yamagishi, "Non-intrusive QoS monitoring method for realtime telecommunication services," NTT technical review, 4(4), 35-40, (2006).

Ning Liao is with Research & Innovation Dept. of Technicolor, Beijing, China. She received the B.E. degree in wireless communication and the Ph.D. degree in Telecommunication Engineering from the Beijing University of Posts & Telecommunications, China, in 1998, and in 2007 respectively. From 1998 to 2001, she was a network engineer in China Telecom. She worked on scalable video compression, error resilient video coding and transmission, and power-efficient service scheduling in WiMAX network from 2003 to 2008. Her research interests are in the areas of video quality assessment, video coding, and wireless multimedia communications.

Objective quality assessment for 3-D video delivery

Maria G. Martini and Chaminda T. E. R. Hewage
Kingston University London, UK
m.martini@kingston.ac.uk, c.hewage@kingston.ac.uk

1. Introduction

Three-dimensional (3-D) video has gone a step beyond conventional two-dimensional (2-D) video, by providing the depth sensation to viewers. Recent advances in 3-D acquisition technologies, image processing, multimedia delivery and displays, made the provision of services such as 3-D television (3DTV) and Free View point Video (FVV) technically feasible. 3-D video applications have the potential to provide high quality immersive experience for a variety of applications in the home, workplace, and public places, with the possibility to use a range of transmission networks including wireless and mobile systems.

For the assessment and design of 3-D video delivery systems, it is crucial to rely on an objective metric well representing the human perception. Even though there are 2-D objective quality models that are highly correlated with the Human Visual System (HVS), these cannot be directly employed in measuring 3-D video quality. Due to the unavailability of an accurate objective quality metric for 3-D video, it is currently difficult to evaluate the perceptual quality of such applications without resorting to full subjective test campaigns. These subjective evaluation tests make use of human observers and therefore require more time and effort than objective quality measurements. Furthermore, these cannot be implemented algorithmically and cannot thus be used for on-line assessment with the purpose of system adaptation. The lack of reliable objective metrics has negative effects on the development and advancement of 3-D video technologies, and the roll-out of 3-D services and commercial products.

This letter addresses recent research on objective 3-D video quality assessment based on subjective observations, i.e., based on the study of the characteristics of the human visual system. The importance of no-reference and reduced-reference metrics for the evaluation and adaptation of media delivery systems is also highlighted.

.

2. Objective 3-D video quality assessment based on subjective observations

The effects of the different artefacts introduced by image capturing, pre-processing, coding and delivery methods on the perceived quality of video are diverse in nature. Pixel-based quality measurements like Mean Absolute Difference (MAD) and Mean Square Error (MSE)/peak signal-to-noise ratio (PSNR) [1] not always reflect the true image quality as perceived by the Human Visual System (HVS). Hence, image quality models which incorporate different dimensions of the HVS have emerged to provide more realistic quality measurements for conventional 2-D video [2] [3]. Such quality models define the relationship between the parameters of the system (e.g., coding and delivery method) and the perceived image quality.

Unlike conventional 2-D video quality, the perceived 3-D video quality is multi-dimensional in nature. 3-D video quality can be considered as a combination of a number of perceptual attributes such as naturalness, presence, stereo impairments, comfort, and perceived depth. The existing 2-D objective quality measures of individual views (e.g., multi-view video captured by clusters of cameras) may not represent the true 3-D image quality as perceived by human viewers. The PSNR of separate views can be used for assessing 3-D video quality and this results in a simple metric that could represent a good starting point for objective quality assessment for 3-D video. Subjective studies have shown indeed that PSNR of colour and depth maps provides a good indication of the global 3D quality [4] [5]. However, some limitations of PSNR have been demonstrated in tests by the Video Quality Experts Group (VQEG) for 2-D video and similar limitations exist for the assessment of image quality in 3-D video. Future research should thus focus on metrics better correlated with human perception.

The attributes associated with 2-D video (e.g., blocking, blurring) cannot be uniquely used in measuring the perceptually important features of 3-D video, such as overall image quality, sharpness, naturalness and depth perception.

Therefore, 3-D quality models which account for depth reproduction in immersive video are required.

A number of 3-D video quality evaluation studies can be found in the research literature. The authors report in [6] that JPEG coding of stereoscopic video has an effect on the overall image quality, but no effect on the perceived depth. Similarly in [7] the added value of depth is not taken into account when assessing the perceived image quality of MPEG-2 coded stereo sequences. However, a positive relationship between depth and perceived image quality for uncompressed stereoscopic images is discussed in [8]. An objective quality metric for 3-D video production in a studio environment is described in [9].

These studies are specific for a certain application and set of system parameters, and do not provide a generic 3-D quality model that covers the whole 3-D application chain from capture to display.

A possible solution proposed for 3-D video quality assessment is the application of 2-D video models to the different components of 3-D video (e.g., left + right view and colour + depth components), including 3-D characteristics, such as depth perception [10][11][4]. A detailed study on quality evaluation of colour plus depth map-based stereoscopic video is presented in [5], where the correlation of the subjective Mean Opinion Score (MOS) with different objective quality metrics is analyzed.

3. NO-reference and reduced-reference metrics

For 3-D media delivery, quality measures can be used as feedback information to adapt video transmission and to change accordingly the system parameters “on the fly”. In this case quality has to be measured at the receiver side [12] [13]. Full Reference (FR) objective quality metrics are not suitable in this case, since the original 3-D video sequence is not available at the receiver side for the purpose of quality assessment. Therefore, Reduced Reference (RR) or No Reference (NR) quality metrics are necessary to tackle this problem with limited or no access to the original 3-D video sequences. RR and NR quality metrics for 2-D video are still hot research topics and some studies can be found in the literature related to the proposed NR and RR quality metrics for 2-D video [14]. Most of these metrics are based on existing objective quality metrics such as PSNR and structural similarity metric (SSIM) [3].

A new reduced-reference metric for 3-D video has been developed by the authors [15] and new metrics are under development. In [15] we proposed a Reduced-Reference quality metric for depth maps associated with colour-plus-depth 3-D video based on edge detection. The depth map of a colour-plus-depth 3-D video sequence determines the position of the corresponding colour image in the 3-D space, in particular in the z-direction (Depth Image-Based-Rendering, DIBR).

Therefore, the quality of depth maps is crucial for 3-D perception since these are used at the receiver side to render novel views. Since edges and contours of the depth map represent discontinuities in depth levels, relying on perceptual studies [16] we considered this as the most critical information for the human visual system and we used such edge information (i.e, binary edge mask generated by using Sobel filtering) as side information for the assessment of the quality at the receiver side. Assuming delivery of compressed 3-D video over a network, we compared the binary edge mask of the original video sequence, transmitted as side information, with the binary edge mask of the received video sequence. A schematic representation is reported in Figure 1. To reduce the overhead associated to side information, binary edge maps could be compressed, for instance through run-length encoding and/or only edge information from selected portions of the depth maps can be transmitted.

In order to evaluate the performance of the proposed Reduced-Reference quality metric for depth maps, we performed experiments for a range of compression levels (i.e., with different QP values). The Orbi and Interview 3-D test sequences (depth maps of these sequences) were encoded using the H.264/AVC video coding standard, with the aid of the JSVM reference software (Version 9.12). Ten seconds long sequences (250 frames) were encoded using different QP values (1, 5, 10, 15, 20, 25, 30, 35, 40, 45 and 50). For each QP value, the quality was measured by using both the PSNR of the depth maps (Full-Reference method) and the PSNR of the edge maps (Reduced-Reference

method) generated for the reference image as well as for the processed image. In [17] transmission of the encoded bit-stream over an IP core network was simulated using IP error patterns generated for Internet experiments [18].

It was observed in [11] [4] [5] that PSNR provides a good indication of the overall image quality and depth perception, when depth maps are impaired by compression and packet losses, for a range of compression rates and PLRs.

We reported our results in scatter plots, with the PSNR metric based on the comparison of the edge maps on one axis and the PSNR metric based on the comparison of the complete depth maps on the other, and subsequently performed polynomial fitting.

The derived metric shows a very good correlation with the PSNR of depth maps, and hence with depth perception, for different sequences. The proposed metric can thus be used in substitution of Full-Reference metrics (that cannot be used for on-the-fly closed loop adaptation of multimedia delivery systems) to represent the perceived 3-D quality when depth maps are compressed and/or received with errors, with a limited overhead due to the transmission of edge maps information.

4. Summary and outlook

We have briefly reviewed the emerging area of objective 3-D video quality assessment for 3-D video delivery, as well as a first study on reduced reference metrics for 3-D video. An important direction for future work is to improve quality models for 3-D video to better represent the global quality experienced by the user, to enable effective assessment of the quality also when the original video sequence is not available as reference.

Acknowledge

The activity of the authors on (3-D) video quality assessment was partially supported by the European Commission (FP7 project OPTIMIX INFSO-ICT-21462).

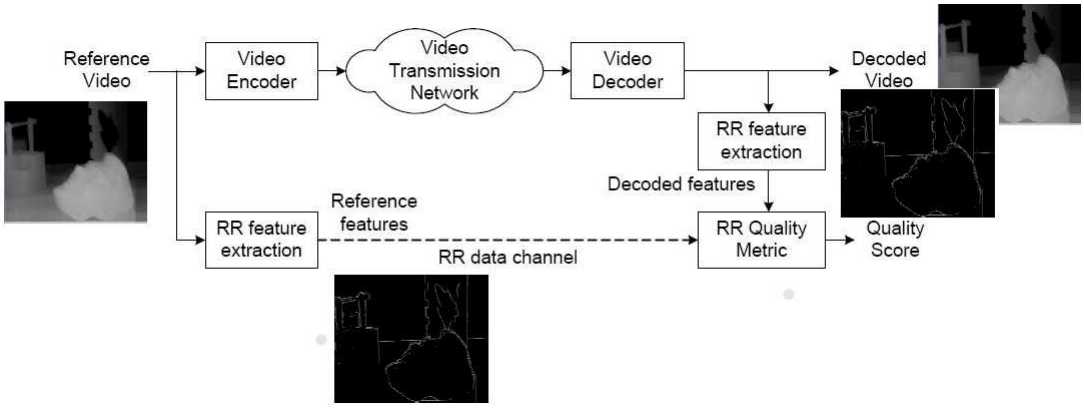


Figure 1: Schematic representation of the proposed reduced reference metric for depth maps associated to colour plus depth 3-D video

References

[1] A. M. Eskicioglu and P. S. Fisher, "Image quality measures and their performance," *IEEE Transactions on Comms.*, vol. 43, pp. 2959–2965, Dec. 1995.

[2] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Transactions on Broadcasting*, vol. 50, no. 3, pp. 312–322, Sept. 2004.

[3] Z. Wang, A. Bovik, H. Sheikh, and E.

Simoncelli, "Image quality assessment: from error measurement to structural similarity," *IEEE Trans. Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[4] S. L. P. Yasakethu, C. T. E. R. Hewage, W. A. C. Fernando, S. T. Worrall, and A. M. Kondoz, "Quality analysis for 3D video using 2D video quality models," *IEEE Transactions on Consumer Electronics*, vol. 54, no. 4, pp. 1969–1976, Nov. 2008.

[5] C. T. E. R. Hewage, S. T. Worrall, S. Dogan, S. Villette, and A. M. Kondoz, "Quality evaluation of colour plus depth map-based stereoscopic video," *IEEE Journal of Selected Topics in Signal Processing - Special Issue on Visual Media Quality Assessment*, vol. 3, no. 2, pp. 308–318, Apr. 2009.

[6] P. Seuntiens, L. Meesters, and W. Ijsselsteijn, "Perceptual evaluation of JPEG coded stereoscopic images," in *Proc. SPIE*, 2003, vol. 5006, pp. 215–226.

[7] W. Tam, L. Stelmach, and P. Corriveau, "Psychovisual aspects of viewing stereoscopic video sequences," in *Proc. SPIE*, 1998, vol. 3295, pp. 226–235.

[8] H. de Ridder W. IJsselsteijn and J. Vliegen, "Subjective evaluation of stereoscopic images: Effects of camera parameters and display duration," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 10, pp. 225–233, 2000.

[9] J. Kilner J. Starck and A. Hilton, "Objective quality assessment in free-viewpoint video production," in *3DTV Conference*, 2008.

[10] A. Tikanmaki, A. Smolic, K. Muller, and A. Gotchev, "Quality assessment of 3D video in rate allocation experiments," *in IEEE International Symposium on Consumer Electronics*, Algarve, Portugal, 2008.

[11] C. T. E. R. Hewage, S. T. Worrall, S. Dogan, and A. M. Kondoz, "Prediction of stereoscopic video quality using objective quality models of 2-D video," *IET Electronics Letters*, vol. 44, no. 16, pp. 963–965, July 2008.

[12] M. G. Martini, M. Mazzotti, C. Lamy-Bergot, J. Huusko, and P. Amon, "Content adaptive network aware joint optimization of wireless video transmission," *IEEE Communications Magazine*, vol. 45, no. 1, pp. 84– 90, January 2007.

[13] C. T. E. R. Hewage, S. Nasir, S. Worrall, and M. G. Martini, "Prioritized 3D video distribution over IEEE 802.11e," *in Proc. Future Network & Mobile Summit*, Florence, Italy, June 2010.

[14] Z. Wang and E. P. Simoncelli, "Reduced-reference image quality assessment using a wavelet-domain natural image statistic model," *Human vision and Electronic Imaging*, pp. 149–159, March 2005.

[15] C. T. E. R. Hewage and M. G. Martini, "Reduced-reference quality evaluation for compressed depth maps associated with colour plus depth 3D video," in *Proc. IEEE Int. Conf. on Image Processing (ICIP 2010)*, Hong Kong, Oct. 2010.

[16] D. Marr and E. Hildreth, "Theory of edge detection," in *Proceedings of the Royal Society of London. Series B*, 1980.

[17] C. T. E. R. Hewage and M. G. Martini, "Reduced-reference quality metric for 3D depth map wireless transmission," in *Proc. 3DTV Conference*, Tampere, Finland, June 2010.

[18] S. Wenger, "Error patterns for Internet video experiments," in *ITU-T SG16 Document Q15-I-16-R1*, 1999.

Maria G. Martini is a Senior Lecturer in the Faculty of Computing, Information Systems and Mathematics in Kingston University, London, where she is also coordinating the Wireless Multimedia Networking Research Group. She received the Laurea in electronic engineering (summa cum laude) from the University of Perugia (Italy) in 1998 and the Ph.D. in Electronics and Computer Science from the University of Bologna (Italy) in 2002. In 1998-2003 she was with DEIS, Univ. of Bologna. In 2004–2007 she was with CNIT, Italy. She has worked as a key person for national and international projects, such as the JSCC project, with Philips Research, the JOCO and PHOENIX European IST projects and she is currently leading the KU team in the ICT-OPTIMIX European project. She is an editor/reviewer for international journals and she is/was in the organizing and programme committee of several international conferences. She was the general co-chair of the ICST/ACM MOBIMEDIA 2009 conference. She is coordinating the edition of the Strategic Applications Agenda (SAA) on mobile health and inclusion applications in the eMobility European Technology Platform. Her research interests include wireless multimedia networks, cross-layer design, joint source and channel coding, error resilient video, video quality assessment and medical applications. She is the inventor of several patents on wireless video.

Chaminda T.E.R. Hewage received the B.Sc. (Hons.) degree in Electrical and Information Engineering from the University of Ruhuna, Galle, Sri Lanka. During 2004-2005, he worked as a Telecommunication Engineer in the field of Data Communication. He completed his Ph.D. at Centre for Communication Systems Research, University of Surrey, Guildford, UK. His research aims at providing QoS support

for 3-D video communication application. He also worked on VISNET II Network of Excellence (NoE) of IST FP6 programme in the area of robust immersive media communications. From September 2009 he is attached to Wireless, Multimedia and Networking Research Group at University of Kingston, London, UK. He is a member of IEEE and IET. He was awarded a gold medal by University of Ruhuna, Sri Lanka, for his achievements in Engineering discipline at the general convocation held in 2004

QoE Aware Mobile Video Delivery

Tasos Dagiuklas, Ilias Politis, and Lampros Dounis
TEI of Mesolonghi, Greece
ntan@teimes.gr, ilpolitis@gmail.com, dounis@gmail.com

1. Introduction

Mobile Video Delivery has received particular attention the recent years, due to rapid growth of wireless systems such as WLANs, 3G, 3G+, LTE, WiMax etc [1]. Video transmission over heterogeneous wireless networks poses many research challenges, including the issues of coping with losses due to both physical impairments and network congestion, as well as maintaining of QoS and session continuity while the user hands off across inter-technology systems (e.g. vertical handover) [2].

In this position paper, we discuss some key research challenges for the QoE (Quality of Experience) maintenance in mobile video delivery across heterogeneous wireless networks. In particular, we discuss the need for QoE-aware mobility and the need for the tight integration between handover, network selection and rate control.

2. Seamless Mobility

Handover is the process of network association and connection maintenance of a Mobile Terminal while it moves across different access points [3]:

- Nomadicity: It is the ability of the user to change his network point of attachment while he/she is on the move.
- Session Continuity: It is the ability that the mobile terminal can switch to a new network point of attachment while maintaining the ongoing session towards the new point of attachment.
- Seamless handoff: It aims to minimize the packet loss while the session is associated with the new point of attachment. It is sometimes referred to as smooth handoff.

One of the standards that have specified a framework for seamless mobility is the IEEE 802.21 standard. IEEE 802.21 proposes the MIH (Media Independent Handover) framework where mobile nodes and the network exchange information and commands for an optimal handover [4]. Moreover, for hiding the heterogeneity of the MAC and physical layers, MIH inserts an intermediate layer between layer 3 (and above) and the divert Layer 2 technology specifics, the Media Independent Handover Function (MIHF). The MIH framework describes three different types of communication that act as services: Event Service, Command Service and Information Service, as illustrated in the following figure.

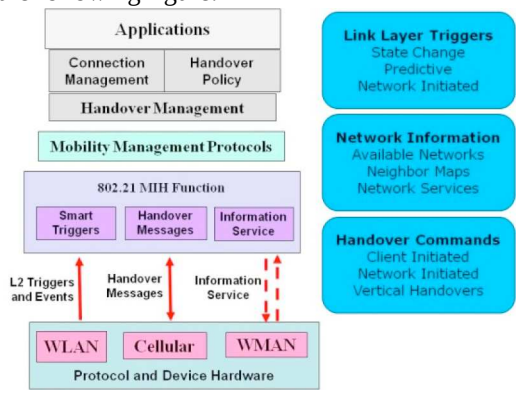


Fig. 2 IEEE 802.21 Media Independent Handover framework

The Media Independent Event Service (MIES) is a communication procedure where indications for events are passed to the MIH users for further handling. The Media Independent Command Service (MICS) provides a set of handover commands. The Media Independent Information Service (MIIS) is a database that contains all the available information from both the Mobile Terminal and the Networks.

QoE Handover Triggering

In heterogeneous wireless networking environment, QoE maintenance of on-going video sessions is a challenging task. This is due to the fact that both networking conditions (congestion, physical impairments) and end-user's mobility may deteriorate the perceived video QoS. These are scenarios for handover from QoE based triggering [5].

- Network Load increases at the current wireless network, leading to congestion and packet loss.
- Application Deteriorates. Video QoS Monitoring (MOS, PSNR) can generate alarms when these parameters are decreased.

Application QoE Monitoring

Monitoring of QoE is based on the RTP Control Protocol Extended Report (RTCP-XR) as defined in [6], which provides a useful set of fields providing information for video performance analysis. Important information that is related to video is the following:

- The packet loss within I/B/P frames.
- Knowledge of GoP structure and key coding parameters to estimate PSNR

RTCP-XR can be used in order to send specific information about the packet loss events in each frame. This information can be used for the PSNR estimation. The PSNR can be estimated using a distortion prediction model. More information about video distortion model and video quality monitoring can be found in [7], [8].

Network Selection

The decision of network selection is based on QoS parameters/criteria that ought to be optimized depending on the available access networks. Such decision can use a cost function that includes the rules and policies for selecting the best candidate network or for adapting ongoing session parameters (increase QoS, reduce the number of handovers).

The network selection scheme can use Multi Attribute Decision Making (MADM) algorithms methods [9], in order to determine the weights of the criteria (packet loss, throughput), which will be used for the final access network ranking.

Handover Process

The handover process will consist of the following steps:

- Handover decision: The mobile terminal related information (e.g. SNR, delay, jitter, PSNR, packet loss, etc), are collected and sent to the MIH server. MIH server is also responsible of collecting corresponding information from neighboring networks.
- Handover initiation: These collected MIH parameters will be evaluated and compared against a set of predetermined threshold values. These thresholds are determined by the network provider and are specified in each user's profile.
- Handover execution: The final stage is the execution of the vertical handover to the decided neighboring network. The Mobile IP platform is responsible of handling the vertical handover and of ensuring seamless service continuity.

3. Seamless Mobility with Rate Control

Seamless handover can be combined with rate adaptation in order to optimize QoS/QoE of an on-going video session. In case, where the mobile user moves to a network with less available bandwidth than the current one, seamless handoff can be combined with rate control so that QoE/QoS is optimized. This can be transformed to an optimization problem (determine the best Quantization Parameter that optimizes PSNR under certain networking conditions).

Rate control aims either to optimize Rate-Distortion at the encoder, or to find an optimum QP by taking into account rate-distortion and frame complexity. Many of these algorithms do not take into account the variation of network conditions and do not operate in real-time. In order to combine seamless handoff with rate control, it is necessary to obtain a formula that describes the relation between PSNR (QP, Available Bandwidth) [10]. This relation can be found through experimentation, as illustrated in Fig. 2. The PSNR reaches a certain peak value, where both coding distortion and packet loss have the least impact on the perceived video quality for a given available bandwidth.

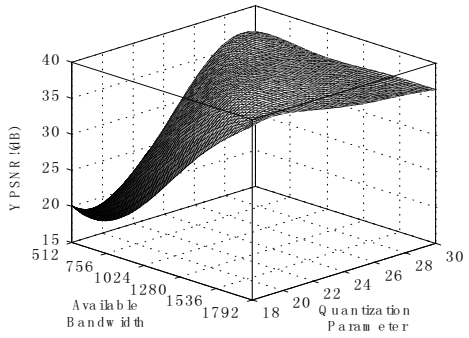


Fig. 3 Surface fitted model of PSNR vs QP under different network conditions (available BW for video transmission)

The above can be verified from the following figure. It illustrates perceived video QoS, while the users hands off among 3 different wireless networks (2 WLANs and 3G) that are also stressed with background traffic. It is demonstrated that the combination of seamless handoff, QoE-handover trigger, network selection and rate control leads to optimizing QoE in mobile video delivery.

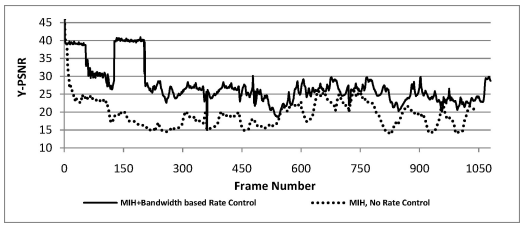


Fig. 4 Perceived QoE/QoS with seamless handoff scenarios (MIH with no rate control vs MIH with rate control)

4. Conclusions

Future challenges in mobile video delivery include: exploit SVC capabilities in mobile video delivery (in case where the user moves to a network with limited bandwidth, several layers may be dropped), combine multi-homing and path diversity with streaming by taking into account end-user's preferences

References

[1] M. Etoh and T. Yoshimura, "Advances in Mobile Video Delivery", *Proc. Of IEEE*, Vol. 93, no. 1, pp. 111–122, June 2005.

[2] Q. Zhang, W. Zhu and Y. Zhang, "End-to-End QoS for Video Delivery Over Wireless Internet", *Proc. Of IEEE*, Vol. 93, no. 1, pp. 123–134, June 2005.

[3] N. Nasser, A. Hasswa and H. Hassanein, "Handoffs in fourth generation heterogeneous networks", *IEEE Communications Magazine*, Vol. 44, no. 10, pp. 96-103, October 2006

[4] IEEE 802.21/D10.0. Draft Standard for Local and Metropolitan Area Networks: Media Independent Handover Services, IEEE Draft, April 2008.

[5] J. Rodriguez, M. Tsagkaropoulos, I. Politis, T. Dagiuklas and S. Kotospoulos, "A middleware architecture supporting seamless and secure multimedia services across an intertechnology radio access network", *IEEE Wireless Communications Magazine*, Vol. 16, no. 5, pp. 24-31, October 2009

[6] IETF draft-ietf-avt-rtcpxr-video-02.txt, "RTCP XR Video Metrics", http://www.ietf.org/internet-drafts/draft-ietf-avt-rtcpxr-video-02.txt, November 2007

[7] S. Tao, J. Apostolopoulos and R. Guerin, "Real-Time Monitoring of video quality in IP networks", IEEE/ACM Trans. on Networking, Vol. 16, October 2008.

[8] I. Politis, M. Tsagkaropoulos, T. Pliakas, and T. Dagiuklas, "Distortion Optimized Packet Scheduling and Prioritization of Multiple Video Streams over 802.11e Networks", Advances in Multimedia, 2007.

[9] C. L. Hwang and K. Yoon, *Multiple Attribute Decision Making: Mehtods and Applications*, Springer Verlag, 1981.

[10] L. Dounis, T. Dagiuklas and I. Politis, "On the Comparison of Real-Time Rate Control Schemes for H.264/AVC Video Streams over IP-based networks using network feedbacks", To be presented at *IEEE ICC*, Kyoto, Japan, 2011

Tasos Dagiuklas (www.tesyd.teimes.gr/cones) received the Engineering Degree from the University of Patras-Greece in 1989, the M.Sc. from the University of Manchester-UK in 1991 and the Ph.D. from the University of Essex-UK in 1995, all in Electrical Engineering. Currently, he is employed as Assistant Professor at the Department of Telecommunications Systems and Net- works, Technological Educational Institute (TEI) of Mesolonghi, Greece. He is the Leader of the Converged Networks and Services Research Group. He is also Senior Research Associate within the Wireless Telecommunications Laboratory of the Electrical and Computer Engineering Department at the University of Patras, Greece. Past Positions include teaching Staff at the University of Aegean, Department of Information and Communications Systems Engineering, Greece, senior posts at INTRACOM and OTE, Greece. He has been involved in several EC R&D Research Projects under FP5, FP6 and FP7 research frameworks, in the fields of All-IP network and next generation services. He has served as TPC member to more than 30 international conferences. His research interests include Future Internet architectures and converged multimedia services over fixed-mobile networks. Dr Dagiuklas has published more than 90 papers at international journals, conferences and standardization for a in the above fields. He is a member of IEEE and Technical Chamber of Greece.

Ilias Politis received his BSc in Electronic Engineering from the Queen Marry College London in 2000, his MSc in Mobile and Personal Communications from King's College London in 2001 and his PhD in Multimedia Communications from University of Patras Greece in 2009. His is currently an Adjunct Lecturer at the Dept. of Computer Science of University of Piraeus, Greece, Scientific Associate with the Dept. of Telecommunications Systems and Networks, at Techological and Educational Institute of Messolonghi, Greece and a post-doc research associate at University of Patras, Greece. Dr. Politis research interests include multimedia communications with emphasis on video transmission optimization, video coding and heterogeneous networking. His is a member of the IEEE, FITCE and the Technical Chamber of Greece.

Lampros Dounis received his B.Sc. degree in networking and data communications from the Technological Educational Institute (TEI) of Mesolonghi, Greece in 2009.

Currently he is a M.Sc student at TEI of Pireaus (partnership with University of Kingston, UK), a researcher at the Wireless Telecommunications Laboratory of the Electrical and Computer Engineering Department at the University of Patras, Greece and a systems integration engineer at Bytemobile European Development Center. His research interests include video transmission analysis and optimization over heterogeneous networks, seamless mobility, converged networks, and programming languages.

Next-Generation Multimedia: 3D and Beyond

Xiaoqing Zhu, Cisco Systems Inc., USA
xiaoqzhu@cisco.com

Recent years have seen rapid advances in multimedia technologies. For success stories, one can readily point to the wide support of H.264/AVC in video codec software and chipsets, the prevalent availability of both high-end and low-end video conferencing systems, as well as the blooming market of 3D movies. A natural question to ask then is: What's next? In this special issue, we shall indulge in such a futuristic mood, and sample a few possibilities as suggested by our invited articles. Given the vast scope of multimedia research, we have confined our discussions in this issue to topics closely related to 3D contents.

The first article is from Professor Masayuki Tanimoto at Nagoya University in Japan. His group has carried out much of the pioneering work on free-viewpoint Television (FTV). As envisioned in the article, "FTV: Free-viewpoint Television", users in the future will be able to look at 3D contents captured by a finite number of cameras from an arbitrarily chosen viewpoint. The key idea behind this is to integrate relevant rays captured by multiple cameras into the target view, while calculating values of the missing rays by means of interpolation. The invited article in this special issue provides an excellent overview of FTV, covering various aspects such as design principles, example systems, and standardization efforts. In addition, the list of references therein provides valuable pointers for a deeper dive into this exciting topic.

For stereoscopic data captured by camera pairs, an accurate and robust depth estimation algorithm is crucial for efficient content representation. Of equal importance is the algorithm used for virtual view synthesis, which allows stereoscopic data to be rendered across different 3D display systems. These problems are discussed at length in the second article, "Next-generation 3D: From Depth Estimation to the Display", by Ramsin Khoshabeh, Can Bal, Ankit Jain, Lam Tran, Stanley Chan, and Prof. Truong Nguyen from University of California at San Diego. The authors survey the pros and cons of existing approaches for depth estimation and virtual view synthesis. Moreover, they describe their recent research on ensuring consistent depth estimation across consecutive video frames, by leveraging smoothing techniques based on total variation minimization. As noted in the article, mature methods for depth estimation and virtual view synthesis will help to push the horizon of innovations forward in the world of 3D research.

Diverting from conventional 3D applications, the third article in this special issue presents the use of augmented reality for virtual try-on of clothes. The enabling technology of this novel application is explained in "Realistic Virtual Try-On of Clothes using Real-Time Augmented Reality Methods", by Prof. Peter Eisert and Anna Hilsmann from Fraunhofer HHI, Humboldt University in Germany. Instead of relying on computationally expensive 3D geometric models, the proposed scheme combines estimation of surface deformation and shading with image-based retexturing. The result is a low-complexity, high-image-quality system that takes in real-time video inputs and computes highly natural image outputs showing virtual try-on effects for the user.

There is no doubt that the above articles can only, if anything, provide a few teasing sampling points of next-generation multimedia. Nevertheless, these are compelling sampling points. They showcase the endless possibilities lying ahead of us in this exciting field of research. Our special thanks to all the contributing authors for sharing with us updates from their research projects, along with their own insightful views for 3D multimedia in the future.

Xiaoqing Zhu is with the Department of Research & Advanced Development at Cisco Systems Inc. She received the B.E. degree in Electronics Engineering from Tsinghua University, Beijing, China, in 2001. She earned both M.S. and Ph.D. degrees in Electrical Engineering from Stanford University in 2002 and 2009, respectively. She was awarded the Stanford Graduate Fellowship from 2001 to 2005, and was recipient of the best student paper award in ACM Multimedia 2007. She interned at IBM Almaden Research Center in the summer of 2003. She was with Sharp Labs of America in the summer of 2006. Her research interests include wireless video networking, Internet video delivery, and resource allocation for distributed systems.

FTV: Free-viewpoint Television

Masayuki Tanimoto, Nagoya University, Japan
tanimoto@nuee.nagoya-u.ac.jp

1. Introduction

FTV (Free-viewpoint TV) [1]-[3] enables us to view a 3D scene by freely changing our viewpoints as if we were there. FTV is the ultimate 3DTV that transmits the infinite number of views and ranked as the top of media. It is also the best interface between human and environment, and an innovative tool to create new types of content and art.

We proposed the concept of FTV and verified its feasibility with the world's first real-time system including the complete chain of operation from image capture to display [4]. FTV with audio was realized by adding free listening-point function [5].

FTV is based on the ray-space method. We developed ray capture, processing, and display technologies for FTV. All-around ray-reproducing 3DTV has been realized by using these technologies [6].

The international standardization of FTV has been conducted as MVC and 3DV in MPEG.

2. Principle of FTV

FTV transmits a finite number of views captured by cameras and the other views at non-camera positions are generated. This free-viewpoint image generation is performed by integration and interpolation of rays. As shown in Fig.1, we need all rays passing through a viewpoint for free-viewpoint image generation. Although some of the rays are captured by cameras, the other rays are not captured by any camera. The rays captured by cameras are integrated and the missing rays are obtained by interpolation.

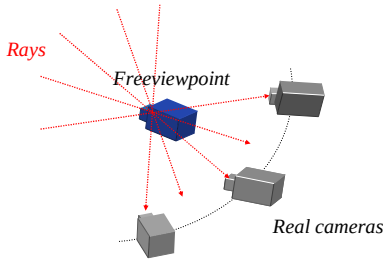


Fig. 1. Rays for free-viewpoint image generation.

This process is performed systematically in the ray-space. For the linear camera arrangement, the ray-space is constructed by placing captured camera views upright and parallel, as shown in Fig. 2, and filling the vacancy between views. An example of filled ray-space is shown in Fig.3. As seen in this figure, the horizontal cross-section of ray-space has a line structure. The line structure is used for the ray interpolation. A free viewpoint image is generated by cutting the ray-space vertically with a knife at a position determined by the viewpoint. Several parallel knifes are used to cut the ray-space to generate view images for a 3D display.

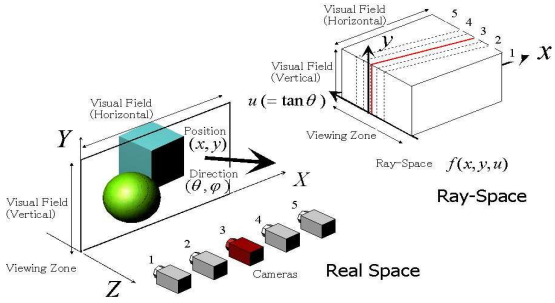


Fig. 2. Acquisition of orthogonal ray-space.

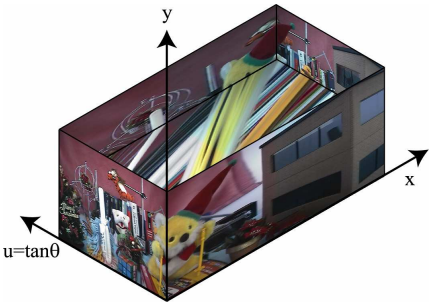


Fig. 3. Typical example of orthogonal ray-space and a horizontal cross-section.

3. FTV System

A "100-camera system" was developed to capture rays of a large 3D space. The camera arrangement is flexible as shown in Fig. 4.

The misalignment and color difference of camera views are corrected for the integration and interpolation of rays.

Free-viewpoint images can be generated in real-time on a laptop PC or a mobile player.

Various types of user interface as shown in Fig. 5 were developed for FTV.

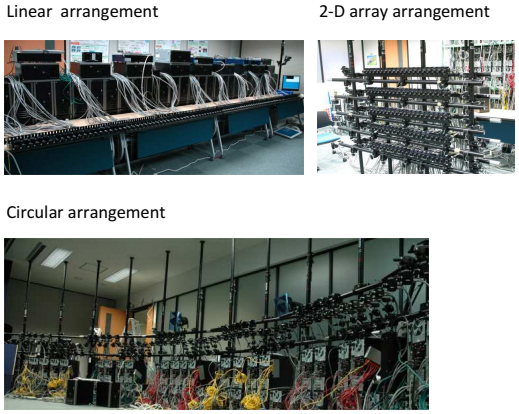


Fig. 4. Flexible arrangement of 100-camera system.

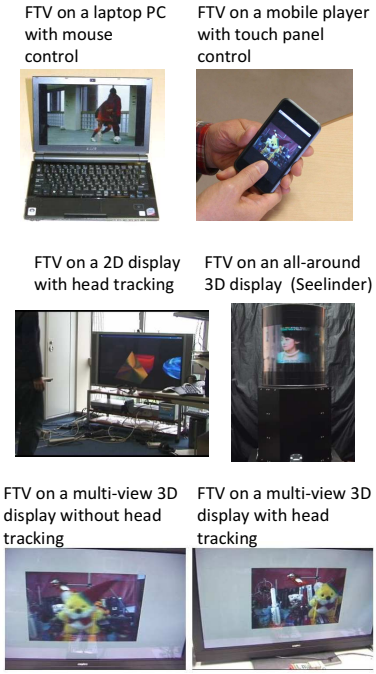


Fig. 5. Various types of FTV user interface.

4. International Standardization of FTV

MPEG has been conducting the international standardization of FTV. The first phase of FTV was MVC (Multi-view Video Coding). MVC was completed in May 2009 and has been adopted by Blu-ray 3D. The second phase of FTV is 3DV (3D Video). 3DV is a standard that targets serving a variety of 3D displays. "Call for Proposals on 3D Video Coding Technology" was issued in March 2011.

5. Conclusions

FTV is cutting the frontier of communications. It transmits whole 3D spatial information and enables the realistic viewing and free navigation of a 3D scene. FTV was adopted as the key concept of 2022 FIFA World Cup bidding to Japan. Japan planed to deliver the replica of a soccer stadium to all over the world by FTV. FTV will find many applications in the wide area such as broadcast, communication, amusement, entertainment, advertising, exhibition, education, medicine and so on.

ACKNOWLEDGMENT

This research was partially supported by Strategic Information and Communications R&D Promotion Programme (SCOPE) of the Ministry of Internal Affairs and Communications, and Grant-in-Aid for Scientific Research (B), 22360151.

References

[1] Masayuki Tanimoto, "Free Viewpoint Television," The Journal of Three Dimensional Images, vol.15, no.3, pp.17-22, September 2001 (in Japanese).

[2] Masayuki Tanimoto, "Overview of Free Viewpoint Television," Signal Processing: Image Communication, vol. 21, no. 6, pp. 454-461, July 2006.

[3] Masayuki Tanimoto, Mehrdad Panahpour Tehrani, Toshiaki Fujii, Tomohiro Yendo, "Free-Viewpoint TV", IEEE Signal Processing Magazine, vol.28, no.1, pp.67-76, January 2011.

[4] M. Sekitoh, T. Fujii, T. Kimoto and M. Tanimoto, "Bird's Eye View System for ITS", IEEE, Intelligent Vehicle Symposium, pp. 119-123, May 2001.

[5] M. Panahpour Tehrani, K. Niwa, N. Fukushima, Y. Hirano, T. Fujii, M. Tanimoto, K. Takeda, K. Mase, A. Ishikawa, S. Sakazawa, A. Koike, "3DAV Integrated System Featuring Free Listening-point and Free Viewpoint Generation", IEEE International Conference on Multimedia signal processing, MMSP, 855-860, Australia, Oct. 2008.

[6] T. Yendo, T. Fujii, M. P. Tehrani and M. Tanimoto, "All-Around Ray-Reproducing 3DTV", IEEE International Workshop on Hot Topics in 3D (Hot 3D), July 2011.

Masayuki Tanimoto received the B.E., M.E., and Dr.E. degrees in electronic engineering from the University of Tokyo in 1970, 1972, and 1976, respectively. He joined Nagoya University in 1976. Since 1991, he has been a Professor at Graduate School of Engineering, Nagoya University. He has been engaged in the research of image coding, image processing, 3D imaging, FTV and ITS.

He is the former president of the Institute of Image Information and Television Engineers (ITE), and a fellow of the Institute of Electronics, Information, and Communication Engineers (IEICE) and ITE. He received the ITE Distinguished Achievement and Contributions Award, the IEICE Achievement Award, and the Commendation for Science and Technology by the Minister of Education, Culture, Sports, Science, and Technology.

Next-generation 3D: From Depth Estimation to the Display

Ramsin Khoshabeh, Can Bal, Ankit Jain, Lam Tran, Stanley Chan, Truong Q. Nguyen
University of California, San Diego, USA
{ramsin, cbal, ankitkj, lat003, h5chan, nguyent}@ucsd.edu

1. Introduction

In recent years, the number of groundbreaking innovations related to traditional display technologies has started to plateau. With frame rates beginning to exceed those visually perceptible by most people and resolutions pushing the limits of network capabilities, display manufacturers have struggled to bring consumers new technologies that merit the need for them to replace their existing models. As such, manufacturers have eagerly tried to adopt new features and modalities into their displays in order to provide a more appealing product, such as internet capabilities and 3D visualization.

This trend was made most evident in 2009 when the movie, *Avatar*, was released in 3D and became the highest grossing movie of all time, earning over $2 billion [1]. Soon after, 3D displays began to go on sale in large quantities, as manufacturers hurried to corner the market. However, the hype behind the 3D revolution masked two crucial problems that had not yet been addressed: 1) the lack of 3D content, and 2) whether consumers would transition to displays that would require them to wear glasses. The second question received an answer fairly quickly when these 3D displays were met with a demand that would at best be classified as mediocre, despite the growing number of companies entering the market. The answer to the first question is still up for debate, as researchers argue whether stereo camera systems will become ubiquitous, 2D-to-3D conversion techniques will pose a viable alternative, or some alternate method will arise.

Currently, some significant research has been going into exploring different technologies that will provide individuals a 3D experience without the need for glasses, so-called autostereoscopic displays. Early prototypes have shown promise and are consistently advancing in quality [2, 3]. However, even with such displays, simply capturing video using a stereo camera system is often insufficient. Every display has different specifications, such as the number of views, the amount of depth they may comfortably show, or even how the data should be formatted.

Due to this fact, two key steps must generally accompany the 3D content generation process: stereo correspondence (or disparity estimation) and virtual view synthesis. In our work, we have explored a number of ways to accomplish these tasks. Our methods can be used to generate 3D content for arbitrary display technologies with the realization that generalization is crucial for the future of 3D media.

2. Stereo Correspondence

Stereo depth estimation is an integral problem associated with the delivery of 3D content. Depth is an important element for an accurate visual representation of 3D content. Motivated by the human visual system, the problem is formulated as the determination of object distance in a scene based on stereo information. Humans see depth by integrating multiple visual cues and processing that information in their visual cortex. By far the most well studied cues come from having two eyes because they intrinsically correlate with depth. A simple experiment to validate the importance of binocular cues is to close one eye and try to grab something in front of you. Stereo depth estimation builds off this observation. Monocular cues (e.g., occlusion, motion, texture, relative size) may aid in the estimation, but stereo cues (e.g., horizontal parallax) are the primary and most robust indicators of depth.

In a two-camera imaging system, disparity is defined as the vector difference between the imaged object point in each image relative to the focal point [4]. It is this disparity that allows for depth estimation of objects in the scene via triangulation of the object point. In rectified stereo, where both camera images are in the same plane, only horizontal disparity exists. In this case, multiview geometry shows that disparity is inversely proportional to the actual depth in the scene. Thus, if disparity can be measured from a rectified stereo image pair, then the relative depth of each object can also be calculated. Fig. 1 shows the relationship between depth and disparity when the cameras are rectified and camera centers are known. The

inverse relationship between depth z and disparity d is identified as: $d = fT/z$ where T is the camera separation (the interocular distance) and f is the camera focal length.

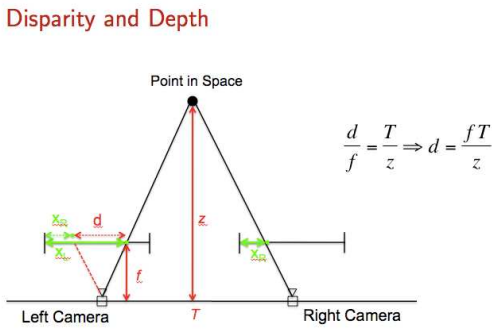


Figure 1. Relating depth to disparity.

Fortunately, excellent methods exist to estimate a disparity map, an image whose locations and intensities correspond to disparity magnitude at a given pixel location. The drawback is that many of these algorithms generally perform poorly on real-world images because they are trained on specific datasets, such as the Middlebury database [5].

For static images, the matching techniques can be generally categorized into local or global methods. Local methods use information within a small window to estimate the disparity for a given pixel. Global approaches, in contrast, incorporate assumptions about depth discontinuities into a global energy function, which is solved through energy minimization techniques such as graph cuts [6] or hierarchical belief propagation (HBP) [7].

Stereo video sequences have been studied less extensively than images. Simply applying the many image-based techniques to each frame of a video produces temporal inconsistencies for the disparity estimates. When these depth estimates are used for view synthesis in autostereoscopic displays, results are poor and contain a great deal of flicker that significantly detracts from the 3D effect. To incorporate time information, some have proposed restructuring the optimization problem by, for example, using a 3-dimensional Markov random field, as in [8], so that temporal information is explicitly modeled. Such methods suffer from the added computational complexity of having to minimize an energy function in a very large space. Other approaches attempt to alleviate the spatio-temporal burden by considering variants of optical flow [9] or median filtering [10], but flow field computation introduces unnecessary errors into the framework and for a robust estimation requires significant computational time as well.

Our approach [11] leverages the advances made with the image-based techniques. The core of the disparity estimation algorithm is fast, robust inferencing using hierarchical belief propagation. HBP maintains the accuracy of global methods, such as traditional belief propagation or graph cuts, but rivals local methods in computational time.

The difficulty with operating on each frame of a video sequence independently is that the consistency between consecutive frames is lost. The noisy estimates for each frame create a flickering effect over time that is highly bothersome to the human visual system. To compensate for the temporal inconsistency, we present a novel, fast, and efficient method. After we compute disparity estimates for each frame individually, we enforce the disparity smoothness assumption between neighboring pixels in space-time by mandating that values should vary smoothly except at object boundaries. With this assumption, we can treat the temporal and spatial inconsistencies as noise and apply Total Variation (TV) minimization, where an unknown signal must be recovered from a noisy observation. Our method is faster than state-of-the-art methods and produces superb results.

2. Virtual View Synthesis

As discussed earlier, the 3D effect is an illusion created as a result of processing a number of visual cues in the visual cortex. The biggest contribution to creating and manipulating depth perception comes from stereo correspondences. Thus by controlling the relative position of the stereo cameras with respect to each other, it is possible to control the amount of depth perception in a scene.

To meet the specifications of different 3D displays, it is necessary to generate content specific to the displays. This includes controlling the number of cameras to be used to capture the scene, and their relative positions with respect to each other. As this is highly cumbersome, it is not desired for content generation. On the other hand, given a stereo pair and the corresponding disparity information, it is possible to render virtual views as if they are captured with virtual

cameras, using the conventional "Depth Image Based Rendering (DIBR)" technique [12] as illustrated in Fig. 2. In the most basic sense, this method selects the objects at the same depth level from the captured views and blends them onto the correct position in the virtual view.

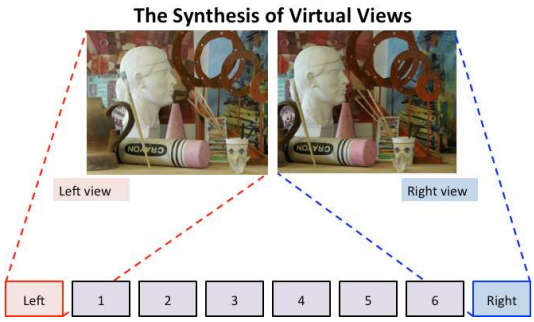


Figure 2. Virtual View Synthesis.

To provide enjoyable depth perception on any display, it is crucial to have the capability of adjusting the depth content from very little depth (for mobile applications) to significant depth (for cinema displays). In order to support this, the best setting for capturing the stereo data is when the cameras are separated by a large baseline. However, for synthesizing the content for a virtual view, a large baseline introduces the occlusion problem, where certain regions that should be visible in the new view do not exist in any of the captured views. This is where most DIBR based view synthesis algorithms separate from each other. So far a significant number of different approaches have been reported in the literature. There exist methods that range between very advanced but slow and very fast but simplistic.

Some of the methods use a circular camera setup where almost all of the occluded areas in one of the captured views appear in another. This type of camera setup allows a layered approach where a background layer can be extracted from the given stereo images and used to fill in missing regions [13, 14]. This approach is especially useful when the virtual cameras are to be positioned arbitrarily in 3D space, as in the case of free viewpoint TV (FTV). Yet, for the case of content generation for autostereoscopic screens or depth content adjustment for stereo systems, the virtual cameras need to be positioned along a line or a wide arc. Thus, these capture systems are not preferable as they bring in unnecessary complexity both to depth estimation and view synthesis.

For rectified camera setups, there are view synthesis methods that process the missing regions pixel-by-pixel and also methods which use patch based inpainting. The pixel based methods tend to use interpolation or simple inpainting methods but they usually suffer from blur in the synthesized region [15]. Some other pixel based methods fill in the missing regions using more advanced methods and optimization techniques [16, 17]. These approaches yield better results but patch-based methods are still usually superior [18, 19]. Some algorithms also take it a step further and provide techniques for achieving spatial, temporal [20] and inter-view consistency [18].

Currently another trend in technology is GPGPU (General Purpose GPU) programming as it can provide significant performance gains over traditional CPU programming, when the algorithms are highly parallelizable. Since a significant portion of virtual view synthesis operations are as such, there are also some reported methods that take advantage of the processing power of the current GPUs to provide very fast view synthesis [21, 22]. Fast view synthesis is highly desirable since it can enable real-time interaction with the user. Unfortunately, these methods provide fast view synthesis at the expense of limiting themselves to simple approaches, and the reported performances are usually below the more advanced methods.

4. Conclusions

Stereoscopic data will soon be ubiquitous. The trend toward 3D has already raised many new concerns as to how the immersive experience may best be brought to viewers. Significant headway is being made to develop display systems that will emulate how people see 3D in the natural world.

Along with these new technologies, 3D video processing algorithms will be central in making this transition take place. Of particular importance are the stereo correspondence and virtual view synthesis algorithms related to the generation of 3D content. As these methods mature, they will expand the horizons of what is possible in the world of 3D, ultimately leading to new innovation. While it is still yet unclear what the future of display technologies will hold, it is certain that 2D media will no longer suffice. 3D is here to stay.

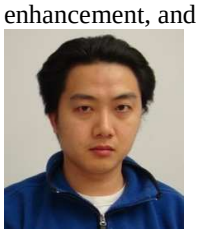
References
[1] http://www.imdb.com/title/tt0499549/
[2] http://alioscopy.com/
[3] http://www.nintendo.com/3ds
[4] P. An, Z. Zhang, and L. Shi, “Theory and experiment analysis of disparity for stereoscopic image pairs,” in Intelligent Multimedia, Video and Speech Processing, 2001.
[5] H. Hirschmuller and D. Scharstein, “Evaluation of cost functions for stereo matching,” in IEEE CVPR, 2007.
[6] Y. Boykov, O. Veksler, and R. Zabih, “Fast Approximate Energy Minimization via Graph Cuts,” PAMI, vol. 23, pp. 1222–1239, February 2004.
[7] P. Felzenszwalb and D. Huttenlocher, “Efficient Belief Propagation for Early Vision,” in CVPR, pp. 261–268, 2004.
[8] O. Williams, M. Isard, and J. MacCormick, “Estimating Disparity and Occlusions in Stereo Video Sequences,” in CVPR, 2005.
[9] F. Huguet and F. Devernay, “A Variational Method for Scene Flow Estimation from Stereo Sequences,” in ICCV, 2007.
[10] M. Bleyer and M. Gelautz, “Temporally Consistent Disparity Maps from Uncalibrated Stereo Videos,” in ISPA, 2009.
[11] R. Khoshabeh, S. Chan, and T. Nguyen, “Spatio-temporal Consistency in Video Disparity Estimation,” in ICASSP, 2011.
[12] C. Fehn, “Depth-image-based rendering (DIBR), Compression and Transmission for a New Approach on 3D-TV," Proceedings of SPIE Stereoscopic Displays and Virtual Reality Systems XI, vol. 5291, pp. 93-104, 2004.
[13] C. L. Zitnick, S. B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski, “High-quality video view interpolation using a layered representation," Proc. ACM SIGGRAPH, Aug 2004.
[14] K. Muller, A. Smolic, and K. Dix, “View synthesis for advanced 3D video systems," EURASIP Journal on Image and Video Processing, 2008.
[15] S. Zinger and L. Do, “Free-viewpoint depth image based rendering," Journal Visual Communication and Image Representation, vol. 21, pp. 533-541, 2010.
[16] A. Jain, L. Tran, R. Khoshabeh, and T. Nguyen, “Efficient Stereo-to-Multiview Synthesis," in ICASSP 2011.
[17] L. Tran, C. Pal, and T. Nguyen, “View synthesis based on conditional random fields and graph cuts," in ICIP, 2010.
[18] L. Tran, R. Khoshabeh, A. Jain, C. Pal, and T. Nguyen, “Spatially Consistent View Synthesis with Coordinate Alignment," in ICASSP, 2011.
[19] P. Ndjiki-Nya, M. Koppel, D. Doshkov, H. Lakshman, P. Merkle, K. Muller, and T. Wiegand, “Depth image based rendering with advanced texture synthesis," in ICME 2010, pp. 424-429.
[20] C.M. Cheng, X.A. Hsu, and S.H. Lai, “Spatio-temporally Consistent Multi-view Video Synthesis for Autostereoscopic Displays," in PCM, pp. 532-542, 2010.
[21] J. Lu, S. Rogmans, G. Lafruit, and F. Catthoor, “High-Speed Stream-Centric Dense Stereo and View Synthesis on Graphics Hardware," MMSP 2007.
IEEE 9th Workshop on, pp. 243{246, 2007.
[22] S. Rogmans, J. Lu, P. Bekaert, and G. Lafruit, “Real-time stereo-based view synthesis algorithms: A unified framework and evaluation on commodity GPUs," Signal Processing: Image Communication, vol. 24, pp. 49-64, 2009.

Ramsin Khoshabeh is a Ph.D. candidate at the University of California, San Diego in the department of Electrical and Computer engineering. He graduated with a B.S. in Electrical Engineering from the same school. As an undergrad, he specialized in computer design and graduated with highest honors (summa cum laude). In graduate school his emphasis has been on computer vision and 3D. His thesis work centers on the application of 3D in the realm of surgical practice. In 2009, he was awarded the Calit2 Strategic Research Opportunities (CSRO) Fellowship.

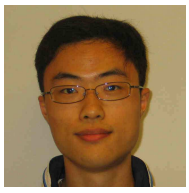
Can Bal is a Ph.D. student at the University of California, San Diego in the department of Electrical and Computer engineering. He got his B.S. and M.S. degrees in Electrical Engineering from Bilkent University, Turkey in 2007 and 2009 respectively. His research interests include

video compression and computer vision applications, focusing on 3D video.

Ankit K. Jain is a Ph.D. student in the Video Processing Lab at the University of California, San Diego. He obtained a B.S. in Electrical Engineering specializing in signal processing from Stanford University in 2005, and worked at MIT Lincoln Laboratory for three years prior to joining UCSD. His graduate research is on the psychovisual bases for artifacts seen in stereo video, as well as 3D video processing, enhancement, and compression.



Lam Tran graduated with a B.S in Computer Science from Rensselaer Polytechnic Institue and B.S in Statistics and Applied Mathameatics from University of Rochester in 2007 and 2009 respectively. His research interests include computational photography, computer vision, and statistical machine learning. Lam was awarded the Powell Fellowship in 2009 and NSF Graduate Research Fellowship is 2011.



Stanley H. Chan is currently a Ph.D. candidate at the University of California, San Diego in the department of Electrical and Computer engineering. He received M.A. degree in applied mathematics from UCSD in June 2009, and B.Eng degree with first class honor in electrical engineering from the University of Hong Kong in June 2007. His research interests are large-scale numerical optimization algorithms with applications to video processing. Mr. Chan is a recipient of Croucher Foundation Scholarship.

Truong Q. Nguyen [F'05] is currently a Professor at the ECE Dept., UCSD. His current research interests are 3D video processing and communications and their efficient implementation. He is the coauthor (with Prof. Gilbert Strang) of a popular textbook, Wavelets & Filter Banks, Wellesley-Cambridge Press, 1997, and the author of several matlab-based toolboxes on image compression, electrocardiogram compression and filter bank design. He has over 300 publications.

Prof. Nguyen received the IEEE Transaction in Signal Processing Paper Award (Image and Multidimensional Processing area) for the paper he co-wrote with Prof. P. P. Vaidyanathan on linear-phase perfect-reconstruction filter banks (1992). He received the NSF Career Award in 1995 and is currently the Series Editor (Digital Signal Processing) for Academic Press. He served as Associate Editor for the IEEE Transaction on Signal Processing 1994-96, for the Signal Processing Letters 2001-2003, for the IEEE Transaction on Circuits & Systems from 1996-97, 2001-2004, and for the IEEE Transaction on Image Processing from 2004-2005.

Realistic Virtual Try-On of Clothes using Real-Time Augmented Reality Methods

Peter Eisert and Anna Hilsmann, Humboldt University Berlin, Germany
{peter.eisert, anna.hilsmann}@hhi.fraunhofer.de

1. Introduction

In the last years, the increase in computational power of regular PCs and smart phones has lead to many new and also commercially successful augmented reality applications. Besides the augmentation of different kinds of information on the display of geo-located devices, the visualization of a user's appearance change caused by particular products has become an important field of research. In such a scenario, a display showing a video or picture of the user can replace a real mirror with the possibility to combine the real scene with graphical models to modify appearance. Possible examples of such virtual mirrors are virtual try-ons for clothes, shoes, watches or jewelry, glasses, or even change of hairstyle or makeup [1,2,3].

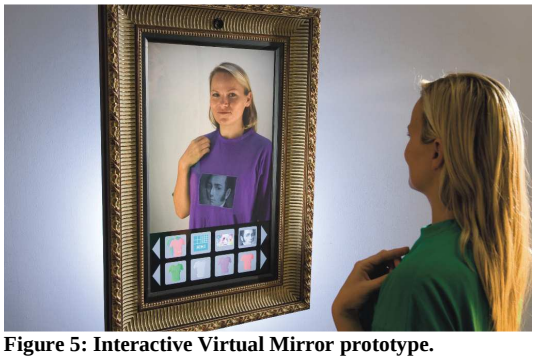


Figure 5: Interactive Virtual Mirror prototype.

In this paper, we focus on interactive virtual try-on of clothes for virtual dressing rooms. A user is tracked by one or more cameras and is visualized on displays showing him/her augmented with new clothes that can be changed and configured, e.g. by touchscreens. The main technological challenges are the accurate tracking of the deformable surface of the user's cloth / body pose and the realistic rendering of new clothes in the correct poses. For the rendering, most solutions are based on textured 3D computer graphics models while foldings and dynamics of the cloth are synthesized. Although realistic results can be achieved [4], a natural simulation with all dynamics and detail is still computational demanding. We therefore follow a somewhat different approach by exploiting as much information from the original video as possible. Rather than rendering a complete synthetic model on top of the video, we segment the piece of clothing from the video and estimate surface deformation and shading, originating from wrinkles in the fabric. A new appearance is then created by retexturing the original video. Given a new pattern and color of the fabric by means of an image, it is deformed according to the estimated motion and shaded by modulation with the shading map. The decomposition of the video and retexturing is illustrated in Figure 6. Since both types of information are recovered from the original video, highly natural results are achieved while preserving relatively low complexity enabling real-time interactive applications. In the next section, we describe our virtual mirror prototype (Figure 5) that allows the change of the color and logo of a t-shirt. Current work on the natural replacement of an entire piece of clothing is illustrated in section 4.

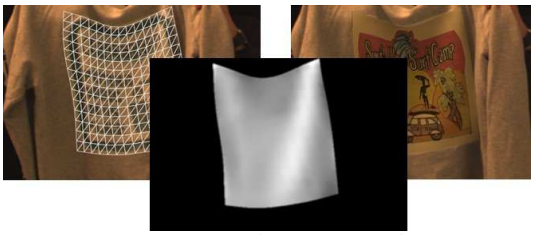


Figure 6: Decomposition of video into deformation and shading map. Retextured results by applying deformation and shading to a new image (right).

2. Virtual Try-On of Clothes

Our virtual mirror prototype for real-time visualization of garments is based on a dynamic texture overlay method for monocular video sequences. Similar to looking into a mirror when trying on clothes, we create the same impression but for virtually textured garments. The mirror is replaced by a large display that shows the mirrored image of a camera capturing e.g. the upper body part of a person (see Figure 5). We segment the piece of clothing of interest and estimate elastic surface deformations and illumination changes in an image-based optimization scheme [5,6]. The feature-less approach matches the entire image region with a reference frame. This optimization scheme starts from a relaxed brightness constancy equation and formulates a pixel-wise error at pixel $\mathbf{x}_i$ as

$$r_i = \Psi_p(\mathbf{x}_i, \boldsymbol{\theta}) \cdot I_{n-1}\left(\Psi_g(\mathbf{x}_i, \boldsymbol{\theta})\right) - I_n(\mathbf{x}_i),$$

where $\Psi_g(\mathbf{x},\boldsymbol{\theta})$ and $\Psi_p(\mathbf{x},\boldsymbol{\theta})$ are geometric and photometric mesh-based warp functions parameterized by vector $\boldsymbol{\theta}$. A cost function based on

the pixel-wise errors and a smoothness term based on the mesh Laplacians is minimized in a robust Gauss-Newton optimization scheme.

Having estimated the deformation and illumination changes, an arbitrary virtual texture can be realistically augmented onto the moving garment such that the person seems to wear the virtual clothing. The result is a combination of the real video and the new augmented model yielding a realistic impression of the virtual piece of cloth (see Figure 7).

Figure 7: Upper row: input camera images. Lower row: retextured results with modified color and logo.

3. Social and Remote Interaction

Besides visualization of the new appearance, a virtual try-on system can offer many novel features for interaction. For example, the usual suggestion of new combinations based on the statistical experience of other users can be enhanced. For that purpose, we analyze the colors and patterns of the user's clothing and can make similar suggestions or hints based on input from users with similar taste. Also, the dressing room's output of a new look can be shared with friends by uploading images on social networks or a life video stream can be sent to remote participants who can even make suggestions via a feedback channel. The usually demanding encoding complexity can here be reduced by exploiting knowledge from the tracking and the graphical objects in the scene [7].

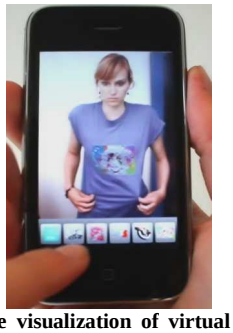


Figure 8: Mobile visualization of virtual mirror output with possibility of remote configuration of clothing.

4. Image-based Representation of Clothes

Current and future work focuses on new image-based representations of clothes that allow realistic real-time rendering of entire pieces of clothing depending only on the pose of the user. Rendering a pure 3-dimensional model of clothes that captures all fine wrinkles is very time-consuming and thus not feasible for a real-time application. However, details at fine wrinkles are very important for the realistic appearance of the clothes. We therefore synthesize new views by interpolating from a large database of previously captured images showing the garment in different poses. The key idea is to model the rough shape with a geometric model accounting for dominant motion / body pose changes and small details like wrinkles, seams, and patterns by a combination of real images to get a perceptually correct model. Hence, our representation combines body-pose-dependent geometry with image-based information consisting of different scale levels (see Figure 9). The rough shape model of fixed topology can be animated to show different poses, surface deformers add rough 3D surface deformation, while fine details are represented by the appearance. The appearance includes alpha maps for a pixel-accurate silhouette, shading for representation of small wrinkles, and texture orientation for retexturing. The information is captured offline in a preprocessing step.

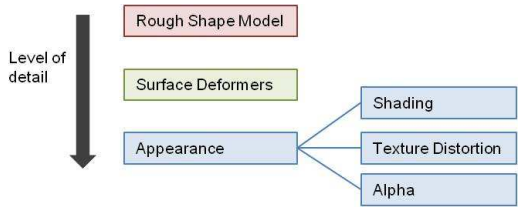


Figure 9: Hierarchical image-based cloth representation.

For rendering, the user's body pose is estimated and drives the interpolation of a new model by means of shape and appearance from samples in the database. Since most computation is performed a-priori for database creation, followed by moderately simple interpolation and rendering, realistic visual output can be achieved in real-time. Due to the separation of the surface color in albedo (i.e. its underlying local color) and shading with associated texture orientation, retexturing of the appearance becomes possible. Given a new fabric pattern, surface texture can be modified by combining the texture distortion and shading information with the new pattern as depicted in Figure 10.

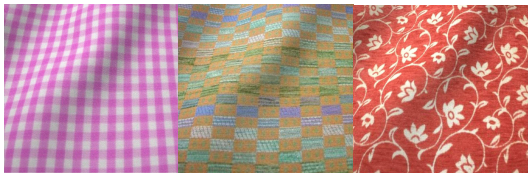


Figure 10: Retexturing results achieved by separating surface albedo from texture distortion and shading information. Left: original image, center and right: retexturing results.

5. Conclusions

Virtual try-on of clothes is an augmented reality application that is going to influence virtual shopping in the next years. We have shown a new method for virtual clothing, which allows for retexturing of a real video of a person using estimates from surface deformation and shading, rather than synthesizing clothes by 3D geometric models. The method can be extended to deal with entire pieces of clothing. Rendering is performed by image interpolation from a large database yielding highly natural results at reasonable complexity enabling high quality real-time virtual try-on.

References

[1] M. Wacker, M. Keckeisen, S. Kimmerle, W. Strasser, V. Luckas, C. Groß, A. Fuhrmann, M. Sattler, R. Sarlette, R. Klein: "Virtual try-on," Informatik Spektrum, pp. 504–511, Dec. 2004.

[2] P. Eisert, P. Fechteler, J. Rurainsky, 3-D Tracking of Shoes for Virtual Mirror Applications, Proc. Computer Vision and Pattern Recognition CVPR, Anchorage, Alaska, June 2008.

[3] N. Magnenat-Thalmann, P. Volino, B. Kevelham, M. Kasap, Q. Tran, M. Arevalo, G. Priya, N. Cadi, An Interactive Virtual Try On. Proceedings of IEEE Virtual Reality Conference (VR), pp. 263-264, March 2011.

[4] S. Pabst, S. Krzywinski, A. Schenk, B. Thomaszewski: "Seams and Bending in Cloth Simulation", Proc. Workshop in Virtual Reality Interactions and Physical Simulation, 2008.

[5] A. Hilsmann, D. Schneider, P. Eisert, Realistic Cloth Augmentation in Single View Video under Occlusions, Computers & Graphics, vol. 34, no. 5, pp. 567-574, June 2010.

[6] A. Hilsmann, P. Eisert, Joint Estimation of Deformable Motion and Photometric Parameters in Single View Video, Proc. ICCV NORDIA Workshop, Kyoto, Japan, pp. 390-397, Sep. 2009.

[7] P. Fechteler, P. Eisert, Accelerated Video Encoding using Render Context Information, Proc. Int. Conf. on Image Processing (ICIP), Hong Kong, pp. 2033-2036, Sep. 2010.

Peter Eisert is Professor for Visual Computing at the Humboldt University Berlin and heading the Computer Vision & Graphics Group in the Image Processing Department of the Fraunhofer HHI, Berlin, Germany. He received the Dipl.-Ing. degree in Electrical Engineering from the Technical University of Karlsruhe and the Dr.-Ing. degree from the University of Erlangen in 2000. In 2001, he worked as a postdoctoral fellow at the Stanford University on 3D image analysis and synthesis as well as facial animation and computer graphics. In 2002, he joined the Image Processing Department at HHI, where he is involved in numerous national and international research projects. He has published more than 100 conference and journal papers on vision, graphics, and communication. In 2002, he received the SPIE VCIP Young Investigator Award and in 2008 the best paper award of the CVPR Nordia Workshop. Since 2006, he is Associate Editor of the International Journal of Image and Video Processing. His research interests include 3D image analysis and synthesis, face processing, image-based rendering, computer vision, computer graphics, as well as image and video processing.

Anna Hilsmann received the Dipl.-Ing. degree in Electrical Engineering and Information Technology from RWTH Aachen University in 2006. In 2007, she joined the Computer Vision & Graphics Group at the Image Processing Department of Fraunhofer HHI in Berlin, Germany. Since 2009 she is also PhD student in Computer Science at Humboldt University Berlin. Her research focuses on realistic visualization of clothes for augmented reality applications. This includes image-based deformation and illumination estimation methods for deformable surfaces as well as dynamic texture overlay methods. These methods are the basis of a new real-time Virtual Clothing system that has been awarded by the German Ministry of Business with the "365 Orte" award in 2009. In 2008

she received the best paper award at the CVPR NORDIA workshop the best poster award at Eurographics in 2011. In the same year she was awarded with the Google Anita Borg Memorial Scholarship.

E-Letter Editorial Board

MMTC Officers