https://doi.org/10.2352/EI.2024.36.11.HVEI-219

Estimating Metric Thresholds for Acceptability and Annoyance of User Generated Video Content

Ali Ak, Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France
Abhishek Gera, Meta, USA
Denise Noyes, Meta, USA
Francois Blouin, Meta, USA
Hassene Tmar, Meta, USA
Ioannis Katsavounidis, Meta, USA
Patrick Le Callet, Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France, Institut Universitaire de France (IUF)

Abstract

User expectation is one of the main factors that drives the user satisfaction for video streaming service providers and online social media platforms. Depending on the context, users may have different expectations of the video quality. Measuring the Quality of Experience (QoE) by taking user expectations into account provide online social media platforms with increased efficiency and users with higher satisfaction. In this work, we explore the relation between video quality and acceptability&annoyance of video quality in online social media platforms context. Moreover we present the methodology to determine the metric thresholds for acceptability&annoyance of video quality. We compare the estimated thresholds with previous studies.

Introduction

Quality of Experience (QoE) in video streaming context defines the user satisfaction and the level of fulfilment of user expectations when viewing a video content. There are many factors that affect the QoE as discussed in detail in Qualinet white paper [1]. Along with the traditional video quality, other factors that affect the QoE include fidelity, display specifications, delivery, storage and processing costs.

Many subjective quality assessment datasets and objective quality metrics were developed and proposed in the last decades to determine and predict the visual quality of professional and user generated video content. However, from the point of view of the video streaming service providers, visual quality of the video content is not the ideal metric to maximize since alone it is not enough to understand whether the delivered content fulfills the user expectations. It is rather important to optimize the imaging pipeline around the users' satisfaction which are driven by their expectations and the quality of the provided service (*e.g.,* video quality). Therefore, it is desirable to understand the relation between visual quality and user satisfaction and filling the gap between the video quality and QoE. To this end, acceptability&annoyance paradigm has been introduced and further improved in recent years[2, 3, 4, 5]. Acceptability&annoyance scale contains three categories as "not acceptable (1)", "acceptable but annoying (2)", and "not annoying (3)". Acceptability&annoyance methodology was initially proposed as a multi-step evaluation scenario and later evolved into a single-step evaluation that would sufficiently capture the same effect [4], as shown in Figure 1

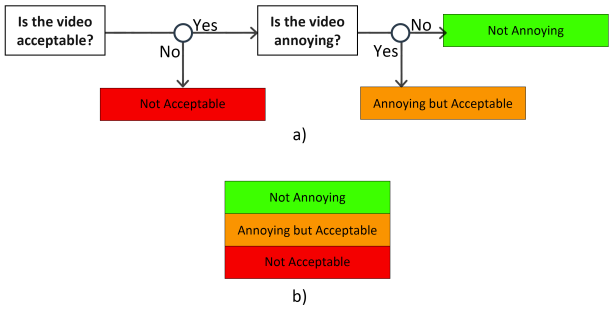

Figure 1. *a) multi-step b) single-step acceptability&annoyance experiment design.*

Previous studies in QoE domain demonstrates that the video quality does not necessarily indicate the acceptability&annoyance of the video content since the thresholds where the video quality drops to annoying and unacceptable levels are heavily influenced by user expectations [6]. However a relation between the video quality and acceptability&annoyance can be drawn for a given context [3, 4, 7]. We define the context here as the context in which the video content is consumed. For example, the context can vary from online social platforms (*e.g.,* Instagram Reels, Tiktok, etc.) to online video streaming services (*e.g.,* Netflix, Amazon Prime Video, etc.) based one where it is consumed. In addition it can define other aspects such as remaining battery or remaining data plan of the users [5].

In this work, we utilize publicly available IPI-VUGC Dataset [1] and estimate the VMAF [8] metric thresholds for different video quality ranges. We then compare the estimated thresholds to previous studies with various contexts. Moreover, we provide a brief introduction of the IPI-VUGC dataset and the algorithm used to determine the metric thresholds.

IPI-VUGC Dataset

IPI-VUGC is a publicly available dataset of video quality and acceptability&annoyance for 336 videos. The subjective opinion scores are collected in laboratory conditions with an

[1] https://zenodo.org/doi/10.5281/zenodo.10475209

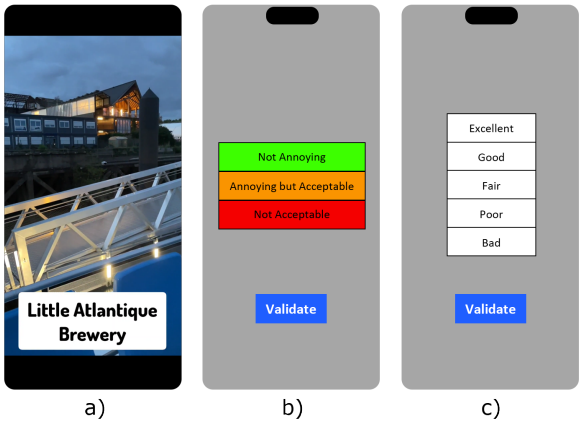

Figure 2. *Sample screens captured during a) stimuli presentation, b) voting screen in acceptability&annoyance experiment, and c) voting screen in video quality experiment. The screen captures are taken from IPI-VUGC dataset under a CC-BY-SA 4.0 license.*

Iphone 14 Pro [2]. Sample screenshots during the stimuli presentation and voting screens of each experiment are visualized in Figure 2.

Content

The content used in the dataset aims to reflect the properties of user generated video content typically found in online social media platforms. To this end 30% of the videos contain sticker or text overlays. The videos are 5 seconds long (typical short term video length) and has vertical orientation. An example stimuli with text overlay is presented in Figure 2-a.

6 Processed video stimuli (PVS) are generated for each SRC by compressing with h264 [9] at various resolutions and constant rate factors(CRF). Thus results in 336 videos, including the SRC videos. The videos are rendered in the native size and resolution of the display and the upscaling is handled by the device. The gaps due to aspect ratio differences between the display and the video content are filled with black color. The device is kept at a fixed brightness during the experiment. Participants could held the device freely during the experiment. Rest of the experiment details follows the ITU recommendations[10].

Experiment Design

Two separate experiments were conducted with the same content to collect subjective opinions on video quality and acceptability&annoyance of the video content.

An Absolute Category Rating with Hidden Reference (ACR-HR) experiment is conducted to collect the video quality scores. Figure 2-c presents the voting screen shown after each video stimuli in this experiment. Classical video quality scale (“Bad”, “Poor”, “Fair”, “Good”, “Excellent”) is used and numerically represented in [1, 5] range where higher values indicate higher video quality. They are represented as Mean Opinion Scores and referred as ACR-MOS in the dataset. ACR-MOS values are calculated with the ZREC [11] MOS recovery algorithm.

To collect acceptability&annoyance labels, the dataset relies on the single step acceptability&annoyance procedure. Instructions are used to simulate the user expectations similar to previous works [5, 4]. Precisely, the IPI-VUGC dataset uses the following instructions to set user expectations in the online social media platform context:

“You are going to participate in an experiment determining the acceptability and annoyance of videos. You will need to imagine yourself scrolling through your preferred social media platform (*e.g.,* Facebook, Instagram, TikTok, etc.) and encountering these videos. Based on your expectations of the video quality in these encounters, you will need to rate the quality of the video in terms of Acceptability and Annoyance.

- The video is not annoying when its quality satisfies or exceeds your expectations.
- The video is annoying but acceptable when its quality is acceptable but not completely satisfies your expectations.
- The video is not acceptable when its quality does not meet your expectations. Such video quality makes you think about skipping to the next video.”

The acceptability&annoyance labels are referred as AccAnn-MOS in the dataset and are numerically in the range [1, 3]. 1, 2 and 3 corresponds to “Not Annoying” (NAnn), “annoying but acceptable” (AA), and “not acceptable” (NAcc), respectively. Similar to ACR-MOS, AccAnn-MOS values are calculated with the ZREC [11] MOS recovery algorithm.

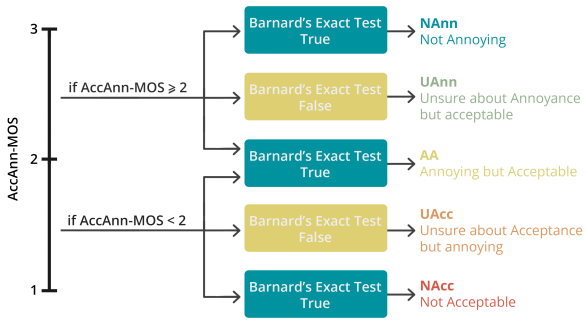

Figure 3. *Overview of the algorithm to determine the Acceptability&Annoyance categories.*

Determining Acceptability&Annoyance Categories and Thresholds

In IPI-VUGC Dataset, each stimuli is rated by 25 unique observers to collect the acceptability&annoyance labels. Simply averaging the numerical representations of these labels provides us with AccAnn-MOS. For certain applications, a simple acceptability&annoyance category might be more desirable. To this end, we can estimate the acceptability&annoyance categories by using Barnard’s test [12]. It is an exact test developed for 2×2 contingency tables. Without loss of generality, Barnard’s test determines the statistically significant difference between the two sets of observations.

We follow the algorithm summarized in Figure 3 similar to those in [4, 5]. In addition to the 3 acceptability&annoyance categories collected in the experiment, we utilize 2 threshold categories as “unsure about annoyance but acceptable” (UAnn) and

[2]https://www.apple.com/go/2022/iphone-14-pro/

	Acceptability	Annoyance
Online Social Media	34	67
Basic Subscription [4]	58	80
Premium Subscription [4]	71	87

VMAF [8] thresholds for acceptability and annoyance of video quality in the contexts of online social media platforms, basic and premium subscription to video streaming services.

"unsure about acceptability but annoying" (UAcc). If the mean acceptability&annoyance score is greater than or equal to 2, we assign the stimuli to one of the NAnn, UAnn, and AA categories based on the Barnard's test result. If the Barnard's test results show that there is not a statistically significant agreement among the participants on one of the neighboring main categories (NAnn or AA), they are assigned to the threshold category (UAnn). Similarly, if the mean acceptability&annoyance score is smaller than 2, we assign the stimuli to one of the AA, UAcc, NAccc categories based on the Barnard's test result.

To estimate the acceptability and annoyance thresholds in terms of objective metric scores, we simply take the average of objective quality metric scores of the stimuli in the threshold categories UAcc and UAnn for acceptability and annoyance thresholds, respectively. In this work, we rely on VMAF [8] for threshold estimation. VMAF is an efficient and reliable full-reference video quality metric that is widely adopted in the industry. It is not a good predictor for video quality when the reference content is heavily distorted as it might be observed in the UGC domain. Despite the drawback, we relied on VMAF to be able to compare the estimated thresholds with the thresholds reported in previous studies exploring different contexts.

Comparison of VMAF Thresholds for Acceptability&Annoyance in different contexts

Table 1 presents the estimated thresholds in terms of VMAF [8] score for acceptability and annoyance of video quality. All thresholds are determined based on the algorithm described above. The thresholds for the online social media platform context is calculated over IPI-VUGC dataset while the thresholds for contexts of basic and premium subscription to video streaming services are calculated in [4].

As table 1 presents, we see that the threshold for annoyance of the video quality in terms of VMAF score is at the highest value of 87 for the premium subscription to a video streaming service context. When a basic subscription is taken into account we see that subjects are more forgiving in terms of video quality required for not annoying content. Even as low as 67 VMAF score is found to be enough for a not annoying video quality in online social media platforms. We also see a similar trend in the acceptability thresholds where subjects in the premium subscription context are more demanding in terms of video quality. Surprisingly, a satisfying video quality in the contexts of the online social media platforms can be seen not acceptable in the context of premium subscription to video streaming services.

Conclusion and Discussion

We showed that the acceptability&annoyance defines the QoE as a combination of user expectation and video quality. We provided a brief introduction to the publicly available IPI-VUGC dataset, that contains video quality ratings and acceptability&annoyance labels in the online social media platforms context. We provided a detailed guideline to the methodology used for determining the acceptability and annoyance thresholds in terms of objective quality metric predictions. Moreover, we used VMAF predictions of video quality to estimate these thresholds and compared them with other datasets developed on video streaming service context.

VMAF score thresholds for acceptability and annoyance reveal that people are more forgiving of the lower video quality content in online social media platform context compared to the context of paid subscription to video streaming services. The VMAF range of 67 to 71 can be seen as not annoying for the context of online social media platforms while can be categorized as not acceptable in the context of using video streaming platforms on television with a premium subscription. This striking difference in thresholds reveal the impact of context on QoE.

The choice of VMAF for user generated videos is not ideal due to variance in source content quality. Despite this, we utilized VMAF to be able to compare with the results of previous studies. Therefore, the study could be further improved with other more appropriate video quality metrics such as UVQ [13]. As future work, the study can be extended with more modern coding algorithms and high dynamic range content. The results of this work may help online social media platforms and streaming service providers to optimize their video delivery services without sacrificing from users' satisfaction.

References

[1] P. Le Callet, S. Möller, A. Perkis, *et al.*, "Qualinet white paper on definitions of quality of experience," *European network on quality of experience in multimedia systems and services (COST Action IC 1003)*, vol. 3, no. 2012, 2012.

[2] M. Angela and H. Knoche, "Quality in context-an ecological approach to assessing qos for mobile tv," 01 2006.

[3] A. Oeldorf-Hirsch, J. Donner, and E. Cutrell, "How bad is good enough? exploring mobile video quality trade-offs for bandwidth-constrained consumers," 10 2012.

[4] J. Li, L. Krasula, Y. Baveye, Z. Li, and P. Le Callet, "Accann: A new subjective assessment methodology for measuring acceptability and annoyance of quality of experience," *IEEE Transactions on Multimedia*, vol. 21, no. 10, pp. 2589–2602, 2019.

[5] A. Ak, A. F. Perrin, D. Noyes, I. Katsavounidis, and P. Le Callet, "Video consumption in context: Influence of data plan consumption on qoe," in *Proceedings of the 2023 ACM International Conference on Interactive Media Experiences*, IMX '23, (New York, NY, USA), p. 320–324, Association for Computing Machinery, 2023.

[6] S. Jumisko-Pyykkö and M. M. Hannuksela, "Does context matter in quality evaluation of mobile television?," in *Proceedings of the 10th International Conference on Human Computer Interaction with Mobile Devices and Services*, MobileHCI '08, (New York, NY, USA), p. 63–72, Association for Computing Machinery, 2008.

[7] T. De Pessemier, K. De Moor, W. Joseph, L. De Marez, and L. Martens, "Quantifying subjective quality evaluations for mobile video watching in a semi-living lab context," *IEEE Transactions on Broadcasting*, vol. 58, no. 4, pp. 580–589, 2012.

[8] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, "Toward a practical perceptual video quality metric," *The Netflix Tech Blog*, 2016.
[9] ITU-T, "Reference software for itu-t h.264 advanced video coding." ITU-T H.264-2, 2016.
[10] ITU-R, "Methods for the subjective assessment of video quality, audio quality and audiovisual quality of internet video and distribution quality television in any environment." ITU-R Recommendation Recommendation P.913, 2021.
[11] J. Zhu, A. Ak, P. Le Callet, S. Sethuraman, and K. Rahul, "Zrec: Robust recovery of mean and percentile opinion scores," in *2023 IEEE International Conference on Image Processing (ICIP)*, pp. 2630–2634, 2023.
[12] G. A. Barnard, "A new test for 2×2 tables," *Natur*, vol. 156, no. 3954, p. 177, 1945.
[13] Y. Wang, F. Yang, B. Adsumilli, N. Birkbeck, J. G. Yim, J. Ke, H. Talebi, P. Milanfar, R. Wolf, J. Jayaraman, C. Church, and J. Lin, "Uvq: Measuring youtube's perceptual video quality," *The Google Research Blog*, 2022.