A CROWDSOURCING APPROACH TO VIDEO QUALITY ASSESSMENT

Babak Naderi, Ross Cutler
Microsoft Corporation, Redmond, USA
firstname.lastname@microsoft.com

ABSTRACT
We propose an open-source extension of the ITU-T Rec. P.910 subjective video quality test based on crowdsourcing principles. This extension addresses the speed, usage cost, and barrier to usage issues of P.910. We implement Absolute Category Rating (ACR), ACR with hidden reference (ACR-HR), Degradation Category Rating (DCR), and Comparison Category Rating (CCR), and include rater, environment, hardware, and network qualifications, as well as gold and trapping questions to ensure quality. We have validated that the implementation is both accurate and highly reproducible.

Index Terms: Subjective Quality, Video Quality, Crowdsourcing

1. INTRODUCTION
Measuring the quality of video is an important task in many engineering areas, such as video codec development, video enhancement, and video telecommunication systems. As a result, there are many standards and metrics for measuring video quality, though the gold standard is the subjective test done in a controlled lab environment. The most prevalent of these standardized subjective tests is the ITU-T Rec. P.910 [1]. However, using P.910 in practice is slow due to the recruitment of test subjects and the limited number of test subjects, and expensive due to paying qualified test subjects and the cost of the test lab. The speed and cost results in the vast majority of research papers not using P.910 but rather objective functions that are not well correlated to subjective opinion.

An alternative to lab-based subjective tests is to crowdsource the testing, and there are several such systems that do this (see Table 1). However, none of these systems have the rater, environment conditions, and hardware qualifications that P.910 requires. In addition, these existing systems have not been rigorously validated to show they are accurate compared to a P.910 lab-based study and give reproducible results. We provide a crowdsourced implementation that includes rater, environment, hardware, and network qualifications, as well as gold and trapping questions to ensure quality. We include a validation study that shows it is both accurate and highly reproducible compared to existing P.910 lab studies. The tool is open-sourced and can be used on the Amazon Mechanical Turk platform for wide-scale usage.

We recently used the tool in the CVPR 2022 CLIC challenge (http://compression.cc) to provide the challenge metric for machine learning-based video codecs.

Section 2 provides an overview of related work, Section 3 describes our tool’s implementation, Section 4 gives the tool’s validation, and Section 5 gives conclusions. The user interface, ablation studies, implementation details, and other analyses are given in [2].

2. RELATED WORK
A recent review of subjective image and video quality assessment tools is given in [3]. P.910 [1] provides a general subjective video quality assessment standard for multimedia applications. P.910 includes ACR, ACR-HR, DCR, and paired comparison (PC) methods, as well as rater qualifications, environment conditions, and video playback procedures. ITU-T Rec. P.911 is a counterpart of P.910 but for audiovisual signals. ITU-T Rec. P.912 [4] provides a target-specific subjective video quality assessment standard, such as for faces, license plates, etc. ITU-T Rec. P.913 [5] considers different displays and testing environments and provides flexibility on the rating scale and modality with mandatory reporting of test requirements [6]. ITU-T Rec. P.918 [7] details subjective assessment methods for five perceptual video quality dimensions, which can provide diagnostic information on the source of observed degradation. Finally, ITU-R BT.500 [8] focuses on the video quality of broadcast television signals in a highly controlled environment.

Tominaga et al. [9] conducted a comparison of eight different subjective video quality assessment methods and found that ACR was the most suitable for statistical reliability, assessment time, and ease of evaluation.

Keimel et al. [10] describes an open-source tool QualityCrowd that supports ACR video quality assessment. QualityCrowd is extended by [11] to include a Double Stimulus Continuous Quality Scale (DSCQS). Rainer et al. [12] describe the tool WESP, an open-source tool that supports ACR, ACR-HR, DCR, and PC.

Jung et al. [13] provide a methodology to conduct remote subjective video quality assessment studies in which the raters
download videos and view them manually on their devices. The methodology includes no tests for visual acuity, color blindness, environmental conditions, or hardware setup.

3. IMPLEMENTATION

Our implementation at https://github.com/microsoft/P910 can be used either as an integrated survey in Amazon Mechanical Turk (AMT) or as an external survey deployed on a dedicated Web server. We also provide a lightweight container-based web application that can serve the experiment. The web application can easily be deployed on any Linux virtual machine. As a result, this implementation can be used for both crowdsourcing tests or remote testing with a dedicated panel of participants.

The open-source implementation includes ACR, ACR-HR, DCR, and CCR. All methods can be used with either a five or nine-point discrete Likert scale. We also followed and extended best practices on video and speech quality assessment in crowdsourcing [15, 16] in our implementation.

3.1. Tools

We provide a set of program scripts to ease the interaction with the system and avoid operation errors. The scripts are used to create trapping sequences, process the submitted answers, and interact with AMT. Figure 1 shows the data flow diagram of the system.

Trapping sequences are customized to the dataset under the test and created by adding a text message asking participants to select a specific score. A test configuration and URLs for test sequences, trapping sequences, gold, and training clips are provided to the master script, which creates the HTML template of the test, a list of variables, and a configuration file for the result parser. Gold clips are video sequences whose quality is known to the experimenter. A test can be created in the HIT App Server by providing the HTML template and a list of variables. A generic project description and list of URLs can be downloaded and used to create a new test in AMT.

The submitted answers are provided to the result parser script along with the configuration file. The script performs data cleansing and aggregates the valid and reliable ratings over the test sequence and over the test condition (i.e., Hypothetical Reference Circuits - HRCs). Reports on the list of bonus assignments and lists of the submissions to be accepted/rejected or extended are also generated.

3.2. Test components

The subjective test is organized in different sections from the participant’s perspective (see Figure 2). Each section is designed to instruct the test participant, qualify the participant, their environment, their setup, and collect their votes.

The instructions and ratings sections are included in all tests, whereas the qualification, and calibration only need to be performed once per test. The setup and training sections are shown (e.g., once per hour).

Fig. 1. Data Flow Diagram.

A set of automatic measurements are performed on the test’s loading time. The experimenter can restrict test participants to specific viewing devices (i.e., mobile, PC, or both), minimum screen refresh rate, and minimum resolution.

3.2.1. Video playback

We developed an HTML5 video playback component that downloads all videos to the browser’s local storage to avoid network latency. Videos are played in full-screen mode and participants must watch the entire video before voting. The player records the playback duration, which is used in the data cleansing process. The experimenter can choose to present videos in their original size or scaled to fill the participant’s viewing device. For DCR or CCR tests, the reference and processed video sequences are shown sequentially with a one-second gray screen in between.

3.2.2. Qualification

Within the qualification section, the eligibility of test participants is evaluated. According to P910, participants should be screened for normal color vision and normal or corrected-to-normal visual acuity (i.e., no error on the 20/30 line of a standard eye chart) [17]. The standard Ishihara test for color vision [18] includes 15 plates in the normal and 6 plates in the short version, which are both too long for a crowdsourcing test. In a prestudy, we invited 300 participants from AMT.
Table 1. Open-source crowdsourcing video quality assessment systems

| Tool              | Measures          | Rater qualification | Viewing conditions | Hardware | Network | Accuracy | Reproducible |
|-------------------|-------------------|---------------------|--------------------|----------|---------|----------|--------------|
| QualityCrowd [10] | ACR, DSCQS        | N                   | N                  | N        | N       | Y        | N            |
| WESP [12]         | ACR, ACR-HR, DCR, PC | N               | N                  | N        | N       | N        | N            |
| avrateNG [14]     | ACR               | N                   | N                  | N        | N       | Y        | Y            |
| Ours              | ACR, ACR-HR, DCR, CCR | Y                | Y                  | Y        | Y       | Y        | Y            |

(34% male) and 191 participants (91% male) from two internet communities dealing with color vision deficiencies. Both groups participated in the full Ishihara test in which 6% of participants from AMT and 96% from the online forums were detected as color blind. Applying a decision tree classifier with entropy as the criterion revealed that only using Plates 3 and 4 reached 98% accuracy (sensitivity 0.996, specificity 0.95). Consequently, we only use these two plates in the qualification section.

We use the Landolt ring optotypes to measure visual acuity, as recommended by ISO 8596:2017 [19]. The participant is presented with broken rings (like the letter C) with a gap in 8 directions. The diameter of the ring is 5 times the size of the gap. The visual acuity is the inverse value of the gap size (in arc minutes) of the smallest identified Landolt ring. In each row (i.e., ring size), 5 samples are presented and the participant must answer 3 or more correctly to pass that size. Our implementation of the visual acuity test consists of two steps: setup and answering up to 5 Landolt rings at a specific size. In the setup section, the participant is asked to adjust the size of a given picture (here a credit card) on their screen until it is the same size as a real credit card. This is used to estimate the size of a pixel on their screen. The participant is then asked to sit in the range of 50 to 75 cm from the screen. The corresponding Landolt ring size is calculated, and the participant must correctly identify the direction of 3 out of 5 rings at that size to pass the test.

3.2.3. Display calibration and instructions

Participants are asked to set the resolution of their device to default or recommended value suggested by their operating system. We also ask them to perform display color calibration using methods provided by their operating system. This section provides information on how to perform these tasks for Windows and Mac operating systems and will only be shown once during the test.

In the instruction section, a sample video for some of the perceptual quality dimensions [20,7] (i.e., fragmentation, discontinuity, and uncleanness) is presented to the participants. They are also informed that impairments can happen in a specific area or time within a video clip and the impairments are not limited to the presented samples.

3.2.4. Setup

The setup section evaluates the viewing conditions, including the brightness of the screen, room light, and viewing distance. A brightness/light calibration task is used, in which participants count the number of geometric shapes in a picture. If they get the answer wrong, they are encouraged to adjust their screen brightness, lighting, or repeat the task. The picture is a matrix of 4x4 squares, with each square having a different gray background and a triangle or circle in different sizes and locations. The foreground color of the shapes is close to the background color of the square.

The crowdsourcing test also includes a short paired-comparison task to evaluate participants’ viewing distance. This task is inspired by [21], which is the recommended method to evaluate crowd workers’ setup and environment in a speech quality assessment test [22]. In this task, three pairs of images are presented to the participants, who are asked to select the image of better quality. One of the images in each pair is distorted with a blur effect. The blur effects are selected based on [23], which found that participants can correctly distinguish between blurred images at different distances.

3.2.5. Training and rating

The training section includes videos from the training set, which cover the entire range of the rating scale. It also includes a trapping clip to test the participant’s understanding. Participants are alerted if they provide a wrong answer and asked to watch the video again. The training section is shown periodically to keep its anchoring effect, following best practices in the speech quality domain [24].

The rating section includes ten video clips, one trapping clip, and one gold clip. The trapping and gold clips are inserted automatically and used in the data cleansing step in post-processing.

4. VALIDATION

We used the VQEG HDTV datasets [25] to evaluate the validity and reliability of our implementation. The datasets contain coding only and coding plus transmission error impairments. They were created to validate objective video quality models that predicted the quality of High Definition Television (HDTV). The video materials and subjective data from...
Table 2. Comparison between laboratory and crowdsourcing experiments (sequence level).

| Dataset          | MOS PCC | MOS SPCC | RMSE | FOM PCC | FOM SPCC | RMSE FOM |
|------------------|---------|----------|------|---------|----------|----------|
| VQEG HDTV3-ran1  | 0.956   | 0.949    | 0.333 | 0.948   | 0.949    | 0.362    |
| VQEG HDTV3-ran2  | 0.964   | 0.951    | 0.302 | 0.946   | 0.939    | 0.370    |
| VQEG HDTV3-ran3  | 0.959   | 0.949    | 0.323 | 0.940   | 0.942    | 0.389    |
| VQEG HDTV3-ran4  | 0.917   | 0.913    | 0.455 | 0.904   | 0.922    | 0.489    |
| VQEG HDTV3-ran5  | 0.947   | 0.923    | 0.367 | 0.932   | 0.909    | 0.415    |
| VQEG HDTV5       | 0.970   | 0.957    | 0.278 | 0.965   | 0.958    | 0.299    |

Table 3. Comparison between laboratory and crowdsourcing tests in HRC level.

| Dataset          | MOS PCC | MOS SPCC | RMSE | RMSE FOM |
|------------------|---------|----------|------|----------|
| VQEG HDTV3-ran1  | 0.967   | 0.980    | 0.655| 0.248    |
| VQEG HDTV3-ran2  | 0.977   | 0.982    | 0.618| 0.211    |
| VQEG HDTV3-ran3  | 0.968   | 0.981    | 0.577| 0.245    |
| VQEG HDTV3-ran4  | 0.940   | 0.975    | 0.706| 0.333    |
| VQEG HDTV3-ran5  | 0.965   | 0.972    | 0.671| 0.257    |

Table 4. Correlation coefficients between five runs of the VQEGHD3 dataset. Pearson correlation coefficients are on the upper triangle and Spearman’s rank correlation coefficients are on the lower triangle.

| Run 1 | Run 2 | Run 3 | Run 4 | Run 5 |
|-------|-------|-------|-------|-------|
| Run 1 | 0.984 | 0.987 | 0.957 | 0.977 |
| Run 2 | 0.959 | 0.985 | 0.957 | 0.977 |
| Run 3 | 0.974 | 0.969 | 0.952 | 0.972 |
| Run 4 | 0.943 | 0.941 | 0.942 | 0.956 |
| Run 5 | 0.954 | 0.947 | 0.942 | 0.933 |

4.2. Reproducibility

On average 63 unique workers participated in each run. We observed a $PCC = 0.971$ and $SRCC = 0.95$ on average between the MOS values of sequences in the five different runs. The correlation coefficients between the runs are reported in Table 4. We also fitted a linear mixed-effects model (LMMs) with random intercept in which test sequences and runs are considered as fixed factors and participants as a random factor. The result shows there was no significant effect of runs on the subjective ratings ($\chi^2(4) = 2.691, p = 0.611$).

5. DISCUSSION AND CONCLUSIONS

We have created the first crowdsourced implementation of P910 which is shown to be as accurate as a lab study and is highly reproducible. In a set of experiments, we showed that each test component significantly improves the validity of the collected ratings. Analysis in [2] shows the qualification, calibration, trapping, and gold questions all give statistically significant improved accuracy.

We plan to include the simultaneous presentation of sequence pairs for future work and include P910 Annex E: An advanced data analysis technique for tests under challenging conditions. In addition, this framework can be easily extended to measure image quality assessment.
6. REFERENCES

[1] ITU-T Recommendation P.910, “Subjective video quality assessment methods for multimedia applications,” 2021.

[2] Babak Naderi and Ross Cutler, “A crowdsourcing approach to video quality assessment,” Sept. 2022, arXiv:2204.06784 [eess].

[3] Michela Testolina and Touradj Ebrahimi, “Review of subjective quality assessment methodologies and standards for compressed images evaluation,” in Applications of Digital Image Processing XLIV. Aug. 2021, vol. 11842, pp. 302–315, SPIE.

[4] ITU-T Recommendation P.912, “Subjective video quality assessment methods for recognition tasks,” 2016.

[5] ITU-T Recommendation P.913, Methods for the subjective assessment of video quality, audio quality and audiovisual quality of Internet video and distribution quality television in any environment, International Telecommunication Union, Geneva, 2021.

[6] Margaret H Pinson and Lucjan Janowski, “A new subjective audiovisual & video quality testing recommendation,” An Era of Change, p. 51, 2014.

[7] ITU-T Recommendation P.918, Dimension-based subjective quality evaluation for video content, International Telecommunication Union, Geneva, 2021.

[8] ITU-R Recommendation BT.500-14, Methodologies for the subjective assessment of the quality of television images, International Telecommunication Union, Geneva, 2019.

[9] Toshiko Tominaga, Takanori Hayashi, Jun Okamoto, and Akira Takahashi, “Performance comparisons of subjective quality assessment methods for mobile video,” in 2010 Second International Workshop on Quality of Multimedia Experience (QoMEX), June 2010, pp. 82–87.

[10] Christian Keimel, Julian Habigt, Clemens Horsch, and Klaus Diepold, “QualityCrowd-A Framework for Crowd-based Quality Evaluation,” in Picture Coding Symposium, 2012.

[11] “Large-Scale Crowdsourcing Subjective Quality Evaluation of Learning-Based Image Coding,” IEEE Visual Communications and Image Processing (VCIP 2021), 2021.

[12] Benjamin Rainer, Markus Waltl, and Christian Timmerer, “A web based subjective evaluation platform,” in 2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX), July 2013, pp. 24–25.

[13] J. Jung, M. Wien, and V. Baroncini, “ISO/IEC JTC 1/SC 29/AG 5: Guidelines for remote experts viewing sessions,” Nov. 2021.

[14] Rakesh Rao Ramachandra Rao, Steve Göring, and Alexander Raake, “Towards High Resolution Video Quality Assessment in the Crowd,” in 2021 13th International Conference on Quality of Multimedia Experience (QoMEX), June 2021, pp. 1–6, ISSN: 2472-7814.

[15] Tobias Hossfeld, Christian Keimel, Matthias Hirth, Bruno Gardlo, Julian Habigt, Klaus Diepold, and Phuoc Tran-Gia, “Best Practices for QoE Crowdtesting: QoE Assessment With Crowdsourcing,” IEEE Transactions on Multimedia, vol. 16, no. 2, pp. 541–558, Feb. 2014.

[16] ITU-T PSTR-CROWDS, Subjective evaluation of media quality using a crowdsourcing approach, International Telecommunication Union, Geneva, 2018.

[17] ITU-T Recommendation P.910, Subjective video quality assessment methods for multimedia applications, International Telecommunication Union, Geneva, 2021.

[18] JH Clark, “The ishihara test for color blindness.,” American Journal of Physiological Optics, 1924.

[19] “ophthalmic optics — visual acuity testing — standard and clinical optotypes and their presentation,” Standard, Geneva, CH.

[20] Falk Schifflner and Sebastian Möller, “Defining the relevant perceptual quality space for video and videotelephony,” in 2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX). IEEE, 2017, pp. 1–3.

[21] Babak Naderi and Sebastian Möller, “Application of just-noticeable difference in quality as environment suitability test for crowdsourcing speech quality assessment task,” in 2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX). IEEE, 2020, pp. 1–6.

[22] ITU-T Recommendation P.808, Subjective evaluation of speech quality with a crowdsourcing approach, International Telecommunication Union, Geneva, 2021.

[23] Bernhard Treutwein, “Adaptive psychophysical procedures,” Vision research, vol. 35, no. 17, pp. 2503–2522, 1995.

[24] Babak Naderi and Ross Cutler, “An Open Source Implementation of ITU-T Recommendation P.808 with Validation,” INTERSPEECH, pp. 2862–2866, Oct. 2020.

[25] Video Quality Experts Group, “Report on the validation of video quality models for high definition video content,” 2010.