https://doi.org/10.2352/EI.2025.37.11.HVEI-211

Cloud gaming quality based on a passive video quality experiment and bootstrapped analysis

Kjell Brunnström[a,b], Linnea Runsten Fredriksson[a] and Shirin Rafiei[a,b]
[a]Department of Industrial Systems, RISE Research Institutes of Sweden, Kista, Sweden
[b]Department of Computer and Electrical Engineering (DET), Mid Sweden University, Sundsvall, Sweden

Abstract

The International Telecommunication Union has a project for developing objective quality models called Parametric Bitstream-Based Quality Assessment of Cloud Gaming services. The model will be divided into an interaction quality module and a video coding impairment module. To evaluate these two modules an experimental campaign was conducted where labs from different parts of the world performed user studies to collect data for the evaluation. This paper describes an experiment for collecting data to evaluate the video coding impairment module. The analysis is based on a bootstrapping approach.

Keywords

Cloud gaming, Video quality, Visual perception, Standardization, Bootstrapping

Introduction

Cloud gaming, also known as game streaming, represents a significant shift in the gaming industry by allowing users to play high-quality games on various devices without the need for powerful hardware. This technology leverages cloud servers to process and render games, streaming the gameplay to users over the internet. As a result, players can enjoy a seamless gaming experience on devices ranging from smartphones to low-end PCs. The rapid growth of cloud gaming services, such as Google Stadia, NVIDIA GeForce Now, and Microsoft's Xbox Cloud Gaming, highlights the increasing demand for accessible and flexible gaming solutions. However, ensuring a high Quality of Experience (QoE) [1, 2] for users remains a critical challenge, necessitating robust methods for assessing and optimizing the performance of these services.

The International Telecommunication Union (ITU) Telecommunication Standardization Sector (ITU-T) initiated a project in 2020 to develop a Quality of Experience (QoE) model that uses the bitstream metadata for Parametric Bitstream-Based Quality Assessment of Cloud Gaming services (P.BBQCG). The model will be divided into an interaction quality module and a video coding impairment module. To evaluate these two modules an experimental campaign was conducted where labs from different parts of the world performed user studies to collect data for the evaluation.

This paper describes an experiment for collecting data to evaluate the video coding impairment module. Another objective is to evaluate bootstrapped analysis for this type of experimental data.

Method

For the passive test a number of conditions were included the experiment. These were:

- Codecs: 3 codecs (AVC, HEVC, AV1)
- Bitrate: 5 levels of bitrate
- Resolution: 540p, 720p, 1080p, 2160p
- Framerate: 30fps, 60fps, 120fps
- Source video: 14

As making a full matrix design incorporating all the conditions, would be too many, a subset was selected using VMAF [3]. The focus for encoding was HEVC, but for the high-quality range AV1 was used and for easy characterization of baseline model H.264 was included. Out of the selection 16 databases with 90 processed video sequences were defined. Five were in HD and had high framerate of 120 frame per seconds (FPS) included and 11 had maximum resolution of 4K with the framerates of 30 and 60 FPS. This paper describes the experiment based on one of the 11 4K databases.

Procedure

Test participants were invited to the Lab. On arrival they were welcomed and asked to sign the consent form for the study. They were informed that they could stop at any time without giving a reason and without suffering any negative consequences. They were then given the instructions to read and pre-questionnaire was filled in. If the instructions were clear, the test participants' visual acuity (Snellen 3 meters) and color vision (Ishihara) were tested. The test participants were invited to the lab room and seated at 1.5 screen heights from the screen i.e. 1.2 meters. There was a training session consisting of five videos to familiarize the participants with the procedure and the voting process. After the training, the main session took place, where 90 videos were watched and scored. The scoring was done on a scale from Extremely bad to Ideal as shown in Figure 1. The test was ended by performing a short interview with the test participants.

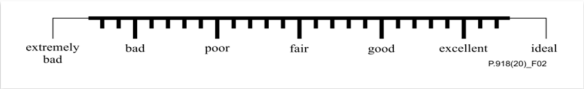


Figure 1: Scale for scoring the quality, that was used in the study

Apparatus

The test room was set up to comply with the requirements of the ITU-R Rec. BT.500-15 [4]. A high-end consumer-grade 65 inch 4K TV (Ultra HD, LG OLED65E7V) was used for the experiments, having a resolution of 3840 by 2160 pixels. Each video was randomized for each participant. For the video playback and randomization, the video player and rating program AVRateNG [5] was used on a Windows 10 computer with a Intel Core i7-6850 K and RAM of 24 GB.

Test participants

Twenty-six test participants (26) took part in the study, with equal numbers of males and females. The age varied from 19 to 59 years old with an average of 32. The gaming skill level varied among the participants from beginners to experts

Analysis

Of interest for the evaluation of objective quality models from this type of experiment are the means of the ratings for each PVS, a.k.a. Mean Opinion Score (MOS), and its confidence interval [6]. This is to be able to make inferences about whether the differences between different MOS are significant or not.

The analysis of this data set was primarily done using bootstrapped analysis. This is a technique to analyze different parameters of data by repeatedly resampling the data [7, 8]. If the data collected is not Normally distributed, parametric statistics should not be used. With bootstrapping calculations of key characteristics of the data can be done without making any assumptions about the underlying distribution. For example, mean, confidence intervals and statistical hypothesis tests. We will compare the direct calculations with the bootstrapped results for a better understanding of the experimental data and results. Comparisons will be made for mean values, confidence intervals and shape of the distribution. We used the Shapiro-Wilk's test to test the Normality, both using the p-values and with the W-statistics from the test. A problem with the Normality tests in general are that they get overly sensitive when the number of samples increases but analyzing the W-statistics which approach 1 when the distribution becomes more Normal, we can see if this was indeed the case for the bootstrapping distributions[9, 10].

Results

In Figure 2Figure 2, the distribution of votes is presented as a bar chart, with each bar corresponding to one of the seven levels of the scale shown in Figure 1. It can be noted that as expected the midrange are more used than the ends. If we consider the ends of this particular scale being designed to encourage the use of the whole scale between bad and excellent, then adding together the end point with Bad and Excellent, we get an almost uniform voting pattern, as shown in Figure 3.

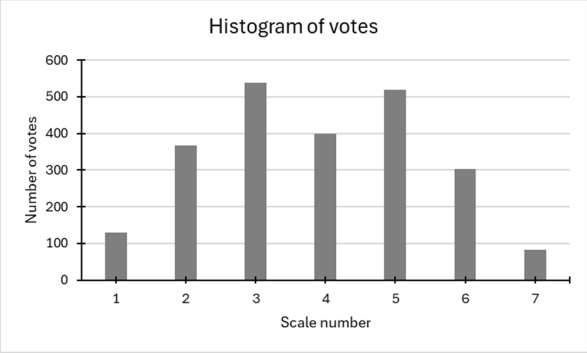


Figure 2: Distribution of votes across the scale used shown in Figure 1. Here showing the number of votes per scale level.

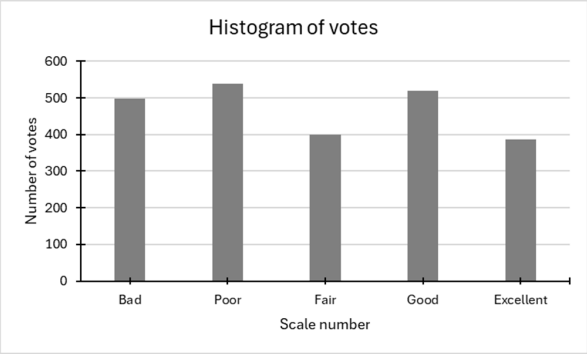


Figure 3: Distribution of votes across the scale used shown in Figure 1. The number of votes when aggregating the lower end to Bad and the higher end to Excellent.

In Figure 4 a comparison between non-bootstrapped and bootstrapped means and confidence intervals based on 1000 resampling for each PVS grouped by SRC, are shown. We can see that although there are several cases that violates the Normality assumptions (Table 1), we get good correspondence between the non-bootstrapped and bootstrapped mean values and confidence intervals. An example of the shape change in distributions is shown Figure 5. The closeness to a Normal distribution are shown in Figure 6 and in Table 1, where a value closer to 1 is indicating a more Normal distribution.

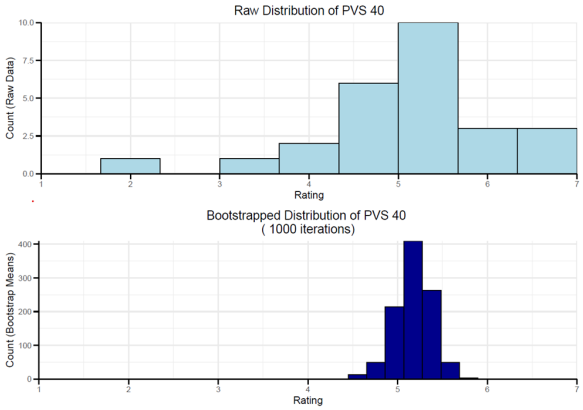


Figure 5: Example of reshaping of a distribution based on bootstrapping. The upper graph shows the raw data distribution, and the lower graph shows the bootstrapped distribution.

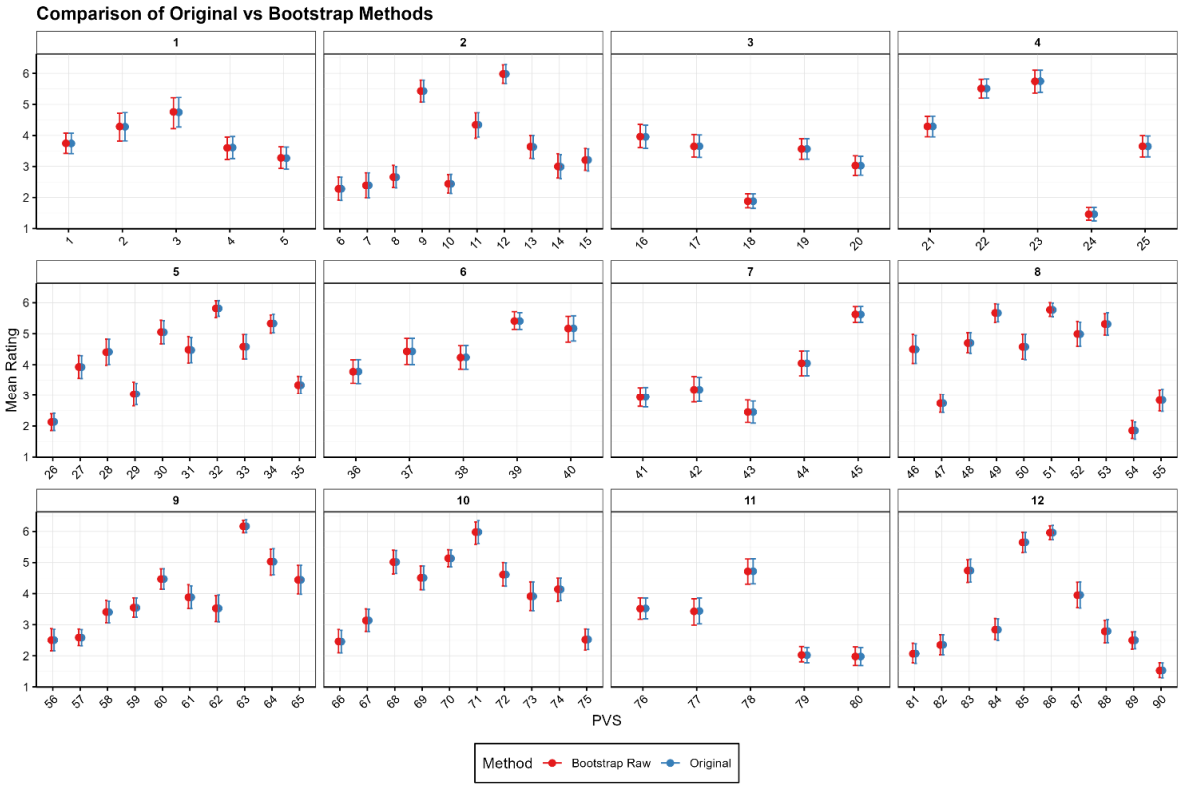


Figure 4: Comparison between non-bootstrapped and bootstrapped means and confidence intervals based on 1000 resampling. Each sub figure represents one SRC.

Table 1: P-values and W-statistics from Shapiro-Wilk's test for the significant cases both for non-bootstrapped and for bootstrapped. The number of iterations was 1000.

	P-value Non-Bootstr.	P-value Bootstr.	W-Statistics Non-bootstr.	W-Statistics Bootstr-
7	0.0448	0.0019	0.920	0.995
12	0.0156	0.0003	0.900	0.994
14	0.0010	0.0000	0.842	0.989
23	0.0220	0.0150	0.907	0.996
24	0.0002	0.0001	0.802	0.993
40	0.0085	0.0003	0.888	0.994
41	0.0015	0.0001	0.851	0.993
43	0.0298	0.0492	0.912	0.997
45	0.0045	0.0003	0.875	0.994
51	0.0251	0.0003	0.909	0.994
54	0.0072	0.0160	0.885	0.996
66	0.0070	0.0000	0.884	0.991
71	0.0000	0.0000	0.709	0.975
90	0.0003	0.0009	0.811	0.994

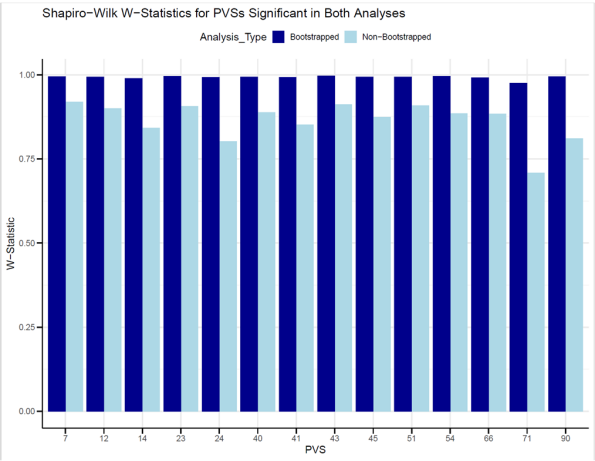


Figure 6: Shapiro-Wilk's W-statistics for the PVS:s (video in the test) that was significantly not Normal based on Shapiro-Wilk's Normality test. Bootstrapping has been done for 1000 iterations.

In Figure 7 the Mean Opinion Scores (MOS) vs bitrate along logarithmic scaled x-axis, is shown for the estimated actual bitrate of the encoded videos in the experiment. We can see that the MOS follows a linear relationship with a correlation of 0.84.

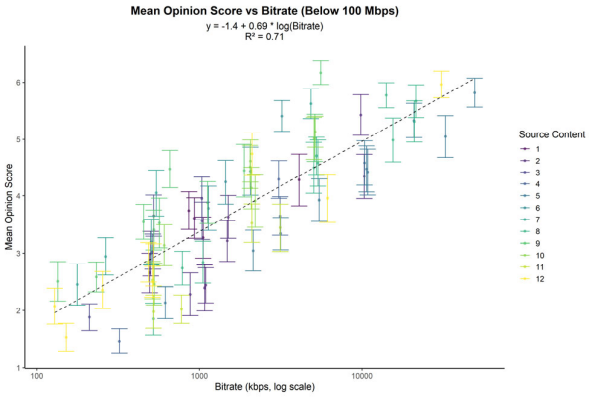


Figure 7: Mean Opinion Score vs log bitrate for the encoded videos. The error bars represent 95% confidence intervals. Each source are drawn with different colors, but the fitting is based on all PVS except the uncompress reference video (not shown in the graph).

Conclusions

This work investigated the perceived quality of cloud gaming in a passive view situation. An experiment with 26 test participants were conducted. The study was a part of a larger effort by the ITU-T for developing a QoE model that use the bitstream meta data for cloud gaming services. The data could not be confirmed to fully follow a Normal distribution, so a bootstrapping approach was adopted, that does not require any assumption on the underlying distribution. We could show that the bootstrapping results were very similar to the results based on the raw data. The bootstrapping made also the data that was not tested to be Normal when non-bootstrapped became quite close to Normal after bootstrapping. Indicating that the underlying distribution was indeed Normal. Based on the calculated Mean Opinion Scores of the PVS:s and the logarithm of the bitrates of the encoded video, we could show that the quality had a linear dependency.

References

[1] Le Callet, P., S. Möller, A. Perkis, K. Brunnström, K. De Moor, A. Dooms, S. Egger, M.-N. Garcia, T. Hossfeld, S. Jumisko-Pyykkö, C. Keimel, C. Larabi, B. Lawlor, F. Pereira, M. Pereira, A. Pinheiro, U. Reiter, P. Reichl, R. Schatz, P. Schelkens, L. Skorin-Kapov, D. Strohmeier, C. Timmerer, M. Valera, J. You, and A. Zgank. (2012). *Qualinet White Paper on Definitions of Quality of Experience* (Version 1.2 (http://www.qualinet.eu/images/stories/QoE_whitepaper_v1.2.pdf)). European Network on Quality of Experience in Multimedia Systems and Services (Qualinet), Lausanne, Switzerland.

[2] ITU-T. (2017). *Vocabulary for performance, quality of service and quality of experience* (ITU-T Rec. P.10/G.100). International Telecommunication Union (ITU), Place des Nations, CH-1211 Geneva 20.

[3] Li, Z., A. Aaron, I. Katsavounidis, A.K. Moorthy, and M. Manohara (2016). *Toward A Practical Perceptual Video Quality Metric*. Netflix Technology Blog. Available from: https://medium.com/netflix-techblog/toward-a-practical-perceptual-video-quality-metric-653f208b9652, Access Date: Oct 23, 2018.

[4] ITU-R. (2023). *Methodology for the subjective assessment of the quality of television pictures* (ITU-R Rec. BT.500-15). International Telecommunication Union (ITU).
[5] Göring, S., R.R.R. Rao, S. Fremerey, and A. Raake. (2021). *AVrate Voyager: an open source online testing platform*. in *2021 IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP)*. 2021.
[6] ITU-T. (2020). *Statistical analysis, evaluation and reporting guidelines of quality measurements* (ITU-T P.1401). International Telecommunication Union, Telecommunication standardization sector, Geneva, Switzerland.
[7] Davison, A. and D. Kuonen, (2003). *An Introduction to the Bootstrap with Applications in R.* Statisticial Computing and Statistical Graphics Newsletter. **13**.
[8] Efron, B., (1979). *Bootstrap Methods: Another Look at the Jackknife.* The Annals of Statistics. **7**(1): p. 1-26, 26.
[9] Shapiro, S.S. and M.B. Wilk, (1965). *An Analysis of Variance Test for Normality (Complete Samples).* Biometrika. **52**(3/4): p. 591-611, DOI: 10.2307/2333709.
[10] Razali, N.M. and Y.B. Wah. (2011). *Power comparisons of Shapiro-Wilk , Kolmogorov-Smirnov , Lilliefors and Anderson-Darling tests*. 2011.

Acknowledgement

This research has been funded by Sweden´s Innovation Agency (VINNOVA, dnr. 2021-02107 and 2023-00755). We would also like to thank Saman Zadtootaghaj, and David Lindero, for inviting and guiding us to contribute to ITU-T Study Group (12 work items - ITU-T P.1204.3), known as P.BBQCG, a Bitstream-Based Quality Assessment Model for Cloud Gaming Services and the rest of all the labs, companies and organizations that contributed to the effort.

Author Biography

Kjell Brunnström is a Senior Scientist at RISE Research Institutes of Sweden AB and Adjunct Professor at Mid Sweden University. He is leading development for video quality assessment as Co-chair of the Video Quality Experts Group (VQEG). His research interests are in Quality of Experience for visual media especially immersive media. He is area editor of the Elsevier Journal of Signal Processing: Image Communication and has co-authored > 100 peer-reviewed scientific articles including conference papers.

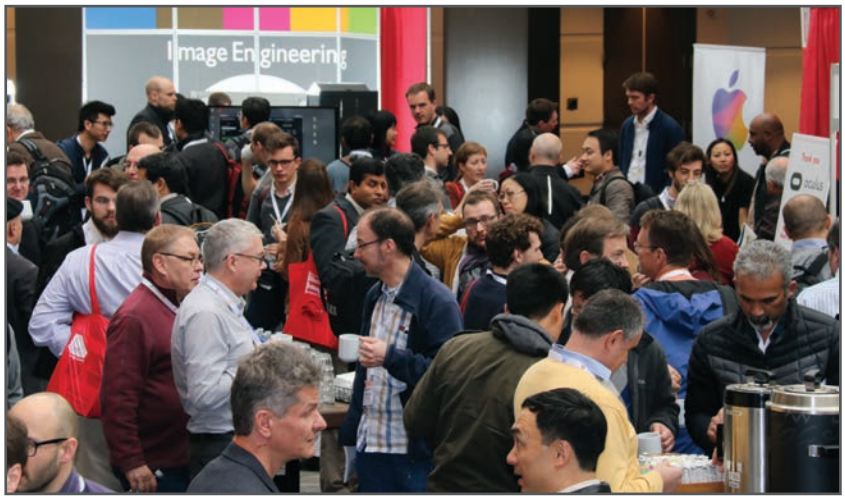
Linnea Runsten-Fredriksson earned her B.Sc. in Computer and Systems Sciences from Stockholm University in 2023. She conducted and facilitated user testing for this study during an internship at RISE Research Institutes of Sweden. Since 2024, she has been working as a user researcher at Telia, focusing on UX research related to user interaction, usability, and accessibility.

Shirin Rafiei received her B.Sc. and M.Sc. degrees in Electronics and Telecommunications in 2009 and 2014, respectively. Since 2020, she has been a researcher and Ph.D. fellow at RISE Research Institutes of Sweden and Mid Sweden University. Her research focuses on interdisciplinary mixed-method approaches, integrating UX and QoE in industrial remote-control systems. She also works in extended reality applications and explores advanced user interaction paradigms using visual interfaces for remote control systems, with emphasis on visual perception.

JOIN US AT THE NEXT EI!

electronic IMAGING

Imaging across applications . . . Where industry and academia meet!



• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

ADVANCING IMAGING SCIENCE 75 1947 IS&T 2022 imaging.org