Video Quality Measurement Standards – Current Status and Trends

Stefan Winkler
Symmetricom
San Jose, CA 95131, USA

***Abstract*—With the wide-spread use of digital video, quality considerations have become essential, and industry demand for video quality measurement standards is rising. Various organizations are working on such standards. This paper reviews some of the existing standards documents, on-going work and future trends in this space.**

***Index Terms*—Video quality measurement, subjective experiments, objective quality metrics, Mean Opinion Score (MOS), standardization**

I. Introduction

In the field of signal processing and telecommunication, speech quality measurement has quite a long history. More recently, quality assessment has been extended to audio and video as well [1]. The industry's need for accurate and reliable objective video metrics has become more pressing with new digital video applications and services such as mobile broadcasting, Internet video, and IPTV. Quality measurement has a wide range of uses, including equipment testing (e.g., codec evaluation), transmission planning and network-dimensioning tasks, head-end quality assurance, in-service network monitoring, and client-based quality measurement.

Standards address a growing number of issues related to video quality measurement. These include definitions and terms of reference, requirements, recommended practices, test plans, and many more. In this paper, I will focus on work specifically related to defining and evaluating quality measurement algorithms, also known as objective quality assessment. In this regard, we can distinguish loosely between Quality of Service (QoS) and Quality of Experience (QoE). Traditional QoS, which is relatively well understood and established, focuses on network performance and data transmission. QoE on the other hand is still an active area of research and standards work and thus less well-defined; the term is meant to describe quality from the perspective of the user or consumer (i.e. viewer), with a focus on perceived quality of the content (or more comprehensively, user experience).

The objective of this paper is to provide some guidance with respect to which standards organization does what, as well as to highlight what has already been accomplished and what is still on the roadmap. The paper is organized as follows. Section II discusses some terminology of quality metrics. Section III briefly outlines common standards for subjective quality assessment. Section IV reviews existing standards on objective quality assessment as well as standards currently under development. Section V provides some concluding thoughts on current activities and trends.

II. Terminology

Objective quality metrics are algorithms designed to characterize the quality of video and predict viewer opinion. Different types of objective metrics exist [2]; the following is an attempt at classification, as illustrated in Figure 1 [3]:

- *Data metrics*, which measure the fidelity of the signal without considering its content (not shown in the Figure). In the image processing community, mean squared error (MSE) or peak signal-to-noise ratio (PSNR) are being used as quasi-standard fidelity metrics [4]. The network QoS community has equally simple metrics to quantify transmission errors, such as bit error rate (BER) or packet loss rate (PLR). None of them take into account the content, i.e. the meaning and thus the visual importance of the pixels or packets concerned.
- *Picture metrics*, which treat the video data as the visual information that it contains. They specifically account for the effects of distortions and content on perceived quality, either based on direct models of the human vision system and its components, or based on the extraction and analysis of certain features or artifacts in the video.
- *Packet-* or *bitstream-based metrics* for compressed video delivery over packet networks, which look at the packet header information and the encoded bitstream directly without fully decoding the video. This has the added advantage of much lower bandwidth of data to be analyzed compared to metrics looking at the fully decoded video, resulting in much lower processing requirements. Using such metrics, it is thus possible to measure the quality of many video streams/channels in parallel. On the other hand, these metrics are necessarily specific to selected codecs and network protocols.
- *Hybrid metrics*, which use a combination of packet information, bitstream or even decoded video as input.

Additionally, metrics can be classified into full-reference, no-reference and reduced-reference metrics based on the amount of reference information they require [5]:

- *Full-reference (FR) metrics* measure the degradation in a test video with respect to a reference video. They require the entire reference video to be available, usually in unimpaired and uncompressed form, and generally impose precise spatial and temporal alignment as well as calibration of luminance and color between the two videos, so that every pixel in every frame can be matched with its counterpart in the other clip for direct comparison.

978–1–4244–4657–5/09/$25.00

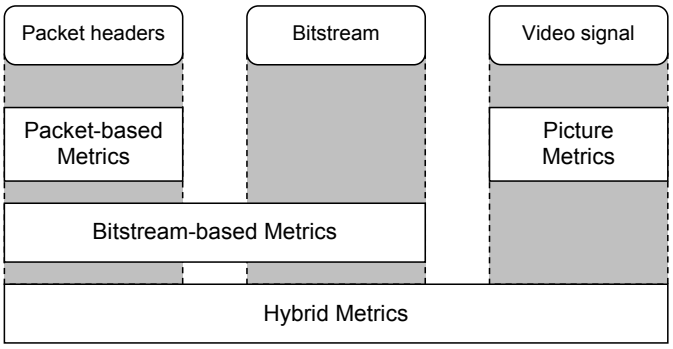

Fig. 1. Classification of packet-based, bitstream-based, picture and hybrid metrics [3], adapted from ITU-T.

- *No-reference (NR) metrics* analyze only the test video, without the need for an explicit reference clip, which makes them completely free from alignment issues. Their main challenge lies in telling apart distortions from content – NR metrics have to make assumptions about the video content and the types of distortions.
- *Reduced-reference (RR) metrics* offer a compromise between FR and NR metrics in terms of reference information. They extract a number of features from the reference and/or test video, and the comparison of the two clips is then based only on those features. This approach makes it possible to avoid some of the assumptions of pure no-reference metrics while keeping the amount of reference information manageable.

These three classes of metrics also have different operational uses. FR metrics are most suitable for offline video quality measurement such as codec tuning or lab testing, whereas NR and RR metrics are better suited for monitoring of in-service video at different points in the system. RR metrics still require a back-channel and access to the reference at some point.

III. Subjective Quality Assessment

Procedures and standards for subjective assessment of speech, audio and video have been around for many years. Due to the proliferation of digital audio and video content, conducting subjective experiments to measure its quality has become relatively commonplace.

Subjective testing for visual quality assessment has been formalized in ITU-R Rec. BT.500 [6] and ITU-T Rec. P.910 [7], which suggest standard viewing conditions, criteria for the selection of observers and test material, assessment procedures, and data analysis methods. The former has a longer history and was written with television in mind, whereas the latter is intended for multimedia applications. Naturally, the experimental setup and viewing conditions differ in the two recommendations, but the procedures from both should be considered for any experiment.

These recommendations define some of the most commonly used procedures for subjective quality assessment. Some examples include:

- Double Stimulus Continuous Quality Scale (DSCQS), where subjects rate short sequence pairs, consisting of a test and corresponding reference video.
- Double Stimulus Impairment Scale (DSIS), also referred to as Degradation Category Rating (DCR), where subjects rate the amount of impairment in the test video with respect to the known reference video.
- Single Stimulus Continuous Quality Evaluation (SSCQE), where subjects watch a program of typically 20-30 minutes duration and continuously rate the instantaneously perceived quality on a slide.
- Absolute Category Rating (ACR), a single-stimulus method, where subjects rate each test video individually without comparison to an explicit reference.
- Pair Comparison (PC), where test videos from the same scene but different conditions are paired in many possible combinations, and subjects make a preference judgment for each pair.

The testing procedures mentioned above generally have different applications [1]. There are also a variety of different rating scales, continuous and discrete, numerical or categorical, from 5 levels to 100 – see [8] for an analysis and more details.

The outcome of any subjective experiment are quality ratings from viewers, which are then averaged for each test clip into Mean Opinion Scores (MOS).

IV. Objective Quality Assessment

The purpose of objective quality measurement standards is three-fold:

- Defining the meaning of MOS for a given application (e.g. people know what a MOS of 4.1 means in terms of video quality).
- Defining a method for MOS prediction that is reliable, i.e. the MOS estimations given by the tool should be closely related to viewer experience.
- Defining a method for MOS prediction that is reproducible, i.e. two people using the tool for the same test clips should obtain the same results.

Existing standards have achieved some, but not all of these objectives.

Various standards bodies, industry forums and other groups are working on video quality assessment in one form or another. In this section, I will review the most active ones as far as objective quality measurement algorithms and their evaluation are concerned.

A. Video Quality Experts Group (VQEG)

The Video Quality Experts Group (VQEG)[1] was founded in 1997 by a group of ITU-T and ITU-R study group members. The group is composed of experts in the field of video quality assessment from industry and academia. The general goal of VQEG is to advance the field of video quality assessment by evaluating objective quality metrics and investigating new subjective assessment methods [9].

[1] See http://www.vqeg.org/.

VQEG brings together algorithm developers and users to plan and execute validation tests of objective perceptual quality metrics, with the help of independent labs. VQEG's approach to validation testing includes creating video databases and conducting subjective experiments. The test videos are largely unknown to the model developers. The subjective ratings are then to be predicted by the objective quality metrics under consideration. Model evaluation is based on prediction performance according to a number of statistical criteria.

An approximate timeline of past and present VQEG projects is shown in Figure 2.

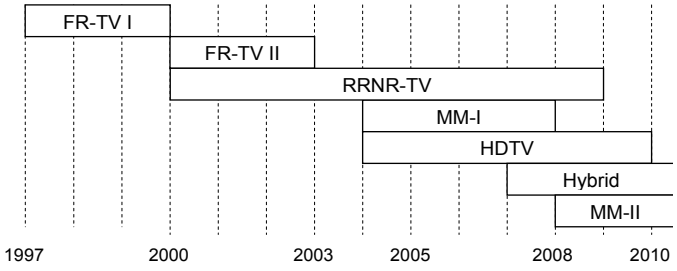

Fig. 2. Approximate timeline of VQEG projects.

In the first test (FR-TV Phase I), which was completed in 2000, the emphasis was out-of-service testing (i.e. full-reference metrics) for production- and distribution-class standard-definition (SD) video. The test conditions comprised mainly MPEG-2 encoded sequences with different profiles, levels and other parameter variations, including encoder concatenation, conversions between analog and digital video, and transmission errors. A set of 8-second scenes with different characteristics were selected by independent labs; the scenes were disclosed to the proponents only after the submission of their metrics. In total, 20 scenes were encoded for 16 test conditions each. Subjective ratings for these sequences were collected in large-scale experiments using the DSCQS method. Unfortunately, the results of this first phase were inconclusive; the performance of most models as well as PSNR were statistically equivalent, so that no algorithm could be recommended. Nonetheless, this phase has provided the research community with a valuable, publicly available, annotated database of test clips, still the only one of its kind today (see also Section V). The findings are described in detail in the final report [10] and in a paper by Rohaly et al. [11].

As a follow-up to this first phase, VQEG carried out a second round of tests for full-reference metrics ("FR-TV Phase II") for SD TV applications, which was completed in 2003. In order to obtain more discriminating results, this second phase was designed with a stronger focus on secondary distribution of digitally encoded television quality video and a wider range of distortions. New source sequences and test conditions were defined, and a total of 128 test sequences were produced. Subjective ratings for these sequences were again collected using the DSCQS method. The best metrics in the test achieved correlations as high as 94% with MOS, thus significantly outperforming PSNR, which had a correlation of about 70%. The results of this VQEG test [12] were the basis for ITU-T Rec. J.144 [13] and ITU-R Rec. BT.1683 [14], which comprise the four top-performing algorithms from this test.

Unfortunately, both phases of tests dealt only with full-reference metrics and focused on MPEG-2 compression for digital TV broadcast, and neither included IP impairments.

Last year, VQEG completed an evaluation of metrics in a "multimedia" scenario (MM Phase I), which is targeted at lower bitrates and smaller frame sizes (QCIF, CIF, VGA) as well as a wider range of codecs and transmission conditions [15]. This comprises video services delivered at 4 Mb/s or less, with a focus on broadband Internet and mobile video streaming. The MM Phase I set of tests was used to validate full-reference, reduced-reference, and no-reference objective models. 13 VGA, 14 CIF, and 14 QCIF subjective tests with 168 test sequences each were created. Subjective video quality was assessed using the ACR method. Based on this test, two new standards for multimedia quality assessment were published, namely ITU-T Rec. J.247 [16], which defines four FR models, and ITU-T Rec. J.246 [17], which defines three RR models. NR models did not achieve satisfactory performance in this test.

The most recent test conducted by VQEG was the reduced-reference and no-reference test for standard-definition television (RRNR-TV). Each experiment included 12 source clips and 34 test conditions for a total of 156 test sequences. MPEG-2 and H.264 codecs were used, together with IP transmission errors. Subjective video quality was assessed using the ACR method. The RRNR-TV final report, which was completed recently [18], describes the performance of seven RR models, some of which may become part of a new ITU recommendation.

VQEG's current project is an evaluation of models for high-definition television (HDTV). The test will comprise full-reference, reduced-reference, and no-reference objective models. Currently, the test video database is being compiled, and the subjective tests are being prepared. Model submission is scheduled for September 2009, and first results should be available by year-end. It is also planned to release a subset of the annotated test sequences after test completion.

Thus far, VQEG has examined only models that consider only decoded video frames, as seen by the viewer. The next VQEG project, termed "Hybrid", will evaluate objective models capable of using packet and bitstream information together with decoded video data. VQEG is currently working on the test plan for this project. This activity is related to some projects of ITU-T Study Group 12 (see below).

Finally, VQEG is planning another multimedia test, MM Phase II, which will be aimed at models that can predict audiovisual quality.

VQEG has also begun a joint effort to develop objective quality assessment models that combine the best parts of existing models. This effort may lead to the establishment of a reference implementation of an objective metric.

B. ITU-T

ITU-T Study Group 9 issues many of the recommendations based on the results and reports compiled by VQEG. Additionally, it develops standards for subjective quality assessment, as well as other related issues, such as the calibration and alignment of video sequences [19].

ITU-T Study Group 12 is working on a non-intrusive parametric model for the assessment of multimedia streaming (P.NAMS for short), which uses packet and codec information as inputs, but explicitly excludes any payload information. A follow-up project called P.NBAMS (B for bitstream) has similar goals, but will allow models to take payload information into account. The group also standardized an opinion model for videophone applications [20], and is extending the work to a planning model for audiovisual applications (G.OMVAS).

Furthermore, there is an ITU IPTV Global Standards Initiative (GSI),[2] whose task is to coordinate existing IPTV standardization activities. Among other things, it is working on recommendations for performance monitoring and quality of experience requirements for IPTV.

C. ATIS IIF

The ATIS IPTV Interoperability Forum (IIF) has a QoS Metrics (QoSM) committee that deals with QoS and QoE issues for IPTV. Among others, it has issued ATIS-0800008 [21], a standard for QoS metrics for linear broadcast IPTV. The list includes video, audio, and audiovisual MOS, but it does not specify how they are to be computed.

Some of the most interesting work currently being done in this committee is the creation of a test plan and a test process for the evaluation of quality metrics. In contrast to the VQEG test plans, each of which are tailored for a specific test, this test plan is intended as a basis for anyone who conducts an evaluation of objective quality models by subjective tests. The test plan is nearing completion and should become available as an ATIS standard later this year.

The test process complements the test plan and proposes a standardized *process* for the evaluation of quality metrics, rather than the standardization of the quality measurement *algorithms*. The premise is that it is sufficient for an algorithm to be validated and to perform well, but the algorithm itself does not need to be standardized. Parallels to this can be found in the voice over IP (VoIP) application space, for example, ITU-T Rec. P.564 [22], which includes a conformance test plan for VoIP quality models. The process enables on-demand validation of metrics at any time, thus encouraging innovation and more rapid advancement of the state of the art. Furthermore, the standardized test process and test plan provide opportunities for independent entities that conduct subjective tests on demand.

Other current work items of the ATIS IIF QoSM committee are an implementer's guide for QoS metrics and an IPTV QoE requirements document.

[2] See http://www.itu.int/ITU-T/gsi/IPTV/ for more information.

D. Other Committees

Some other groups also look at video QoE from various angles.

The Broadband Forum[3] (formerly known as the DSL Forum), for example, has published a report on QoE requirements [23].

The Video Services Forum (VSF)[4] has recommended transport-related metrics for video over IP [24]. It also started a QoE Activity Group last year, which is working on expanding the IP metrics defined in its existing report to taking into account the payload in terms of video and audio content.

V. Discussion and Conclusion

A variety of groups and organizations are working on QoS and QoE standards today, and significant progress has been made, both in terms of standardized approaches to QoE, as well as our understanding of the evaluation and performance of quality metrics.

One big contribution has been the video sequences from the VQEG FR-TV Phase I test, which as of today still represent the only database of video clips annotated with subjective ratings that is publicly available. VQEG also plans to release a subset of the annotated sequences from the HDTV test after its completion. There are additional efforts outside the standards groups to evaluate VQMs and make them more easily comparable through open databases of video sequences annotated with subjective ratings. One example for still images is the image quality database compiled by the Laboratory for Image & Video Engineering (LIVE) of the University of Texas at Austin,[5] which has become very popular in the research community. LIVE now plans to release an annotated video quality database based on its own subjective experiments. The Consumer Digital Video Library (CDVL)[6] currently under construction has a similar goal.

One of the issues today is the sheer abundance of VQMs, from free algorithms to commercial products, which are increasingly used by industry for various applications. Even standards such as the ITU Recommendations discussed above specify more than one algorithm, adding to the confusion of users. A more flexible on-demand validation process, such as the one in preparation by ATIS, may further increase the number of available choices in objective quality measurement. All this makes it harder to achieve the objectives of standards for meaningful and reproducible MOS values mentioned above in Section IV. Therefore, to assure comparability of quality measurement results across many platforms and VQM choices, there is a need to translate (or cross-calibrate) from the output of one VQM to that of another. A possible method using a common database of test sequences was proposed by Brill et al. [25] and also published in ATIS technical reports, but it still needs to be implemented more widely in practice.

[3] See http://www.broadband-forum.org/.
[4] See http://www.videoservicesforum.org/.
[5] See http://live.ece.utexas.edu/research/quality/.
[6] See http://www.cdvl.org/ (under construction).

REFERENCES

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

[2] ——, "Video quality and beyond," in *Proc. European Signal Processing Conference*, Poznań, Poland, September 3–7, 2007, invited paper.

[3] S. Winkler and P. Mohandas, "The evolution of video quality measurement: From PSNR to hybrid metrics," *IEEE Transactions on Broadcasting*, vol. 54, no. 3, pp. 660–668, September 2008, invited paper.

[4] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it?" *IEEE Signal Processing Magazine*, vol. 26, no. 1, pp. 98–117, January 2009.

[5] S. Winkler, "Perceptual video quality metrics – a review," in *Digital Video Image Quality and Perceptual Coding*, H. R. Wu and K. R. Rao, Eds. CRC Press, 2005, ch. 5.

[6] ITU-R Recommendation BT.500-11, "Methodology for the subjective assessment of the quality of television pictures," International Telecommunication Union, Geneva, Switzerland, 2002.

[7] ITU-T Recommendation P.910, "Subjective video quality assessment methods for multimedia applications," International Telecommunication Union, Geneva, Switzerland, 2008.

[8] S. Winkler, "On the properties of subjective ratings in video quality experiments," in *Proc. International Workshop on Quality of Multimedia Experience (QoMEX)*, San Diego, CA, July 29–31, 2009.

[9] K. Brunnström, D. Hands, F. Speranza, and A. Webster, "VQEG validation and ITU standardization of objective perceptual video quality metrics," *IEEE Signal Processing Magazine*, vol. 26, no. 3, pp. 96–101, May 2009.

[10] VQEG, "Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment," April 2000, available at http://www.vqeg.org/.

[11] A. M. Rohaly, P. Corriveau, J. Libert, A. Webster, V. Baroncini, J. Beerends, J.-L. Blin, L. Contin, T. Hamada, D. Harrison, A. Hekstra, J. Lubin, Y. Nishida, R. Nishihara, J. Pearson, A. F. Pessoa, N. Pickford, A. Schertz, M. Visca, A. Watson, and S. Winkler, "Video Quality Experts Group: Current results and future directions," in *Proc. SPIE Visual Communications and Image Processing*, vol. 4067, Perth, Australia, June 21–23, 2000, pp. 742–753.

[12] VQEG, "Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment – Phase II," August 2003, available at http://www.vqeg.org/.

[13] ITU-T Recommendation J.144, "Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference," International Telecommunication Union, Geneva, Switzerland, 2004.

[14] ITU-R Recommendation BT.1683, "Objective perceptual video quality measurement techniques for standard definition digital broadcast television in the presence of a full reference," International Telecommunication Union, Geneva, Switzerland, 2004.

[15] VQEG, "Final report from the Video Quality Experts Group on the validation of objective models of multimedia quality assessment," September 2008, available at http://www.vqeg.org/.

[16] ITU-T Recommendation J.247, "Objective perceptual multimedia video quality measurement in the presence of a full reference," International Telecommunication Union, Geneva, Switzerland, 2008.

[17] ITU-T Recommendation J.246, "Perceptual visual quality measurement techniques for multimedia services over digital cable television networks in the presence of a reduced bandwidth reference," International Telecommunication Union, Geneva, Switzerland, 2008.

[18] VQEG, "Final report from the Video Quality Experts Group on the validation of reduced-reference and no-reference objective models for standard definition television, Phase I," June 2009, available at http://www.vqeg.org/.

[19] ITU-T Recommendation J.244, "Calibration methods for constant misalignment of spatial and temporal domains with constant gain and offset," International Telecommunication Union, Geneva, Switzerland, 2008.

[20] ITU-T Recommendation G.1070, "Opinion model for video-telephony applications," International Telecommunication Union, Geneva, Switzerland, 2007.

[21] ATIS Standard 0800008, "QoS metrics for linear broadcast IPTV," Alliance for Telecommunications Industry Solutions, Washington, DC, USA, 2007.

[22] ITU-T Recommendation P.564, "Conformance testing for voice over ip transmission quality assessment models," International Telecommunication Union, Geneva, Switzerland, 2007.

[23] DSL Forum, "Triple-play services quality of experience (QoE) requirements," DSL Forum Architecture and Transport Working Group, Tech. Rep. TR-126, 2006.

[24] Video Services Forum, "Recommended video over IP metrics," VSF Test and Measurements Activity Group, Tech. Rep., 2006.

[25] M. H. Brill, J. Lubin, P. Costa, S. Wolf, and J. Pearson, "Accuracy and cross-calibration of video quality metrics: New methods from ATIS/T1A1," *Signal Processing: Image Communication*, vol. 19, no. 2, pp. 101–107, February 2004.