

## Studying the impact of the content selection method on the video QoE on mobile devices

Nikolas Wehner, Nils Mertinat, Michael Seufert, Tobias Hoßfeld

### Angaben zur Veröffentlichung / Publication details:

Wehner, Nikolas, Nils Mertinat, Michael Seufert, and Tobias Hoßfeld. 2020. "Studying the impact of the content selection method on the video QoE on mobile devices." In Twelfth International Conference on Quality of Multimedia Experience (QoMEX), 26-28 May 2020, Athlone, Ireland, edited by Andrew Hines and Niall Murray, 1-4. Piscataway, NJ: IEEE.  
<https://doi.org/10.1109/qomex48832.2020.9123088>.

### Nutzungsbedingungen / Terms of use:

licgercopyright

Dieses Dokument wird unter folgenden Bedingungen zur Verfügung gestellt: / This document is made available under these conditions:

**Deutsches Urheberrecht**

Weitere Informationen finden Sie unter: / For more information see:

<https://www.uni-augsburg.de/de/organisation/bibliothek/publizieren-zitieren-archivieren/publiz/>



# Studying the Impact of the Content Selection Method on the Video QoE on Mobile Devices

Nikolas Wehner, Nils Mertinat, Michael Seufert, Tobias Hoßfeld  
University of Würzburg, Institute of Computer Science, Würzburg, Germany  
{nikolas.wehner | nils.mertinat | michael.seufert | tobias.hossfeld}@uni-wuerzburg.de

**Abstract**—When conducting video QoE studies participants are usually asked to rate the QoE of prepared test videos. However, participants are given no choice to select content, which they like or in which they are interested. This may cause annoyance or frustration when conducting the QoE study, which eventually might affect the QoE results of the study. The consequent question is whether the content liking has a direct impact on the submitted ratings by the participants and whether the freedom of choosing the video content in QoE studies results in better ratings. To investigate this research question, CroQoE, an existing framework for crowdsourced video testing, is extended and used in a pilot field study. In this work, the results of a QoE study with individual and dynamic content selection are compared to a QoE study with pre-selected contents. Moreover, this work includes a comparison to a previous QoE study for validation. As the previous study was conducted on desktop PCs, the CroQoE study further allows to identify differences in the stalling perception between studies on desktop PCs and mobile devices.

## I. INTRODUCTION

Quality of Experience (QoE) describes the perceived quality of an end user when utilizing any kind of service or application. It depends highly on the user, the system, the service, or the context [1]. Multiple QoE studies have been conducted on HTTP Adaptive Streaming (HAS), the nowadays dominant video streaming technology, in order to find out the most important QoE influence factors [2]. These studies have revealed that especially initial delay, stalling, and visual quality influence the QoE for HAS strongly. However, in each of the previous QoE studies, test persons were subject to videos, which the researchers selected beforehand. As the space of possible QoE influence factors is highly dimensional, it is not clear whether the method of video content selection has any impact on the QoE. Another unresolved aspect is the impact of the end device type, i.e., desktop and mobile, on the QoE. The ITU-T standards P.1203 and P.1204 [3], [4] also differ between the device types of desktop and mobile and the new P.1204 even considers tablet-type devices and TVs. However, the device types are only considered for the visual quality, but the interaction between end device type and stalling perception has not been investigated yet, although stalling is the major QoE degradation of HAS.

This work addresses these questions by conducting a pilot field study with CroQoE [5]. It is an existing app for crowd-

sourced QoE studies of HAS, which runs on mobile devices. Users can submit keywords to consume video contents fitting their interests. The selected video contents are dynamically prepared by inserting the test conditions into the crawled videos. Users can then download and watch the videos, and rate them with respect to content liking and subjective quality.

The results of a QoE study with individual and dynamic content selection are compared to a QoE study with pre-selected contents. Moreover, they are compared also to an existing study [6] for validation. As the validation study has been conducted in a crowdsourced fashion on desktop PCs, it is also possible to quantify the impact of the end device type on the QoE when comparing the results of the studies.

The paper is structured as follows: background and related work are presented in Section II. The CroQoE framework, the conducted field study, and the evaluation are described and discussed in Section III, and Section IV concludes.

## II. BACKGROUND AND RELATED WORK

In [2], a survey on the QoE influence factors of HAS is presented. The main findings are that stalling, i.e., the depletion of the video buffer, results in the strongest degradation of the QoE and subsequently has to be avoided at any cost. Multiple short stalling events also degrade the QoE stronger than one long stalling event. Further, a high quality adaptation frequency and amplitude should be avoided as it degrades the QoE independent of the switching direction. Metrics have also been introduced to estimate the visual quality, e.g., VMAF [7], [8], or the QoE as a whole, e.g., the ITU-T standards P.1203 and P.1204 [3], [4].

Focusing on stalling, the authors of [9] reveal that longer stalling events degrade the quality. Furthermore, they state that the position of a stalling event has no importance. However, the authors of [10] contradict this last finding by stating that there is an impact of the last stalling position. Both stalling and frame rate reduction are investigated in [11] where the authors state that frame rate reduction is preferred to stalling. As a second finding, they show that periodic stalling patterns are preferred to irregular stalling patterns. The authors of [12] compare stalling to quantization by using a random neural network model that is able to estimate QoE based on stalling and quantization. With subjective studies, they find out that users perceive stalling events stronger than variations in the quantization parameter.

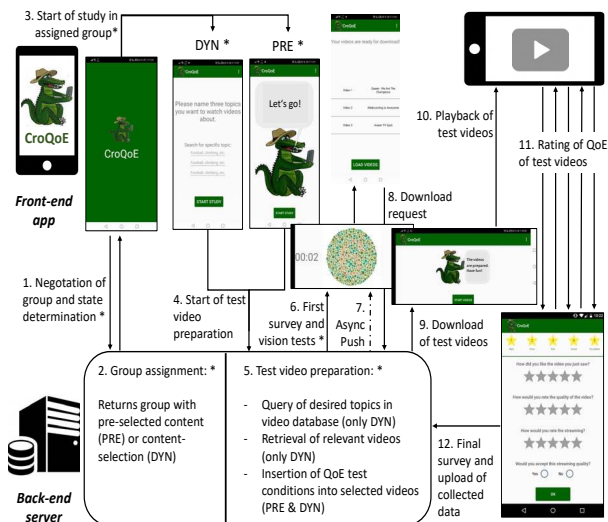


Fig. 1: Extended workflow of CroQoE

The authors of [6] conducted crowdsourcing studies in the desktop environment to assess the impact of stalling and model YouTube QoE. They observed that stalling is a main influence factor and showed that the fitted Mean Opinion Score (MOS) follows an exponential function for a varying number of stalling events, which confirmed the IQX hypothesis presented in [13]. In this work, their results are compared to the collected ratings of stalling on mobile devices.

When it comes to QoE on mobile devices there are only few apps. Most of them do not consider subjective feedback by their users and solely try to estimate the QoE based on monitored parameters, e.g., [14]. YoMoApp [15] can be considered as an exception because it additionally asks for subjective feedback after a user ended a video. But also YoMoApp is not able to conduct controlled QoE studies. Thus, the CroQoE app is used in this work.

### III. FIELD STUDY WITH CROQOE

#### A. CroQoE

The front-end of CroQoE connects via REST API to the back-end server and is implemented as Android application. The QoE study is presented in the front-end app. The video content preparation and the storage of the data are performed on the back-end server. The extended workflow of CroQoE is shown in Figure 1. All added or modified components compared to [5] are tagged with an asterisk.

To allow comparisons between the QoE of dynamically selected contents and pre-selected contents, CroQoE is extended for an additional group where the participants are shown pre-selected videos. The back-end decides to which group a participant belongs when the front-end is started (step 1 + 2). At startup, users are thus either assigned to the group with dynamically selected contents (DYN) or to the group with pre-selected contents (PRE) and are then shown the matching start screen (step 3). While users of DYN can then submit keywords for video contents of their interests, users of PRE

just continue with a button. Afterwards, the back-end starts to prepare the test videos (step 4).

For DYN the back-end first crawls a major video streaming provider for matching videos based on the submitted user's keywords. A video matches when it fits specific guidelines, i.e., the top five short HD videos which are sorted by view count. By selecting a random video ID from the returned list repetition of the same video is avoided. The selected videos are then downloaded by the back-end and the test conditions are dynamically inserted with FFmpeg for both groups. First, the video is cut to a specified length. Then, initial delay, stalling events, or visual quality changes can be added to the video (step 5). However, these processing tasks can take a long time, especially, when processing multiple videos.

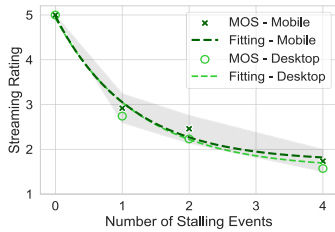
As shown in [16], long waiting times during a study can result in an annoyance of the participants which can directly influence the participant's QoE. To avoid any negative bias, the waiting times (for this study up to five minutes) are bridged with surveys and vision tests in the front-end application. As soon as the preparations in the back-end are finished, CroQoE utilizes Google's Firebase Cloud Messaging (FCM) to notify the user's device that the server has finished its task (step 7).

The front-end app can then download the videos from the back-end (step 8 + 9) after which the video playout can be performed (step 10). Users are not able to control the media during playback and the videos are played out full-screen in landscape orientation. When a video ends, the user has to submit ratings on visual quality, streaming quality, and content liking on the ACR scale and on streaming acceptance on a binary scale (step 11). Finally, the collected data are uploaded to the back-end. (step 12).

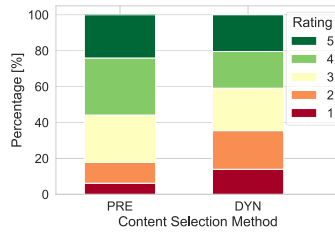
#### B. Pilot field study

The field study took place over three days in the beginning of January 2020 in which 150 people in total utilized the app on a University campus. As each user watched three videos, the obtained dataset consists of 450 videos in total. The videos were watched either on a Google Pixel 3a, a Google Pixel 2 XL, or a Google Pixel XL. The participants were mainly students and University employees with a mean age of 22.5 years.

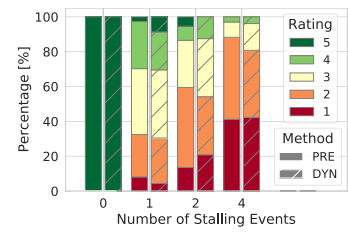
Each participant was assigned either to the group with pre-selected contents or to the group with dynamic content selection. All videos were downloaded in the best available quality to avoid any visual bias and were cut to an exact playout length of 30 seconds. The playback of the video started at 20% of the actual playback to avoid any introducing scenes, e.g., the studio names in a movie trailer. Either zero, one, two, or four stalling events were added to the videos in a regular pattern, where each stalling event had a length of four seconds which is in alignment to the baseline study [6]. Further, each of the three videos watched by a single participant showed a different regular stalling pattern. Initial delay and visual quality were not modified in this study. All stalling condition and group assignments were performed with a water filling algorithm to guarantee a balanced dataset.



(a) Exponential fitting of the datasets.



(b) Ratings for content liking.



(c) Streaming ratings by method.

Fig. 2: Evaluation results.

Based on the submitted ratings, the questionnaire results, and the vision test results, the dataset was filtered for outliers. Sessions where users did not pass the vision tests were excluded from the dataset as well as videos where users submitted contradictory ratings. After the filtering 324 videos remained, whereby 156 videos belonged to DYN and the remaining 168 videos to PRE. The stalling conditions were distributed almost evenly within PRE and DYN.

### C. Evaluation

1) *Impact of end device type on the QoE:* For the validation of CroQoE and the evaluation of the impact of the end device type on the QoE, the obtained ratings for the pre-selected contents of this study (mobile) are compared to the original results of the reference QoE study (desktop) [6].

The exponential fitting of the streaming ratings of both datasets can be seen in Fig. 2a. The 95% confidence interval of the CroQoE data is shown in form of the light grey band. The goodness of fit is determined by the coefficient of determination  $R^2$  which is around 0.994 for the CroQoE data. The figure shows that both fittings are quite similar and that they are for most of the time within the confidence band. This indicates no differences between the streaming ratings of both datasets when considering stalling events. The two-sided Mann-Whitney U test also rejects the hypothesis that there are any differences.

Further, the IQX hypothesis formulated in [13] is again confirmed by the CroQoE data and the ratings align well with the ratings from [6]. Thus, these results also indicate that stalling events do not impact the streaming perception differently on mobile and desktop devices.

2) *Impact of content selection method on the QoE:* The impact of the content selection method on the QoE is evaluated by comparing the content liking ratings and the streaming ratings of the content selection methods PRE and DYN.

The distributions of the content liking ratings for both content selection methods (cf. Figure 2b) reveal that the participants of PRE showed a higher satisfaction with their video content than participants of DYN. The mean content liking for PRE is 3.51 with a median of 4, while the mean for DYN is 2.92 with a median of 3. When performing a Mann-Whitney U test for the content liking, a statistically significant difference could be observed between the methods ( $p$ -value  $< 10^{-5}$ ).

However, this is contradictory to our original goal of providing users content which is fitting their interests. These differences in the content liking can be explained by the fact that participants often did not know what content they would like to watch right now and then ended up with submitting generic categories, e.g., food or climbing. One reason for this could be that they felt watched during the study or that they were simply startled so much that they really had no idea what they would like to watch at the moment. Since there are tons of videos for generic topics, these submissions resulted in highly diverse videos, which not always suited a participant's interest. Further, the randomized way in which the videos were selected, i.e., a random video of the top five videos ordered by view count, increased this effect.

Figure 2c depicts the distributions of the observed streaming ratings for both content selection methods depending on the observed number of stalling events side by side. No differences can be observed for zero stalling events and only minor differences can be observed for the other numbers. Again, the Mann-Whitney U test is performed for the streaming ratings of the methods to check whether there are any differences between the methods. A  $p$ -value of 0.343 could be observed and hence no differences between the streaming ratings of the content selection methods could be determined. This would reject the hypothesis that the used video content has any impact on the QoE.

## IV. CONCLUSION

This paper was a first step into investigating the impact of the content selection method on the QoE. For this purpose, a study on the impact of stalling on the QoE of video streaming was conducted, which did not show evidence that stalling events impact the streaming perception differently on mobile and desktop devices. A key aspect of the study was that a decrease of the content liking did not result in a lower QoE. However, to determine if an increase of the content liking impacts the streaming ratings, the dynamic content selection process has to be modified and extended. One option would be to guide the user through a content selection process, which is started based on the user submitted keywords. By giving the user a choice to select from a list of matching videos the content liking might be increased. With an increased content liking, it is then possible to further evaluate the impact of the content selection on the results of QoE studies.

## REFERENCES

- [1] P. Le Callet, S. Möller, and A. Perkis (eds), “Qualinet White Paper on Definitions of Quality of Experience,” European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003), Lausanne, Switzerland, Tech. Rep., 2013, version 1.2.
- [2] M. Seufert, S. Egger, M. Slanina, T. Zinner, T. Hoßfeld, and P. Tran-Gia, “A Survey on Quality of Experience of HTTP Adaptive Streaming,” *IEEE Communications Surveys & Tutorials*, vol. 17, no. 1, 2015.
- [3] International Telecommunication Union, “ITU-T Recommendation P.1203: Parametric Bitstream-based Quality Assessment of Progressive Download and Adaptive Audiovisual Streaming Services over Reliable Transport,” 2016. [Online]. Available: <https://www.itu.int/rec/T-REC-P.1203/en>
- [4] P.1204 : Video quality assessment of streaming services over reliable transport for resolutions up to 4k. [Online]. Available: <https://www.itu.int/rec/T-REC-P.1204-202001-I/en>
- [5] M. Seufert, N. Wehner, and P. Casas, “App for Dynamic Crowdsourced QoE Studies of HTTP Adaptive Streaming on Mobile Devices,” in *2018 Network Traffic Measurement and Analysis Conference (TMA)*. IEEE, 2018, pp. 1–2.
- [6] T. Hoßfeld, R. Schatz, M. Seufert, M. Hirth, T. Zinner, and P. Tran-Gia, “Quantification of YouTube QoE via Crowdsourcing,” in *Proceedings of the International Workshop on Multimedia Quality of Experience - Modeling, Evaluation, and Directions (MQoE)*, Dana Point, CA, USA, 2011.
- [7] R. Rassool, “Vmaf reproducibility: Validating a perceptual practical video quality metric,” in *2017 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*. IEEE, 2017, pp. 1–2.
- [8] Z. Li, C. Bampis, J. Novak, A. Aaron, K. Swanson, A. Moorthy, and J. Cock, “Vmaf: The journey continues,” *Netflix Technology Blog*, 2018.
- [9] Y. Qi and M. Dai, “The Effect of Frame Freezing and Frame Skipping on Video Quality,” in *Proceedings of the 2nd International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP)*, Pasadena, CA, USA, 2006.
- [10] T. N. Minhas and M. Fiedler, “Impact of Disturbance Locations on Video Quality of Experience,” in *Proceedings of the 2nd Workshop of Quality of Experience for Multimedia Content Sharing (QoEMCS)*, Lisbon, Portugal, 2011.
- [11] Q. Huynh-Thu and M. Ghanbari, “Temporal Aspect of Perceived Quality in Mobile Video Broadcasting,” *IEEE Transactions on Broadcasting*, vol. 54, no. 3, pp. 641–651, 2008.
- [12] K. D. Singh, Y. Hadjadj-Aoul, and G. Rubino, “Quality of Experience Estimation for Adaptive HTTP/TCP Video Streaming Using H.264/AVC,” in *Proceedings of the IEEE Consumer Communications and Networking Conference (CCNC)*, Las Vegas, NV, USA, 2012.
- [13] M. Fiedler, T. Hossfeld, and P. Tran-Gia, “A Generic Quantitative Relationship Between Quality of Experience and Quality of Service,” *IEEE Network*, vol. 24, no. 2, pp. 36–41, 2010.
- [14] Q. A. Chen, H. Luo, S. Rosen, Z. M. Mao, K. Iyer, J. Hui, K. Sontineni, and K. Lau, “QoE Doctor: Diagnosing Mobile App QoE with Automated UI Control and Cross-layer Analysis,” in *Proceedings of the Internet Measurement Conference (IMC)*, Melbourne, Australia, 2014.
- [15] M. Seufert, N. Wehner, F. Wamser, P. Casas, A. D’Alconzo, and P. Tran-Gia, “Unsupervised QoE Field Study for Mobile YouTube Video Streaming with YoMoApp,” in *Proceedings of the 9th International Conference on Quality of Multimedia Experience (QoMEX)*, Erfurt, Germany, 2017.
- [16] D. Strohmeier, S. Jumisko-Pyykkö, and A. Raake, “Toward task-dependent evaluation of web-qoe: Free exploration vs. ?who ate what??” in *2012 IEEE Globecom Workshops*, 2012, pp. 1309–1313.