Impact of test condition selection in adaptive crowdsourcing studies on subjective quality

Michael Seufert, Ondrej Zach, Tobias Hoßfeld, Martin Slanina, Phuoc Tran-Gia

Angaben zur Veröffentlichung / Publication details:

Seufert, Michael, Ondrej Zach, Tobias Hoßfeld, Martin Slanina, and Phuoc Tran-Gia. 2016. "Impact of test condition selection in adaptive crowdsourcing studies on subjective quality." In *Eighth International Conference on Quality of Multimedia Experience (QoMEX), 6-8 June 2016, Lisbon, Portugal*, edited by Karel Fliegel, 1–6. Piscataway, NJ: IEEE. https://doi.org/10.1109/qomex.2016.7498939.

Impact of Test Condition Selection in Adaptive Crowdsourcing Studies on Subjective Quality

Michael Seufert*, Ondrej Zach†, Tobias Hoßfeld‡, Martin Slanina†, Phuoc Tran-Gia*

*University of Würzburg
Institute of Computer Science
Würzburg, Germany
Email: seufert@informatik.uni-wuerzburg.de

†Brno University of Technology
Department of Radio Electronics
Brno, Czech Republic
Email: ondrej.zach@phd.feec.vutbr.cz

‡University of Duisburg-Essen
Chair of Modeling of Adaptive Systems
Essen, Germany
Email: tobias.hossfeld@uni-due.de

***Abstract*—Adaptive crowdsourcing is a new approach to crowdsourced Quality of Experience (QoE) studies, which aims to improve the certainty of resulting QoE models by adaptively distributing a fixed budget of user ratings to the test conditions. The main idea of the adaptation is to dynamically allocate the next rating to a condition, for which the submitted ratings so far show a low certainty. This paper investigates the effects of statistical adaptation on the distribution of ratings and the goodness of the resulting QoE models. Thereby, it gives methodological advice how to select test conditions for future crowdsourced QoE studies.**

I. Introduction

Quality of Experience (QoE), i.e., the subjective perception of the quality of a service as a whole, is becoming increasingly important for network and service providers. Both want to deliver the service to the end user in the best manner to achieve a high customer satisfaction. In order to understand what factors influence the QoE, extensive subjective studies have to be conducted.

Recently, crowdsourcing, i.e., the outsourcing of small tasks to a large crowd, has been widely used for subjective QoE assessments (e.g., [1], [2]). The advantages of crowdsourcing over classical laboratory studies are its price, speed, and the more realistic setting of service consumption (especially, context and system factors). Nevertheless, a crowdsourcing QoE study has to be well designed to avoid typical pitfalls from the heterogeneous environment and unsupervised nature of such studies [3]. The researcher wants to investigate the impact of a parameter x on QoE (in terms of Mean Opinion Score (MOS)), thus, a QoE model $f(x)$ is to be derived, which returns the MOS for any (realistic) $x \in \mathbb{R}$. Typically, a fixed rating budget (i.e., total number of user ratings) is given, which determines the costs for the study. The researcher selects discrete test conditions (TC), which are distributed over the tested range, and the test participants see a pre-allocated or random TC when they access the study. This results in (approximately) the same number of ratings per TC. Eventually, the MOS of each TC will be fitted to obtain $f(x)$.

The new concept of adaptive crowdsourcing challenges this traditional approach, and aims at adaptively distributing users to TCs at the moment participants access the study. The main idea of the adaptation is to allocate the next rating to a TC, for which the submitted ratings so far show a low certainty, e.g., in terms of confidence intervals (CI) of the MOS. This adaptive distribution of the rating budget is expected to increase the overall certainty of the QoE model. As indicated in [4], very high and very low quality TCs are rated more homogeneously because ratings concentrate close to the edge of the rating scale. The CIs of the MOS become small after few ratings, and gathering more ratings for these TCs will only bring a negligible gain. In contrast, TCs with a medium quality will foster a high diversity of subjective ratings, and thus, a lower certainty of QoE model $f(x)$. Shifting rating budget from the extreme quality TCs is expected to increase the certainty for these TCs.

The goal of this work is to introduce adaptive crowdsourcing and to investigate different TC selection strategies for both traditional and adaptive study design. It will examine the impact of expert knowledge on the selection of TCs and demonstrate the effects of statistical adaptation. A crowdsourcing study on the impact of encoding bitrate on the QoE of H.264 videos was conducted to obtain a ground truth pool of ratings. Drawing from this pool, we simulate different traditional and adaptive TC selection strategies, analyze the effects, and investigate the goodness of the resulting QoE models. Finally, we conducted live crowdsourcing studies to show the behavior of the strategies in a realistic setting.

Therefore, this work is structured as follows. Section II presents related work and Section III describes the methodology of our study, the investigated strategies, and the simulation framework. Section IV presents the effects of the different strategies and the results of the live crowdsourcing study, and Section V concludes.

II. Related Work

Crowdsourcing is a widely adopted methodology for subjective quality assessment and was used among different application domains, such as video (e.g., [5], [6]) and image quality (e.g., [2], [7]), and QoE of video streaming (e.g., [1], [8]) or other web services (e.g., [9], [10]).

From a methodological point of view, works have been conducted with respect to motivation and incentives (e.g., [10], [11]), reliability methods and screening mechanisms (e.g., [12]), result quality monitoring (e.g., [13]), and the development of crowdsourcing frameworks and platforms (e.g., [14]). A comprehensive report of best practices and lessons learned for crowdsourced QoE studies is given in [3].

In this paper, we propose the idea of adaptive test designs in crowdsourcing studies. However, it has not been researched how to select TCs in such an adaptive crowdsourcing study in

order to obtain best results. Therefore, this work will investigate different strategies for TC selection using the example of a video quality study, namely, investigating the impact of encoding H.264 videos to different bitrates on the QoE of video streaming. This problem has already been studied, e.g., in [15] and [16]. The authors of these studies offer a objective assessment of H.264 coded sequences with different encoding parameters. A study of perceived video quality is presented in [17]. This study uses lower bitrates as it is focused mainly on QoE when encoding sequences with low resolution. Thus, we want to point out that the focus of our work is not on providing quantitative results on QoE, but on investigating different general methodologies for adaptive crowdsourcing, which can be applied to obtain such results.

III. Study Description

In this section, the used methodology is described. First, the investigated TC selection strategies are presented. Second, the focus is on the crowdsourcing study, which was conducted to obtain ground truth ratings. Finally, the simulation framework is described.

A. Test Condition Selection Strategies

The first option for test designers is to determine a parameter and select between discrete and continuous TCs. Note that some parameters might not allow for continuous TCs due to their character, technical limitations, or study characteristics. For both cases, the following three classes of TC selection strategies can be applied:

a) Baseline: A baseline strategy tries to allocate the same number of users per condition in the discrete case, or to fully cover the parameter range in the continuous case. Several variants of baseline strategies can be implemented given a fixed rating budget. In the discrete case, these include an a-priori-selection of TCs per user (distribute the users equally to TCs before the test starts), a waterfilling-based selection (always select condition with fewest ratings), or a randomized selection (select a TC randomly according to a uniform distribution). In the continuous case, the parameter range could be equally spaced (discretized) in subranges according to the budget and the above mentioned algorithms can be applied, or TCs could be selected using a uniform distribution over the whole parameter range or over the largest subrange without ratings.

b) Statistical Adaptation: Statistical adaptation refers to an adaptive test design, in which TCs are selected based on a statistic computed from already given ratings. The goal is to allocate more ratings to TCs with high uncertainty of measurement. Typically, a baseline strategy is used beforehand to avoid the cold start problem, i.e., to obtain a minimum number of ratings before the adaptive strategy is applied. In the discrete case, for example, the TC with the largest CI or the highest variance can be selected as the next TC to be tested. Having a continuous parameter, the parameter range can be split into subranges, and the next TC(s) can be selected from the subrange with the largest CI or variance. Additionally, the TCs/subranges can be excluded from the selection process when the statistic indicates a sufficient level of certainty with the measurement.

c) Expert Knowledge Selection/Adaptation: Expert knowledge can be applied when any knowledge about the investigated parameter is already available. It can be exploited for both discrete and continuous scenarios, but its implementation depends largely on the specific parameter and available knowledge. If a functional relationship about the impact of the parameter is known and a study is conducted to learn the exact parameters of the relationship, the characteristics of the function can be used for the selection of TCs. For example, studying an exponential relationship of a parameter, it is more beneficial to allocate (more) TCs where the slope is expected than in the range of the (almost) constant tail. The expert knowledge is typically used in combination with a baseline or a statistical adaptation strategy.

For the remainder of this work, we consider the video encoding bitrate parameter over the TC range 500 kbps to 3000 kbps and have selected the following strategies in the discrete (D) and continuous (C) case, respectively. The notation also indicates whether a baseline (B) strategy or statistical adaptation (S) is used, respectively.

In the discrete case, we limit ourselves to a crowdsourcing study with five TCs. The following algorithms are considered:

1) Fixed number of ratings per TC (D-B)
2) Statistical adaptation based on CI width (D-S)

For the baseline algorithm (D-B), the rating budget is divided by the number of available TCs and each TC is selected equally often.

The statistical adaptation (D-S) aims at minimizing the CI width. It comes into effect after each TC was rated five times, and subsequently selects the TC with the highest CI width.

The continuous crowdsourcing studies are based on uniformly distributing the TCs over the investigated parameter range:

1) Uniform distribution over investigated range (C-B)
2) Statistical adaptation of distribution subrange based on CI width (C-S)

The baseline algorithm (C-B) selects the next TCs based on a uniform distribution on the range of investigated TCs.

We use a conceptual algorithm for the statistical adaptation in the continuous case (C-S) aiming to split the whole TC range into subranges and to minimize CI widths of subranges. Initially, 10 ratings each are allocated to both edge TCs (i.e., 500kbps and 3000kbps) to anchor the tested range, and 10 ratings each are distributed over both halves of the TC range. Then, the range is divided into these four subranges, and a statistic of each subrange is computed. The algorithm selects the subrange with the largest statistic and the next four ratings are uniformly distributed over this subrange. The statistics are recomputed and the subrange with the largest statistic is selected next. When a subrange is selected for the second time, this subrange is halved and 4 ratings each are distributed over both new subranges before recomputing the statistics. The algorithms repeats the described process until the rating budget is consumed. The statistic used is not the classical CI width, but the standard deviation divided by the square of the number of ratings, which proved to promote the splitting. Note that the

exemplary algorithm described here might be easily improved, which is out of scope of this work[1].

B. Crowdsourcing Study

We conducted a crowdsourcing study, in which the users had to rate the quality of H.264 video sequences having different bitrates. Five source sequences were used, which cover wide variety of characteristics. All source video sequences were available in 1080p resolution at 25 frames per second, the length of the sequences was adjusted to 10 seconds. The source video sequences were downscaled using *ffmpeg*[2] tool to standard resolution (576p) to meet the possibly low Internet connections of the crowd workers and were encoded using the *x264*[3] implementation. All source video sequences were encoded to 51 quality levels. These levels were determined by bitrate and the selected values ranged from 500 to 3000 kbps in steps of 50 kbps. This range of bitrates is commonly used on the Internet for video streaming of Standard Definition video.

We used an online test framework similar to [1], which follows the best practices described in [3], and thus, includes monitoring of test execution and reliability checks. Every participant watched five videos with different content each, and rated the quality afterwards on a 5-point ACR scale. To avoid network influences during the playback (e.g., initial delay, stalling), all videos were downloaded to the local browser cache before the playback. Unreliable users were filtered out according to the clicking behavior during a pre-test, which indicated if the user read the instructions or not. Moreover, ratings of users who did not watch all videos in their full lengths or answered simple content questions incorrectly were discarded. Finally, ratings of users with different answers to personal questions, which were presented twice at the beginning and at the end of the test, were also eliminated.

The study was available as a micro job on the crowdsourcing platform Microworkers[4]. Every user could participate and was rewarded with 0.30$ upon completion of the test. Altogether, the test used in this study was completed 2047 times by 789 unique users from 80 different countries. Following the strict consistency checks of the framework, we observe a consistency rate of 48.12%, which means that slightly less than half of the participants conducted the test properly.

We monitored that some users experienced stops during playback of the sequences due to playback problems in the browser. We conducted a t-test confirming that there is a significant difference between the ratings for disturbed sequences (sequences with stops during playout) and undisturbed sequences (p value in the order of 10^{-12}). Therefore, we also discarded those ratings from the data set. After filtering out unreliable users and ratings from disturbed sequences, we had together 2817 scores from 563 unique users.

C. Simulation Framework

The simulation framework uses all reliable ratings obtained by the crowdsourcing study for one of the tested videos (i.e., all ratings belong to the same video content). These are 638 ratings gathered from 476 unique users, which gives us approximately 12 ratings per processed video sequence on average. The restriction to one video content can be done in this case as the focus of our work is on the adaptive crowdsourcing strategies and the goodness of the resulting models, and not on the actual characteristics of the model or the influence on the content. Nevertheless, similar results with respect to the investigated crowdsourcing strategies could be obtained for all contents. Note that we also do not consider unreliable ratings in the course of our study, but leave this interesting aspect for future work. In this case, a differentiation would be needed based on when unreliable ratings can be identified by the system (in momento or a posteriori, cf. [12]).

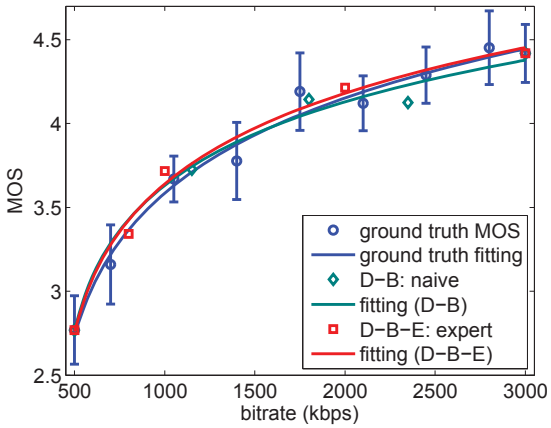


Fig. 1: Ground truth model, and models for discrete baseline strategies (D-B).

For each TC (i.e., for each bitrate), a pool is formed, which contains all ratings given by the study participants for that TC. Adaptive crowdsourcing studies can be simulated for each of the investigated algorithms (described in Section III-A) with respect to a given rating budget n_{max}, i.e., the number of ratings for the crowdsourcing study. Until the n_{max} ratings are obtained, a two step process is repeated. First, the next TC is selected according to the investigated algorithm. Then, the simulation framework uses the empirical distribution to draw user ratings for that TC (drawing with replacement) and adds it to the set of ratings for that run. The full set of ratings (containing n_{max} ratings) is considered the outcome of one crowdsourcing study (i.e., one simulation run).

IV. Results

We will investigate the impact of different bitrates on the QoE of H.264 videos in both the discrete case with five TCs, and the continuous case. Note that, although it would be technically feasible to have a continuous test design for this quality study, the continuous case is approximated by 51 TCs ($\{500, 550, \ldots, 3000\}$ kbps).

A. Impact of Expert Knowledge on Test Condition Selection

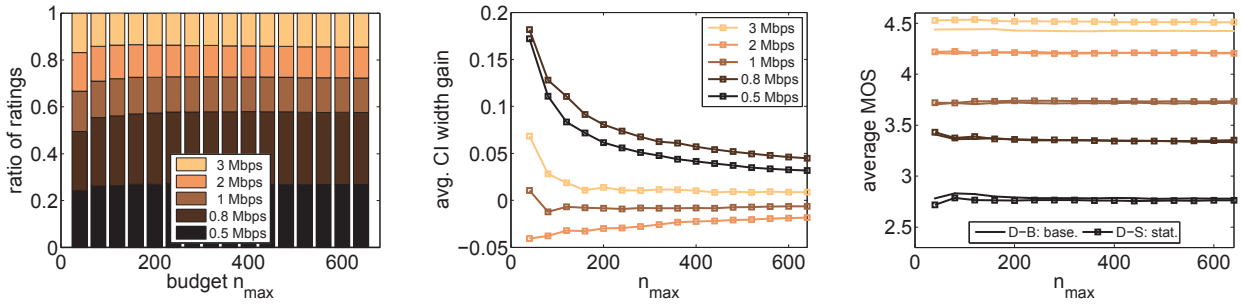
Figure 1 shows the impact of TC distribution of a naive baseline approach and a baseline approach with expert knowledge on the fitted model. The figure presents the ground truth model (blue), which is based on all ratings gathered in the crowdsourcing study aggregated per TC. To obtain the blue

[1]In future work, we will consider a study to find better parameters and statistics for the C-S strategy.

[2]http://www.ffmpeg.org

[3]http://www.videolan.org/developers/x264.html

[4]https://microworkers.com



(a) Average ratio of ratings per TC for statistical adaptation strategy.

(b) Gain of D-S in terms of confidence interval width per TC.

(c) Comparison of ground truth MOS to MOS obtained by both strategies.

Fig. 2: Comparison of baseline (D-B) and statistical adaptation (D-S) strategy for discrete TCs

curve, the mean opinion scores (MOS) for all 51 TCs were fitted by a logarithmic function according to the method of least squares. Grouping the TCs in regular intervals, it can be seen that the blue curve lies well within the confidence intervals (CI) for the MOS of the groups (blue markers). We consider this function as the ground truth model, i.e., the target function of the researcher $f(x)$, throughout the remainder of this work.

The green line shows the naive baseline model (D-B) obtained by fitting all ratings from five regular TCs (green markers). For the D-B-E model, the fitting was conducted based on all ratings of different TCs (red markers), which were distributed based on the expert knowledge that the model is a logarithmic function. It can be seen that, for the given example, both the green and the red line resemble the blue ground truth function and lie within the CIs. Thus, the different distribution of the three inner fitting TCs did not have a clear effect. Also looking at other fittings out of our ground truth, we see no benefit of the expert knowledge. For the remainder of this work, we will nevertheless use the "expert" TCs for D-B and D-S as they show nicely distributed MOS values.

B. *Effects of Statistical Adaptation*

To assess the effects of statistical adaptation of TC selection, we will first focus on the discrete case. Figure 2a shows a stacked bar chart of the average distribution of selected TCs for different rating budget resulting from the D-S algorithm. It can be seen that the D-S algorithm allocates more ratings to the low bitrate TCs, which have a lower MOS around 3. For the higher bitrate TCs, the pool of ratings contains more consistent ratings, which results in low confidence intervals already after few ratings, such that the TC is not selected again. These results are in line with findings from [4].

Figure 2b presents the gain of statistical adaptation in terms of CI width for different rating budgets. Therefore, the differences of the CI widths between D-B and D-S strategies are computed over 50 simulation runs per rating budget. The plot shows the average differences for the five TCs. It can be seen that the differences are positive for the low bitrate TCs, which indicates that the CIs are smaller for D-S. Thus, we can see that the statistical adaptation causes the desired effect that CI widths are decreased as more ratings are given to these TCs (cf. Fig. 2a). Obviously, the gain becomes smaller when the rating budget increases, because more ratings are allocated to each TC, which leads to generally smaller CIs. While the CIs of the 1000kbps and 3000kbps TCs are little affected, the CIs of the 2000kbps TCs become slightly larger for D-S. This does not come as a surprise, as less ratings are allocated to this TC compared to D-B, especially if other TCs face a higher rating diversity. Still, the increase of CI width is rather small, because the ratings for this good quality TC are quite homogeneous.

Looking at the resulting MOS for the different TCs in Figure 2c, there is a small catch. The figure compares the MOS obtained by the D-B (plain line) and D-S (line with boxes) strategies to the ground truth, which is the right-most point of the D-B line. While D-B and D-S results completely conform to the ground truth for the 500kbps to 2000kbps TCs, the MOS for 3000kbps shows a different behavior. It can be seen that D-S always leads to a higher MOS than D-B and ground truth for this TC. The reason for this effect could be the high quality of this TC and the decreasing variance of ratings close to the edge of the rating scale (cf. SOS hypothesis [4]). This setting facilitates small CIs around high MOS values after few ratings, which will prevent D-S from allocating more ratings to that TC. This means, the D-S algorithm is likely to receive several "Excellent" (5) ratings for a high quality TC, which quickly shorten the CI around a too high MOS. As the CI widths of other TCs decrease slower because of a higher variance of ratings, other TCs are preferred, such that this too high MOS can hardly be rectified.

To get comparable results, in the continuous case, the CIs are computed per subrange, i.e., the x-axis is split in nine parts of 450kbps width and the CI is computed over all ratings within a subrange. Figure 3 shows the average CI width over all TCs (discrete)/subregions (continuous) and over 50 simulation runs per rating budget. The x-axis depicts the rating budget. It can again be seen that increasing the rating budget leads to smaller average CI widths due to the higher number of ratings per TC/subregion. Due to the same reason, no algorithm can clearly outperform the others if the rating budget is high. However, for small rating budgets, it can be seen that the statistical adaptation algorithms D-S (green) and C-S (red) work nicely and reach smaller CI on average than their baseline counterparts D-B (blue) and C-B (cyan). Thus, in terms of CI widths, the gain of statistical adaptation appears only for low budgets.

Studying the impact of the statistical adaptation on the fitting, we compare the ground truth model to the models obtained by the simulation runs. Figure 4 shows the different rating budgets n_{max} on the x-axis, and the respective average

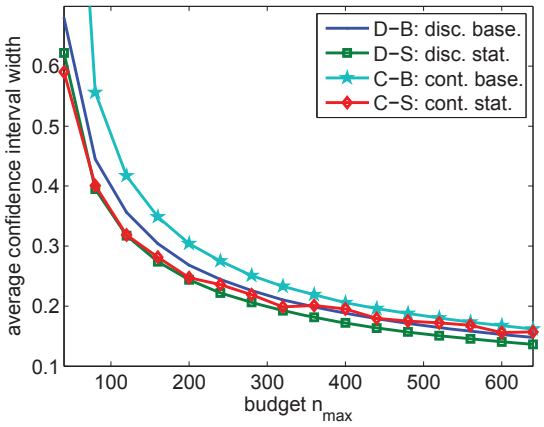


Fig. 3: Average confidence interval width over all TCs (discrete)/subregions (continous).

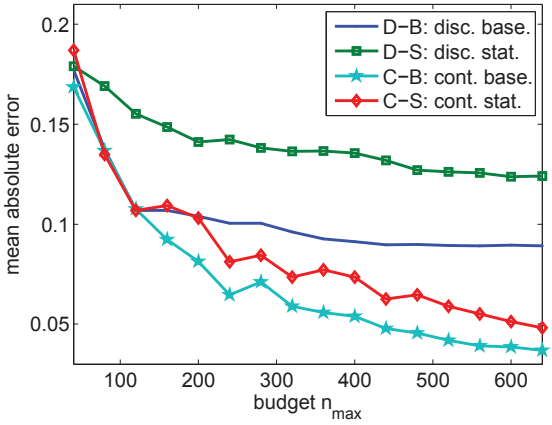


Fig. 4: Average mean absolute error of fitted models compared to ground truth model.

goodness of fit over 50 simulation runs on the y-axis. The mean absolute error is the goodness of fit metric used to compare the D-B (blue), D-S (green), C-B (cyan), and C-S (red) models to the ground truth model. The metric indicates the average of absolute errors at all 51 discretized TCs. It can be seen that an increasing rating budget improves the fitted model as the presented error-based goodness of fit metric becomes smaller. In general, it can be observed that all models are very close as the numeric values of the mean absolute error are very small (below 0.2 on MOS scale). However, two observations are noteworthy. Due to the effect that good quality TCs at the edge of the rating scale are likely to be insufficiently sampled (cf. Fig. 2c), we see that the fitting caused by the statistical adaptation is slightly worse than the baseline TCs. Moreover, the continuous case outperforms the discrete case and reaches better fittings. The reason is that the fitting can be based on a higher number of data points (one per TC), which can compensate inaccuracies better than if it was based on only a few data points. Thus, a continuous test design should be preferred when applicable for the investigated parameter.

C. Performance of Statistical Adaptation in a Real Crowdsourcing Study

To demonstrate the behavior of the investigated algorithms in a realistic setting, live crowdsourcing experiments were conducted with 100 reliable user ratings each. As the rating budget is so small, the parameters of C-S had to be slightly altered, such that only 12 ratings are initially distributed (instead of 40), and the subranges are divided into three new subranges (instead of two).

Figure 5a shows the evolution of the rating budget distribution for the D-S algorithm. The figure shows the ratio of ratings per TC as stacked bar plot after a certain amount of used budget, which is depicted on the x-axis. Starting from a uniform distribution after 20 ratings, the adaptation tends to prefer 500kbps and 800kbps TCs. After the whole rating budget was distributed, it can be seen that those TCs receive a significant share of ratings. In contrast, the more homogeneous ratings for the higher quality TCs result in a smaller ratio of ratings for these TCs.

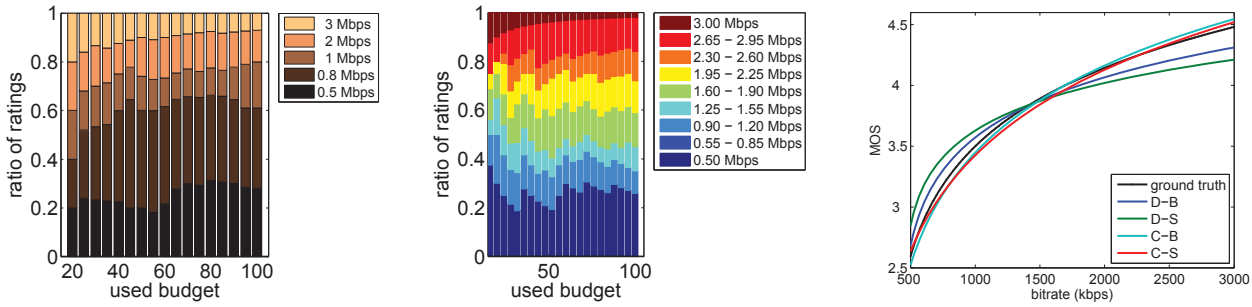
Figure 5b shows the corresponding plot for the C-S algorithm. Therefore, the edge TCs are separated, and the remaining range is divided into seven subranges of 350kbps containing seven discretized TCs each. It can be seen that the edge TCs have a high ratio for low budget due to the initial phase of the C-S algorithm. Moreover, it is evident that by chance the 500kbps TC has a very high rating diversity and was selected frequently. Apart from that, the C-S adaptation distributes the ratings almost uniformly over the whole parameter range, which means that, in this study, at any time the computed statistic is surprisingly similar for all subranges.

The resulting QoE models are presented in Figure 5c. The black ground truth shows the optimal outcome of the crowdsourcing study. It can be seen that both discrete strategies D-B (blue) and D-S (green) result in a less accurate model, which intersects the ground truth model at around 1500kbps and shows a higher MOS for lower bitrate TCs, and a lower MOS for the higher bitrate TCs, respectively. Although both models are less accurate, it can be seen that statistical adaptation (D-S) outputs a worse model compared to the baseline strategy D-B. In the continuous case, the QoE models of both strategies C-B (cyan) and C-S (red) well resemble the ground truth. A negative impact of the statistical adaptation is not visible here.

V. Conclusion

In this work, the effects of statistical adaptation in crowdsourced QoE studies were investigated. Therefore, a ground truth of subjective ratings was gathered via a crowdsourcing study about the impact of encoding bitrate on the QoE of H.264 videos. Four TC selection strategies were discussed and evaluated by simulating crowdsourcing studies based on the ground truth pool. Finally, the investigated strategies were implemented in a real crowdsourcing study.

In a crowdsourcing study with discrete TCs, no effect of the distribution of the inner TCs within the investigated parameter range was visible. Both a regular distribution and the distribution based on expert knowledge about the QoE model resulted a similar quality of resulting models. The statistical adaptation, aiming to minimize the CI for the MOS, showed the expected behavior. The average CI width could be minimized by D-S especially for low rating budget. Nevertheless, it is possible that for some TCs the resulting CI will be slightly larger than for D-B, because other conditions with higher rating diversity can be preferred by D-S. In terms of quality of the



(a) Ratio of ratings per TC for D-S. (b) Ratio of ratings per subrange for C-S. (c) Resulting QoE models.

Fig. 5: Results of live crowdsourcing studies of all investigated strategies.

resulting QoE models, the interesting effect was apparent that high quality conditions close to the edge of the rating scale can output a too high MOS when using statistical adaptation (D-S). This is due to the homogeneity of the ratings, which is likely to quickly decrease the CIs, and thus, prevent to allocate more ratings to that condition. The results also show that crowdsourcing studies with continuous TCs outperform the discrete case and reach better fittings, because the fittings can be based on more data points, such that inaccuracies can be better compensated.

In the live crowdsourcing experiment, all algorithms were tested in a realistic setting. It could be seen that the adaptation of D-S works as expected. In contrast, the adaptation algorithm for the continuous case (C-S) needs to be refined in the future. Especially the QoE models resulting from the continuous strategies well resembled the ground truth model. In the discrete case, the resulting models showed some inaccuracies and did not exactly overlap. Thus, researchers have to be aware that a crowdsourcing study is a single random experiment, and, especially for low budgets, less accurate QoE models might be the outcome. The best way to reduce the probability of this undesired outcome is to increase the rating budget for the crowdsourcing study. In the future, this work will be continued by considering unreliable ratings and allocation of multiple TCs per user (e.g., different contents) in the simulation framework, and a mathematical approach to study the effects of adaptive crowdsourcing will be taken.

Acknowledgment

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) under grants HO4770/1-2 and TR257/31-2 (OekoNet) as well as grants HO4770/2-1 and TR257/38-1 (Crowdsourcing). Research described in this paper was also financed by the National Sustainability Program under grant LO1401. For the research, infrastructure of the SIX Center was used. The authors alone are responsible for the content.

References

[1] T. Hoßfeld, R. Schatz, M. Seufert, M. Hirth, T. Zinner, and P. Tran-Gia, "Quantification of YouTube QoE via Crowdsourcing," in *Intl. Workshop on Multimedia Quality of Experience - Modeling, Evaluation, and Directions (MQoE)*, Dana Point, CA, USA, 2011.

[2] J. A. Redi, T. Hoßfeld, P. Korshunov, F. Mazza, I. Povoa, and C. Keimel, "Crowdsourcing-based Multimedia Subjective Evaluations: A Case Study on Image Recognizability and Aesthetic Appeal," in *2nd Intl. Workshop on Crowdsourcing for Multimedia (CrowdMM)*, Barcelona, Spain, 2013.

[3] T. Hoßfeld, M. Hirth, J. Redi, F. Mazza, P. Korshunov, B. Naderi, M. Seufert, B. Gardlo, S. Egger, and C. Keimel, "Best Practices and Recommendations for Crowdsourced QoE - Lessons learned from the Qualinet Task Force Crowdsourcing," 2014, European Network on Quality of Experience in Multimedia Systems and Services.

[4] T. Hoßfeld, R. Schatz, and S. Egger, "SOS: The MOS is not enough!" in *3rd Intl. Workshop on Quality of Multimedia Experience (QoMEX)*, Mechelen, Belgium, 2011.

[5] P. Hanhart, P. Korshunov, and T. Ebrahimi, "Crowd-based Quality Assessment of Multiview Video plus Depth Coding," in *Intl. Conference on Image Processing (ICIP)*, Paris, France, 2014.

[6] M. Rerabek, P. Hanhart, P. Korshunov, and T. Ebrahimi, "Quality Evaluation of HEVC and VP9 Video Compression in Real-time Applications," in *7th Intl. Workshop on Quality of Multimedia Experience (QoMEX)*, Costa Navarino, Greece, 2015.

[7] P. Hanhart, P. Korshunov, and T. Ebrahimi, "Crowdsourcing Evaluation of High Dynamic Range Image Compression," in *SPIE Applications of Digital Image Processing XXXVII*, 2014.

[8] M. Seufert, T. Hoßfeld, and C. Sieber, "Impact of Intermediate Layer on Quality of Experience of HTTP Adaptive Streaming," in *11th Intl. Conference on Network and Service Management (CNSM)*, Barcelona, Spain, 2015.

[9] P. Amrehn, K. Vandenbroucke, T. Hoßfeld, K. de Moor, M. Hirth, R. Schatz, and P. Casas, "Need for Speed? On Quality of Experience for File Storage Services," in *4th Intl. Workshop on Perceptual Quality of Systems (PQS)*, Vienna, Austria, 2013.

[10] M. Varela, T. Mäki, L. Skorin-Kapov, and T. Hoßfeld, "Increasing Payments in Crowdsourcing: Dont Look a Gift Horse in the Mouth," in *4th Intl. Workshop on Perceptual Quality of Systems (PQS)*, Vienna, Austria, 2013.

[11] A. Sackl, M. Seufert, and T. Hoßfeld, "Asking Costs Little? The Impact of Tasks in Video QoE Studies on User Behavior and User Ratings," in *4th Intl. Workshop on Perceptual Quality of Systems (PQS)*, Vienna, Austria, 2013.

[12] B. Gardlo, S. Egger, M. Seufert, and R. Schatz, "Crowdsourcing 2.0: Enhancing Execution Speed and Reliability of Web-based QoE Testing," in *Intl. Conference on Communications (ICC)*, Sydney, Australia, 2014.

[13] M. Hirth, S. Scheuring, T. Hoßfeld, C. Schwartz, and P. Tran-Gia, "Predicting Result Quality in Crowdsourcing Using Application Layer Monitoring," in *5th Intl. Conference on Communications and Electronics (ICCE)*, Da Nang, Vietnam, 2014.

[14] T. Hoßfeld, M. Hirth, P. Korshunov, P. Hanhart, B. Gardlo, C. Keimel, and C. Timmerer, "Survey of Web-based Crowdsourcing Frameworks for Subjective Quality Assessment," in *16th Intl. Workshop on Multimedia Signal Processing (MMSP)*, Jakarta, Indonesia, 2014.

[15] F. Speranza, A. Vincent, and R. Renaud, "Bit-Rate Efficiency of H.264 Encoders Measured With Subjective Assessment Techniques," *IEEE Transactions on Broadcasting*, vol. 55, no. 4, pp. 776–780, 2009.

[16] S. Pasqualini, F. Fioretti, A. Andreoli, and P. Pierleoni, "Comparison of H.264/AVC, H.264 with AIF, and AVS based on Different Video Quality Metrics," in *Intl. Conference on Telecommunications (ICT)*, 2009.

[17] M. Shahid, A. K. Singam, A. Rossholm, and B. Lövström, "Subjective Quality Assessment of H.264/AVC Encoded Low Resolution Videos," in *5th Intl. Congress on Image and Signal Processing (CISP)*, 2012.