Multimedia Tools and Applications

On Subjective Quality Assessment of Adaptive Video Streaming via Crowdsourcing and Laboratory Based Experiments

--Manuscript Draft--

Manuscript Number:	MTAP-D-15-01190R2
Full Title:	On Subjective Quality Assessment of Adaptive Video Streaming via Crowdsourcing and Laboratory Based Experiments
Article Type:	Manuscript
Keywords:	Adaptive Video Streaming; crowdsourcing; Subjective Quality Assessment; quality of experience
Corresponding Author:	Jacob Søgaard, Ph.D. Technical University of Denmark Kgs. Lyngby, DENMARK
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Technical University of Denmark
Corresponding Author's Secondary Institution:	
First Author:	Jacob Søgaard, Ph.D.
First Author Secondary Information:	
Order of Authors:	Jacob Søgaard, Ph.D.
	Muhammad Shahid, Ph.D.
	Jeevan Pokhrel, Ph.D.
	Kjell Brunnström, Professor
Order of Authors Secondary Information:	
Funding Information:	

Authors' Reply to Reviewers' Comments

On Subjective Quality Assessment of Adaptive Video Streaming via Crowdsourcing and Laboratory Based Experiments

by

Jacob Søgaard, Muhammad Shahid, Jeevan Pokhrel, Kjell Brunnström

We would like to express our sincere gratitude to the editor for handling our paper and for providing us the opportunity to revise the manuscript. We would also like to thank the reviewers for reviewing our paper along with the constructive comments which have been very helpful for improving the quality of our paper. We have revised the manuscript according to the reviewers' comments and the following pages present our point-wise response to the review.

Authors' Response to Reviewer 1

Comment 1: Please include a more clear explanation for the necessity of conducting experiment 2, or the objectives that you wanted to achieve with experiment 2, in the introduction of section 4 (something similar to the reply provided by the authors to my previous comments 58: can you clearly state what are the novel indications that you have obtained by conducting lab experiment 2 (in relation to the findings that you had already published)?

Authors' Response: We are highly thankful to the reviewer for this advice and we have updated the introduction of Section 4 accordingly. The related sentences read the following:

"Note that Laboratory Experiment 2 was mainly conducted to better understand any difference between the results of Lab Experiment 1 and the Crowdsourcing Experiment to confirm whether it is due to the difference in test material or due to the experimental methodology. Based on the findings in Lab Experiment 2 we were able to conclude that the difference in the results is most likely due to the PC methodology as discussed in Section 6."

Authors’ Response to Reviewer 5

Comment 1: Authors reworked their paper according reviewers recommendations and suggestions. They also answered reviewers questions. The structure of the paper is well composed. The Introduction and related works are followed by explanation of proposed solution and experimental results. The paper is well written and I recommend it to publish.

Authors’ Response: We again thank the reviewer for the efforts spent in reviewing our paper and are grateful that the work on our revision has been acknowledged.

Authors' Response to Reviewer 8

Comment 1: if existing tools don't support PC, then why not extending it and contribute back instead of doing it from scratch

Authors' Response: We thank the reviewer for his comment and we agree that it would have been better to improve the functionality of the existing tools by adding the support of PC method. However, for the sake of this project, we found it more straightforward to build a platform by ourselves based on the recommended best practices for conducting crowdsourcing tests. For example, we included the screen tests mentioned in reference [1] in order to increase the reliability of the subjects.

[1] Hossfeld, T., Keimel, C., Hirth, M., Gardlo, B., Habigt, J., Diepold, K., Tran-Gia, P.: Best practices for QoE crowdtesting: QoE assessment with crowdsourcing. IEEE Trans. on Multimedia 16(2), 541–558 (2014)

Comment 2: Fig. 3 as is makes no sense as it just shows the video frame with some additional buttons or whatever instead of the entire application.

Authors' Response: Figure 3 is a screenshot of the entire application when in the state of showing the first pair of the videos. This is stated in the caption. The frame of the internet browser is not included in the screenshot since that can change from viewer to viewer. The previous version of the paper included a more detailed explanation of the page (included below), which was cut due to reviewer comments in the previous review. Several more screenshots of the application have been included in the online software repository.

The evaluation loop consists of 3 pages. Two of them handle the video playback, while the third presents the question regarding preference. ***On the video pages, a single button which starts the video playback is displayed along with information about the progress of the test. When the playback has finished, another button leading to the next page is displayed.*** *If the window containing the video becomes inactive, e.g. another window is opened, the video will automatically pause until the play button is pressed again. An example of the playback page can be seen in Fig. 3. In order to ensure a smooth uninterrupted playback, unless it is intentional, videos are at high frame rate played (invisibly to the user) on previous pages. In this way, we ensure that the videos are fully downloaded, buffered and ready to be displayed locally.*

Comment 3: Fig 8 shows low correlation between crowdsourcing and lab-based study and conclusions are not convincing as they're more questions are left open than has been answered.

Authors Response: We understand that the drawn conclusions mainly hint at the change of methodology from ACR to PC as being one of the possible reasons of the relatively low correlation between laboratory-based and crowdsourcing-based subjective experiments. Additionally, we would like to highlight that we have shared our reflections on it in the light of the Lab Experiment 2 (Section 6.2 of the manuscript) by indicating that the effect of this change in methodology might have been emphasized due the nature of our test video pairs which were obtained from the same original video. Therefore, we have drawn attention to the role of the content dependency in the same section. We expect to present further deep analysis on these issues in our future work, which is beyond the scope of this manuscript.

Comment 4: The paper uses author's own dataset and evaluation setup making results not very reproducible, specifically as these assets exist. Hence, authors are requested to make their assets available for others to cross-check.

Authors' Response: We thank the reviewer for the kind suggestion. However, we are bound to follow the copyrights associated with this dataset that hinders us from making it public. Nonetheless, we have added the following footnote in the manuscript so that any interested reader can obtain the dataset for research purposes.

"The used dataset cannot be made public due to copyright issues of the videos. However, interested researchers can obtain the dataset through a bilateral agreement of its use for reproduction of the results only."

Authors' Response to Reviewer 10

Comment 1: Although authors made a serious revision in the presentation of their paper, however I am still not convised about the novelty of their procedures. I think that the weaknesses of this work can not be recovered by changes in presentation. Are more deep and related to the theoritical base of this work.

Authors' Response: We thank the reviewer for appreciating the improvements in our presentation. We are convinced that our article has ample novelty based on our novel study on the usage of Paired-Comparison based subjective testing in the crowdsourcing environment for adaptive video streaming. We provide some new findings besides confirming the previously known trends. Moreover, we would like to draw attention to the fact that such studies are required to be repeated in order to reach enough confidence for the reproducibility of the results. This need is highlighted by a recently published article [1] which presents analysis of a large scale study to estimate the reproducibility of scientific studies.

1. Estimating the reproducibility of psychological science, Open Science Collaboration, Science 349, aac4716 (2015)

Multimedia Tools and Applications manuscript No.
(will be inserted by the editor)

On Subjective Quality Assessment of Adaptive Video Streaming via Crowdsourcing and Laboratory Based Experiments

Jacob Søgaard · Muhammad Shahid · Jeevan Pokhrel · Kjell Brunnström

Received: date / Accepted: date

Abstract Video streaming services are offered over the Internet and since the service providers do not have full control over the network conditions all the way to the end user, streaming technologies have been developed to maintain the quality of service in these varying network conditions i.e. so called adaptive video streaming. In order to cater for users' Quality of Experience (QoE) requirements, HTTP based adaptive streaming solutions of video services have become popular. However, the keys to ensure the users a good QoE with this technology is still not completely understood. User QoE feedback is therefore instrumental in improving this understanding. Controlled laboratory based perceptual quality experiments that involve a panel of human viewers are considered to be the most valid method of the assessment of QoE. Besides laboratory based subjective experiments, crowdsourcing based subjective assessment of video quality is gaining popularity as an alternative method. This article presents insights into a study that investigates perceptual preferences of various adaptive video streaming scenarios through crowdsourcing based and laboratory based subjective assessment. The major novel contribution of this study is the application of Paired Comparison based subjective assessment in a crowdsourcing environment. The obtained

Jacob Søgaard
Technical University of Denmark, Kgs Lyngby, Denmark
Tel.: +45 45 25 65 68
Fax: +45 45 93 65 81
E-mail: jsog@fotonik.dtu.dk

Muhammad Shahid
Blekinge Institute of Technology, Blekinge, Sweden
Tel.: +46(0)45 538 57 46
E-mail: muhammad.shahid@ieee.org

Jeevan Pokhrel
Montimage, Paris, France
E-mail: jeevanpokhrel@gmail.com

Kjell Brunnström
Acreo Swedish ICT, Kista, Sweden
Mid Sweden University, Sundsvall, Sweden
Tel.: +46 (0)70 841 91 05
E-mail: kjell.brunnstrom@acreo.se

results provide some novel indications, besides confirming the earlier published trends, of perceptual preferences for adaptive scenarios of video streaming. Our study suggests that in a network environment with fluctuations in the bandwidth, a medium or low video bitrate which can be kept constant is the best approach. Moreover, if there are only a few drops in bandwidth, one can choose a medium or high bitrate with a single or few buffering events.

Keywords Adaptive Video Streaming · Crowdsourcing · Subjective Quality Assessment · Quality of Experience

1 Introduction

The increasing trend of the usage of video services for applications related to business, education, and entertainment purposes necessitates better coding technologies for efficient storage and transmission of video data. Further, this requirement is emphasized by the growing size of the display devices capable of playing videos with High Definition or Ultra High Definition resolution. Most of the video streaming is delivered over the Internet and it is estimated that the proportion of video data over the Internet will grow further [9]. This trend of the usage of multimedia services is expected to raise the users' awareness about perceptual quality. Both service providers and consumers are interested in the delivered level of perceptual quality. Besides compression, the perceptual quality of videos can get degraded due to distortions in the transmission medium. In traditional video streaming, effects on users' Quality of Experience (QoE) due to varying network conditions have not been addressed completely. Hypertext Transfer Protocol (HTTP) based Adaptive Streaming (HAS) offers a video streaming solution that is more robust against network induced distortions, so there are only limited or no losses in the transmission. One of the salient features of HAS, which is standardized by Motion Pictures Expert Group (MPEG) as Dynamic Adaptive Streaming over HTTP [1], is the availability of control with the client in order to adapt the video streaming to the varying network conditions. The adaptation is made possible through the provision of multiple bitrate copies in segments of the video being transmitted from the server. The usage of adaptive streaming has been notably observed to be useful in the reduction of video stalling that might occur due to bandwidth constraints [45]. A client might prefer to switch to lower quality video instead of experiencing a halt (buffering) in the video playback. Moreover, such a provision of multiple levels of video quality has other advantages as well, besides offering the possibility of adapting to the consumers' display terminal and the preferred price plan of the services.

As of now, HAS is being developed to find optimal solutions for its various stages. For example, under what conditions, would it be perceptually preferable to switch to a lower quality in order to avoid a stalling in the playback? Also, the options of slowly or rapidly switching to the lowest or highest quality might pose different impacts on the user QoE. To this end, subjective experiments of quality assessment are performed using a test stimuli that is representative of different adaptation scenarios. The research in this area has received rather big attention during recent years in the efforts to obtain an understanding of the perceived quality of the comparably slowly varying quality changes, which HAS

gives rise to, combined with the abrupt halt that occasionally still may occur due to rebuffering. A problem in this research area is that international standards for conducting subjective tests for HAS are still largely missing. The currently available standards, e.g. ITU-R Rec. BT.500-13 [3], ITU-T Rec. P.910 [2], and ITU-T Rec. P.913 [4], only cover methods for short sequences with the exception of Single Stimulus Continuous Quality Evaluation (SSCQE). SSCQE is a method for continuously giving quality scores and could be used for studying HAS. However, it is a method that is hard to setup and carry out, since it requires precisely calibrated scoring devices. The viewer may drift in quality without noticing it and there is usually a delay between the occurrence of an event and the time instant of the viewer response. Furthermore, the aggregation to an overall score is not straightforward. An alternative approach that can avoid such issues is Paired-Comparison (PC) based subjective assessment that we have used in this study and is presented in Section 4.1

Subjective experiments are considered to be the most valid methodology to assess the QoE and are generally conducted in a controlled laboratory environment. Objective or computer software assisted methods [34] have been largely seen as an alternative approach, to get around the complications involved in the laboratory based subjective experiments. However, even the objective methods with state-of-the-art performance are not considered as an optimal replacement of subjective assessment. Crowdsourcing based subjective experiments have gained attention to replace needs of laboratory based tests and these experiments offer promising correlation with the latter [19]. The procedure of crowdsourcing mainly involves collecting subjective assessment of quality through ubiquitous streaming via the Internet. This enables the investigator to receive opinions from a vast variety of viewers; in a time-flexible, test-data size scalable, and swift manner. In this study we investigate how crowdsourcing could potentially be used for studying HAS and how it relates to more traditional laboratory testing.

2 Goals and Contributions

This paper presents a subjective study on HAS through crowdsourcing as briefly reported in [35]. We employ a test stimuli representative of various adaptive video streaming scenarios that are adopted in practice by most service providers to find a good ABR strategy (e.g. how to arrange the bitrate budget). Additionally, we report on the content-dependency of the subjective QoE. We also report a laboratory based subjective experiment in order to further investigate the results of the crowdsourcing based experiment and the correlation with the laboratory based results.

In this paper we only consider bitrate adaption by adjusting the bitrate of the video encoding and not other forms of adaptation such as spatial and temporal scaling. For studies on such adaptation and their influence on the QoE we refer to [20, 32].

In comparison to the related work given in the following section, it becomes evident that more subjective studies on the assessment of various scenarios of adaptive video streaming are required. Especially, it is desirable to conduct studies that are closer to the real-life usage of video services. Therefore, the application of Paired Comparison based subjective assessment for adaptive video streaming

in a crowdsourcing environment is our major novel contribution. Additionally, we have analyzed the results of crowdsourcing experiment in the light of a follow-up laboratory-based experiment as well. The experimental design is deliberately made to not introduce too many new parameters in order to make it possible interpret the results. Therefore, we start from a data set that has already been annotated and introduce the changes from there. Based on [27] it is now clear that experimental results need to be repeated in several independent studies to make a trustworthy result. In that regard our contribution is also that we manage the repetition of some previous studies as well e.g. how a viewer reacts to one vs multiple stalling events.

The remaining part of this article is structured as the following. An account of related work is summarized in Section 3. Section 4 presents an outlook of the test data and methodology used in the subjective experiments. Section 5 summaries the model used to process the user feedback obtained through the crowdsourcing experiment. An analysis of the results is presented in Section 6. A discussion and conclusive remarks are presented in Sections 7 and 8, respectively.

3 Related Work

Robinson et al. [30] conducted a subjective study to evaluate user experience on HAS based video streaming under constrains of varying bandwith, latency, and video-data losses. Various observations were made including the preference of constant bitrate over frequently changing bitrate and a slow drop to the lower quality over oscillating bitrates. Staelens et al. [37] performed a subjective study for long duration videos on tablet devices for investigating the impact of quality switches due to adaptive streaming. They observed that users mainly perceive the change of quality from the highest to the lowest levels and high to medium quality changes remain largely unnoticeable. Moreover, stalling in the playback of videos was observed to be the least preferable. In [31] it was studied how spatial and temporal quality switching had different impact on the QoE and which features of the switching that were most relevant in relation to the QoE. The results reported in [42] investigated the optimum number of coding quality levels that could be used in an adaptive video system by studying the just noticeable difference levels that exist in the quality range of video content. The incorporation of the effects of frame rate and resolution adaptations on the user perception to obtain the encoding configuration that maximizes the QoE for a certain type of content has been investigated in [11]. The study presented in [23] particularly compared the difference of the impact of increase and decrease of quality in response to a variation in the network condition. It is reported that downgrading the quality has a stronger impact on the QoE. Similar results were obtained in the study reported in [14]. Moreover, in order to study the impact of slow or rapid variations of quality in comparison with low or high quality video streaming, [40, 38] presented the results of their subjective assessment of QoE on such test stimuli. In [28] the authors used Youtube and crowdsourcing to conduct a subjective experiment of Adaptive BitRate (ABR) streaming. The results indicated that the delivered representation bitrate and the number of stalls were the main influence factors on the QoE. Finally, [12, 18, 33] presented surveys on the studies related to various influence factors of QoE in HTTP adaptive streaming.

Of the related works mentioned above, the subjective experiments in [18,28] were conducted by crowdsourcing. In [18], authors analyse the effect of switch amplitude (i.e., quality level difference), switching frequency, and recency effects on HAS Quality of Experience (QoE) while in [28], authors analyse the effect of average representation bitrate (i.e., media throughput at the client), average startup time (or startup delay), and average number of stalls on existing DASH-based Web clients. However, the subjective test performed in these experiments do not use PC based methodology and moreover, the subjective test we performed were more intense with highest number of test videos. An introduction to crowdsourcing, a discussion of the differences between crowdsourcing and laboratory experiments, and best practices for crowdsourcing, such as including a screen test and control questions, are presented in [16,17]. In [8,29] web-based platforms for subjective studies of QoE for videos are presented. In [13] various improvements for implementing subjective crowdsourcing experiments are proposed, so-called momento methods that increase the reliability and execution time of the crowdsourcing campaign. This work builds on our previous work [35], where the initial results from the crowdsourcing experiment are presented.

4 Subjective Experiments of Video Quality

Most of the test videos used in this study were previously used in the laboratory based subjective experiment reported in [40,38]. These test videos closely resembles the video quality level and the content types used by service providers. Also, different adaptive scenarios are considered in the experiment to address the service providers' concerns. From now onwards, we refer to this as Laboratory Experiment 1. The original purpose of Laboratory Experiment 1 was to compare the outcomes of a subjective experiment using a traditional and standardized Absolute Category Rating (ACR) test methodology versus a semi-continuous methodology developed to evaluate long sequences in a more realistic setting as explained in [38]. Also, the impact of some of the technical factors of the adaptation scenarios, such as the amplitude of the quality switching and video chunk size was investigated. The impact of including or excluding audio was investigated in [39] and no statistically significant effect was found. Therefore, we chose to exclude audio in the subjective experiments done in this work.

The original videos were all in 1280x720 resolution with a frame rate of 24 fps and encoded using the high profile for H.264/AVC at 4 different bitrates: 600 kbps, 1 Mbps, 3 Mbps, and 5 Mbps. The videos were encoded with closed GOP, maximum 2 B-frames and 3 reference frames. Seven different sources were used; three sources were taken from entertainment action/romance movies (denoted Pirates, Darkhour, Streetdance) and the rest was content from: a soccer match (denoted Football), a sports documentary (denoted ClosetoEdge), a newscast (denoted News), and a concert (denoted Rollingstone). The selection of the source content is motivated in [38].[1] The subjective Laboratory Experiment 1 was carried out at the lab of Acreo Swedish ICT AB (Acreo Lab) in a test room compliant with the ITU-R BT. 500 [3].

[1] The used dataset cannot be made public due to copyright issues of the videos. However, interested researchers can obtain the dataset through a bilateral agreement of its use for reproduction of the results only.

Spatial perceptual Information (SI) and Temporal perceptual Information (TI) as defined in [2] can be used for categorizing the video content. In content with low SI values are found scenes with minimal spatial detail, while content with high SI values contains scenes with the most spatial detail. Content with low TI values consist of still scenes and very limited motion, while in content with high TI values scenes with a lot of motion are found.

The content of the original videos can be described as follows, where the spatial and temporal information are noted as (SI,TI) for each content. The Pirates video (48,29) is from an action movie and features some scenes in smooth motion, some with groups of walking people, and some with camera panning. The Darkhour video (51,28) is from a thriller movie and features scenes with rapid scene changes and cloudy atmosphere. The Streetdance video (46,34) is from a drama movie and consist mostly of scenes with smooth motion with static background and some scenes with groups of dancing people in bright ambient. The Football video (56,29) is from a TV broadcast of a football match and has moderate motion and wide angle camera sequences with panning. The ClosetoEdge video (43,24) is from a documentary and is mostly shot with a handheld camera and features varying characteristics. The Rollingstone video (45,42) is from a concert recording, where the lead singer moves around a lot and the video has some sudden scene changes. The News (49,23) video is from a TV news broadcast and has static scenes with one/two standing/sitting people, some scenes with moving background and some scenes without reporters with more motion and panning.

Several adaptation scenarios for the videos were produced in the Laboratory Experiment 1, such as going from a high to a low bitrate in a stepwise manner. Out of all those scenarios, the following are used in this work: Gradual Decreasing (GD), Rapid Decreasing (RD), constant 600 kbps (N600), constant 1 Mbps (N1), constant 3 Mbps (N3), and constant 5 Mbps (N5). Additional details of these scenarios, such as the timing of the bitrate steps, can be found in [38]. Additionally, we introduced new buffering scenarios to test the quality perception in relation to the aforementioned scenarios. The buffering scenarios include: 1 Freezing event for 2 seconds in the constant 3 Mbps video (1F3M), 2 Freezing events for 1 second each in the constant 3 Mbps video (2F3M), and 1 Freezing event for 2 seconds in the constant 1 Mbps video (1F1M). The freezing events were in most cases placed in an evenly spaced manner, except when this coincided with or were very close to a scene change; in this case the freezing event was moved a few seconds away from the scene change, so the interaction between those two effects was minimized. We did not consider initial delay in our test stimuli as some studies, e.g. [15], noted that it does not seem to pose a significant impact on the QoE for the user. Due to the semi-continuous methodology used in previous work, some of the degradations were applied to the content at different time intervals. Thus, each Processed Video Sequences (PVS) originating from a specific original content as described above might be from different time intervals in that original content.

In total 9 different scenarios were used, resulting in a total of 63 stimuli. Table 1 presents a summary of this test stimuli. Using this test stimuli, we conducted a crowdsourcing experiment that is referred to as Crowdsourcing Experiment in the rest of the paper. Additionally, a laboratory based experiment was performed that we refer to as Laboratory Experiment 2 in this article. Table 2 presents a summary of the usage of this test stimuli in different experiment setups. Note that Laboratory Experiment 2 was mainly conducted to better understand any

Table 1 Test Stimuli

No.	Code	Description of Client Behavior
1	GD	Gradually decreasing the quality by bitrate order 5-3-1-0.6 Mbps
2	RD	Rapidly decreasing the quality by bitrate order 5-0.6 Mbps
3	N600	No change in quality level by keeping the bitrate constant at 0.6 Mbps
4	N1	No change in quality level by keeping the bitrate constant at 1 Mbps
5	N3	No change in quality level by keeping the bitrate constant at 3 Mbps
6	N5	No change in quality level by keeping the bitrate constant at 5 Mbps
7	1F1M	One frame-freeze event of 2 seconds at a constant bitrate of 1 Mbps
8	1F3M	One frame-freeze event of 2 seconds at a constant bitrate of 3 Mbps
9	2F3M	Two frame-freeze events of 1 second each at a constant bitrate of 3 Mbps

Table 2 Use of Test Stimuli in Experiments

Instance	Used Test Stimuli cf. Table 1
Laboratory Experiment 1	1 to 6
Crowdsourcing Experiment	1 to 9
Laboratory Experiment 2	1 to 9

difference between the results of Lab Experiment 1 and the Crowdsourcing Experiment to confirm whether it is due to the difference in test material or due to the experimental methodology. Based on the findings in Lab Experiment 2 we were able to conclude that the difference in the results is most likely due to the PC methodology as discussed in Section 6.

4.1 Crowdsourcing Experiment

Crowdsourcing is a powerful and cost effective tool to perform short and simple tasks online as it facilitates the access to a large number of fairly low price workers in a short period of time. However, performing multimedia subjective quality assessment with crowdsourcing brings many challenges. If the resources at viewers end are limited for instance, low internet connections, low resolution screens etc., it is very difficult to transmit and display high quality multimedia contents. Moreover, having very little control over the viewers environment, such as viewing conditions, viewers mental state etc., makes crowdsourcing tests untrustworthy compared to laboratory tests. In addition, it is very difficult to check the reliability of the viewers. Therefore, in order to deal with these challenges, we have embedded different screen tests in our crowdsourcing tool. Viewers are obliged to perform screen tests and answer survey questions, which helps us to know about the visibility power and personal background of the viewers along with some of the display properties of their screens and current environment. Different dummy questions related to the multimedia content are asked during the test in order to differentiate the unreliable viewers.

Crowdsourcing experiments should be as simple as possible for the viewer, therefore we chose to follow the Paired Comparison (PC) evaluation methodology [2], where the test sequences are grouped into pairs that are presented to the viewer one after the other. Also, to keep it simple for the viewer, after each pair of videos the viewer is simply asked which of the stimuli he or she preferred via the online interface. Since we chose the PC methodology, which is very simple for the viewer

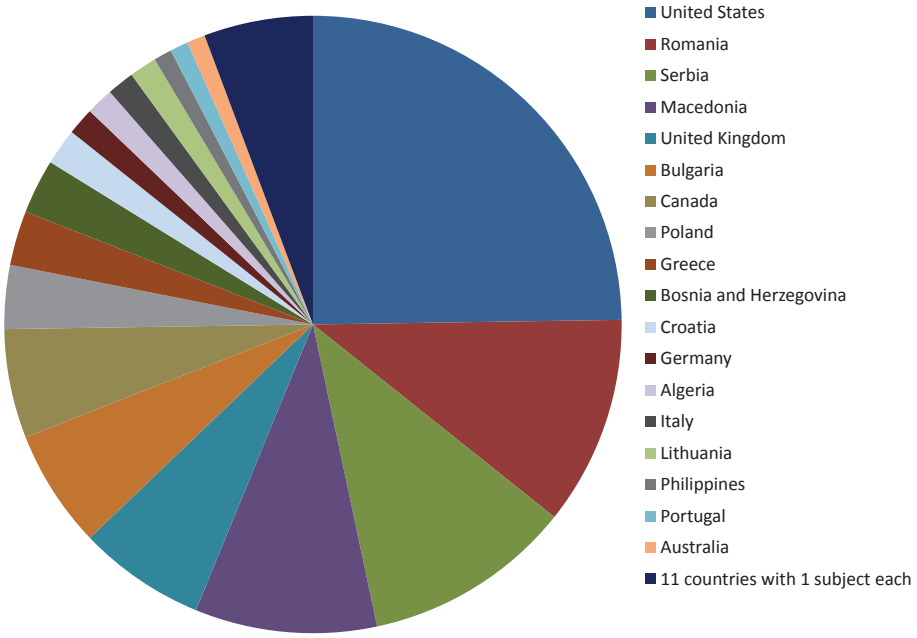

Fig. 1 Distribution of viewers across different countries.

[25], we did not train the viewers before the test, which would have been necessary if e.g. a rating scale had been used [16]. This also has the advantage that we did not influence the viewers during a training session, which can cause the obtained data to be biased.

We used the optimized square design [25] for the pairings based on our assumptions of the quality levels. This was done to reduce the number of pairings needed to get reliable measurements. Using this method, our complete test set consisted of a total of 126 pairings. These pairing were divided into 14 tasks with 3 videos from 3 different contents, i.e., 9 videos for each task. For a random video from each content the viewers also needed to answer a control question about the content. We also used the screen test from [17] prior to the subjective test to filter out potential malicious viewers. In total 215 paid viewers participated as viewers in the crowdsourcing experiment from 30 countries. Fig. 1 shows the distribution of viewers across different countries.

To conduct the crowdsourcing subjective experiment we chose to create our own web-based platform capable of presenting videos to viewers for performing paired-comparisons. In this article, we present a brief documentation of this software. For further documentation and technical implementation the reader is invited to access the Web page of the open source project that served as the basis for our platform [36]. The test videos are required to be in a playable format for internet browsers, but otherwise our platform can be used for any PC subjective video experiment, not only for ABR videos as was done in this work. Alternatives to our platform for PC subjective experiments in crowdsourcing include [8, 29]. Advantages of our platform includes: access to the source code for easy modification, the overall

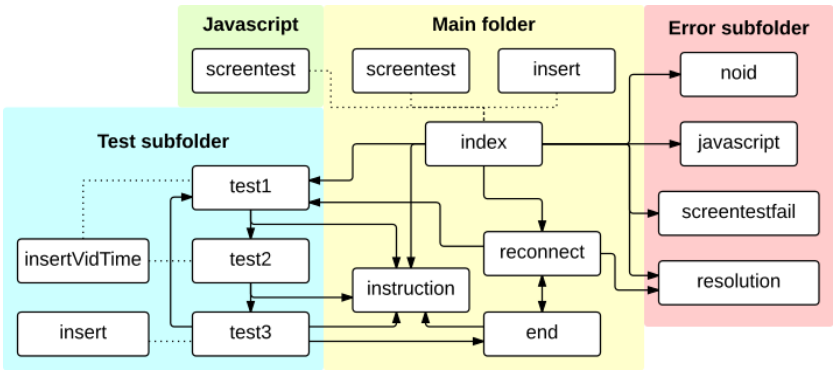

Fig. 2 Flowchart of the interface, the dashed line indicates dependency.

experiment is easily broken into smaller tasks, viewers are dynamically assigned to the current task with fewest views, an unique solution to ensure smooth playback, and the design and setup of the paired comparisons can be entirely defined by the experiment designer.

In the current version of the platform, it is assumed that each viewer should watch 9 comparisons and that each viewer is directed to the front page with a unique id. In this work, the Microworkers platform[2] was used to hire the viewers. The platform was built using the Hypertext Preprocessor (PHP) language and Javascript. The flow of direction of the interface in the interface is illustrated in Fig. 2. Dashed lines indicate dependencies, meaning that boxes connected only with dashed lines are used as part of the pages they are connected to. Most pages of the interface include a link to an instruction page which is a simple web-page with detailed instructions for the viewer.

The front page of the interface consists of a small version of the instructions and the screen test from [17] is shown. Placed below that is a small questionnaire that can be tailored with e.g. demographic questions. At the very bottom of the page is a progress bar showing the status of the loading process of the first pair of videos. When the loading is done a start button will appear, which leads to the test loop. In this way the videos are ensured to be playable without unintended interruptions. The compatibility of the viewer's platform is also tested. If any error is detected, e.g. JavaScript being disabled or the device resolution is too low, the viewer is redirected to the relevant error page automatically. The error pages contain information about the specific errors and what the viewer might be able to do in order to redeem the error.

The evaluation loop consists of 3 pages. Two of them handle the video playback, while the third presents the question regarding preference. An example of the playback page can be seen in Fig. 3. In order to ensure a smooth uninterrupted playback, unless it is intentional, videos are at high frame rate played (invisibly to the user) on previous pages. In this way, we ensure that the videos are fully downloaded, buffered, and ready to be displayed locally.

For 1/3 of the video pairs, a control question related to the content in the video is also asked on the preference page. The answers from the control question can be used when filtering out unreliable viewers. In this way screening of the viewers

[2] https://microworkers.com

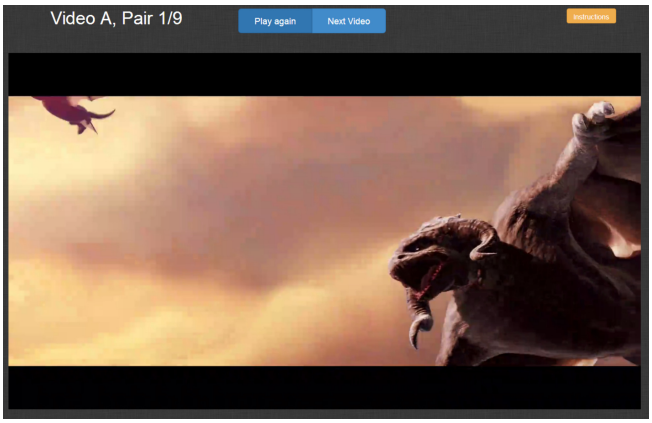

Fig. 3 Screenshot of the first video playback page. The video frame is from [5].

is done twice: first based on the information collected at the front page before the evaluations of the videos as described above, and finally after the experiment has been concluded when information about the control questions is also available. If there are more pairs for the viewer to evaluate, the next pair of videos will be loaded on the preference page, with a progress bar showing the status at the bottom of the page. In this case the viewer will be redirected back to the first video playback page. Otherwise, the viewer will be redirected to the end page. At the end page, the viewer will receive a unique code as documentation of their participation.

The status page contains information meant for the test manager. It provides a quick overview of the current status of the tests with information which is retrieved directly from the database. The interface is connected to a database in order to store and exchange information between the interface and participants. The database stores information about viewer answers to video pair preferences, content control questions, screen test results (as described below), the initial buffer time, the time spent on playback for each video and the current progress for each viewer.

Screen tests are used to find the end user watching conditions. If the watching conditions for the viewers are not favorable then the test scores available are not reliable. In our web based application, we therefore applied two screen test mechanisms as described in [17] [13]. In the screen test the visibility of the symbols depend on different conditions, such as screen orientation, screen resolution, screen brightness, screen color combination, viewer's eyesight etc. Viewers are not allowed to proceed in the test without performing the screen test. In addition, an unreliability score was calculated using the screen test implementation from [17]. Based on the unreliability scores obtained from the screen test, 215 viewers out of 266 potential viewers were allowed to complete a subset of the experiment.

4.2 Laboratory Experiment 2

As the test videos used in the Crowdsourcing Experiment (Section 4.1) were composed of additional test scenarios as compared to Laboratory Experiment 1, it

might be argued that the results of the two setups can not be compared directly due to the difference of the test data. Therefore, in order to better understand any deviation of results from the Laboratory Experiment 1 and the crowdsourcing experiment if any, the videos chosen for Laboratory Experiment 2 were the same as the ones used in the Crowdsourcing Experiment (Section 4.1) while the evaluation methodology used was the same as the Laboratory Experiment 1 [40], namely the Absolute Category Rating (ACR) test methodology [2] with a discrete rating scale from 1 to 5.

As in [40], the subjective experiment was carried out at the Acreo Lab. The lab was equipped with a 46" Hyundai S465D display with the native resolution of 1920x1080 and 60 Hz refresh rate. Viewing distance was 4 times of display height. The peak white luminance of TV was 177 cd/m^2 and the ambient illuminance level in the room was about 20 lux. A modified version of the VQEGPlayer [7] was used to present the randomized PVS and the voting interface after each PVS.

The viewers were initially screened for visual acuity (Snellen chart), color vision (Ishihara), and asked to fill the Simulator Sickness Questionnaire (SSQ) [10] as well as answering some questions about their background and video habits. The viewers were then instructed in the testing procedure and a training session was performed, so the viewers could familiarize themselves with the procedure and the range of the qualities. During the training session some examples of PVSs including quality variation (adaptation scenarios) and videos with buffering events were shown. The actual test with the randomized PVSs were then carried out in one session lasting around 20 to 30 minutes. After the test, the viewers were again asked to fill the SSQ. The viewers were of different ages (mean 32.5, median 28.5, max 60 and min 18) and background. There were 7 female and 15 male.

5 Processing of the pair-comparison data

In order to analyze the results obtained from the pair-comparison tests in the crowdsourcing experiment and being able to compare with the results from the Lab tests, it is required to convert the obtained preference values into quality values for each PVS. As noted in some related studies [24, 25], it is possible to compare the results obtained through pair-comparison with the results obtained via ACR method. To this end, we use the Bradley-Terry-Luce (BTL) model [6], given its popularity in similar studies of video quality assessment, e.g. [21, 22]. If p_{ij} represents the probability that a video stimulus i is preferred over a stimulus j, the related BTL takes the form as given by the following:

$$p_{ij} = \frac{\pi_i}{\pi_i + \pi_j} \tag{1}$$

where π_i is the quality score for stimulus i and it can take on any value between or equal to zero and one. This expression can be reformulated by using the empirical probability of preference values, i.e.,

$$p_{ij} = \frac{m_{ij}}{m_{ij} + m_{ji}} \tag{2}$$

where m_{ij} is the frequency of stimulus i being preferred over stimulus j. The corresponding π_i can be computed by maximizing the log-likelihood function given

by the following expression:

$$L(\pi_1, \pi_2, \pi_3, ..., \pi_n) = \sum_{i=1}^{n} \sum_{j=1}^{n} p_{ij} \ln(\frac{\pi_i}{\pi_i + \pi_j}) \quad (3)$$

where n is the number of stimuli. This expression can be solved by modern computer assisted iterative methods and we adopted the optimization routines in a software package as mentioned in [44]. This packages relies on the Matlab function `fminsearch` to find solution of the model. Specifically, we used the preference matrices of the nine stimuli. Necessary cautions and measures were taken to avoid any local extrema points. To this end, we carefully inspected the covariance matrix so that it did not contain negative values on its main diagonal. Moreover, an initial seed of likelihood values has been provided to the software package and the obtained likelihood values have been used as seed to an iterative call to the underlying function. The iterations were conditioned to a tolerance values of 10^{-4} for the difference of the likelihood values for successive calls.

Among other parameters of the model, this software provides Hessian matrix of the log-likelihood function and that can be used to compute the related covariance matrix. This in turn can be used to estimate standard errors from the main diagonal of the covariance matrix denoted $Diag\big[\widehat{cov()}\big]$. Finally, the 95% confidence intervals are obtained by the following:

$$\pm 1.96 \sqrt{Diag\big[\widehat{cov()}\big]} \quad (4)$$

We transform the obtained estimates of the probability values using the natural logarithm [6] and normalize them to the interval of the ratings in the laboratory experiments for obtaining mean opinion score (MOS) values. The same transformation is used on the bounds of the confidence intervals from equation (4). Thus, the transformed confidence interval will be uneven due to the natural logarithm transform.

6 Results and Discussion

6.1 Crowdsourcing Experiment

To analyze the results we applied the Bradley-Terry-Luce (BTL) model [6] as detailed in Section 5. The viewers were filtered by excluding viewers with too many unlikely preferences. We define an unlikely preference as a preference where the corresponding probability in the BTL model is lower than a threshold θ. In our test, we only allow 2 out of 9 unlikely preferences and we set $\theta = 0.25$. With this approach 6 viewers were excluded from the final results.

In order to validate the results obtained from the crowdsourcing experiment, we compared the opinion scores obtained from the Laboratory Experiment 1 [40] to the crowdsourcing experiment. This is shown in Fig. 4. The results show that the opinion scores obtained from both the experiments are strongly correlated, however not as strongly as could be expected of a repetition of a laboratory test. This can be due to the differences in the test setup, such as evaluation method, viewing environment and the introduction of new distortions. To investigate this,

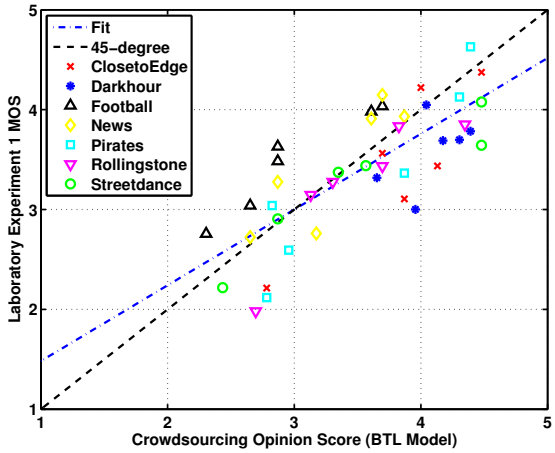

Fig. 4 Comparison between Laboratory Experiment 1 and the crowdsourcing subjective experiment. Linear correlation: 0.77.

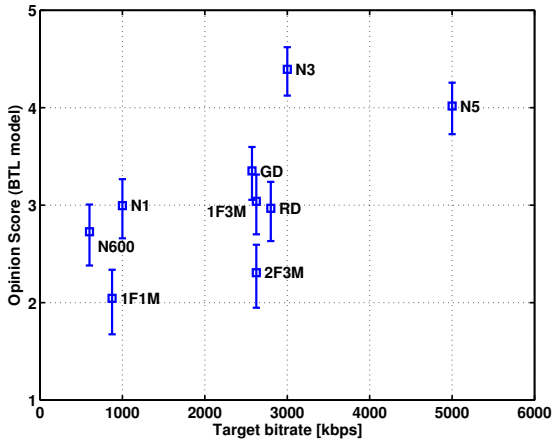

Fig. 5 Opinion Scores versus the average bitrate for the crowdsourcing experiment.

we performed a new laboratory test, Laboratory Experiment 2, as detailed in Section 4.2. The results and comparison to this test can be seen in Section 6.2.

Our experiment verifies the results from earlier studies, e.g., [26], that buffering events have a high impact on the QoE. Due to this, users generally prefer viewing videos at lower bitrates than having buffering events in videos at higher bitrates.

The quality of the videos can also be compared against the average bitrate of the videos. This has been illustrated in Fig. 5, where the mean of the subjective scores has been calculated over the video contents. Generally, users prefer videos at higher bitrates, i.e., 3 or 5 Mbps and the difference between them is probably more due to the difference in content than the difference in compression levels. Users dislike buffering events and it seems that the frequency is more important than the total duration of these events (both videos at 3 Mbps with buffering has a total buffering time of 2s), which is in line with earlier studies e.g. [43]. But if

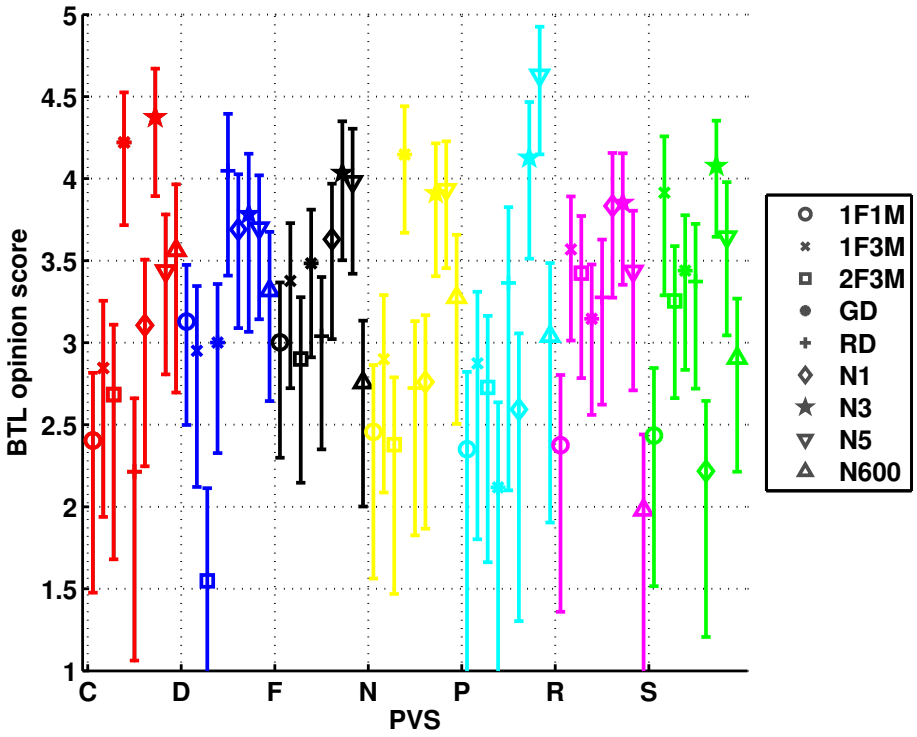

Fig. 6 Opinion Scores with 0.95 confidence intervals for every PVS for the crowdsourcing experiment. The PVSs are indexed by the first letter in their names (see Section 4).

the bitrate is high enough and the frequency of the buffering events is low enough, e.g., the 1F3M video, this seems to be a viable alternative to decreasing the bitrate of the video or having a constant low bitrate, e.g., 600 kbps or 1 Mbps.

We also investigated the impact of the video content on perceptual preference of different adaptation scenarios. In contrast to relying only on Spatial and Temporal perceptual Information (SI and TI) indices [41], we also analyze our results using semantic indicators such a genre (e.g. action movie, concert recording, and newscast as outlined for the test sequences in Section 1) of the video as well. The opinion scores for each PVS can be seen in Fig. 6 and it is evident that the content plays a major role in the perception of quality. The content with lowest standard deviation of the BTL scores across the different degradation types is the Football content (with a standard deviation of 0.47 compared to values from 0.63 to 0.82 for other contents), which might be explained by the high spatial complexity of that content, which makes the different adaptation strategies more attractive compared to lowering the bitrate. The content with the highest standard deviation of the BTL scores across degradation types is the Pirates content (with a standard deviation of 0.82 compared to the second highest of 0.76) due to very high scores for high bitrates and low scores for medium to low average bitrates. This could be due to that the source for this content is high quality, visually pleasing, and from a very well known blockbuster making it easier to distinguish between

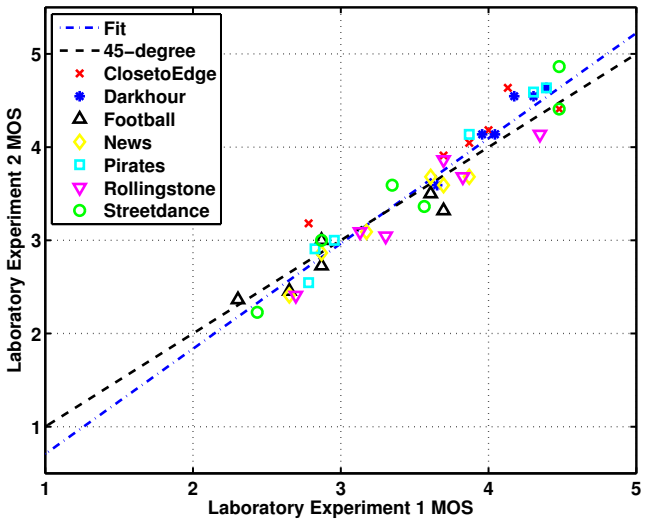

Fig. 7 Scatter plot for overlapping sequences in Laboratory Experiment 1 and Laboratory Experiment 2. Linear correlation: 0.96.

quality levels. For ClosetoEdge and News the PVS with gradual decreasing bitrate has higher BTL scores than other content (statistical significant with an overall confidence level of 0.90 to Darkhour, Pirates, and Rollingstone), which could be due to the general low temporal complexity of these sequences. For Darkhour the 2F3M scores lower than all other PVSs, which might be due to suspense being interrupted by two pauses. In the Football sequence the constant 600 kbps ranks low compared to other Football PVSs, which might be due to high temporal complexity of the sequence, causing a lot of flickering artifacts at low bitrates. This is also the case for the Rollingstone content where the uniform black background contains a lot of very noticeable artifacts at 600 kbps. The value of the BTL scores of the 600 kbps PVS subtracted from the mean of the other PVSs for the Football and Rollingstone contents are respectively 0.67 and 1.38 compared to other contents where this value is in the range of -0.40 to 0.39. We also note that for ClosetoEdge the 5 Mbps video is rated lower than expected, which is probably due to specific content in this PVS.

We also performed statistical tests on all overlapping PVSs between the crowdsourcing test and Laboratory Experiment 1 [40] for significance difference. With an overall confidence level of 90% there were no significant difference in the means of any PVS.

6.2 Lab Experiment 2

Before any analysis of the results, screening of the observers according to ITU-R Rec. BT.500-13 [3] was applied. No observers were eliminated in the screening. The MOS results of this laboratory experiment correlates very well with the original laboratory test as it can be seen in Fig. 7. The linear correlation coefficient with the crowdsourcing results are on the other hand only 0.69 if the same subset of

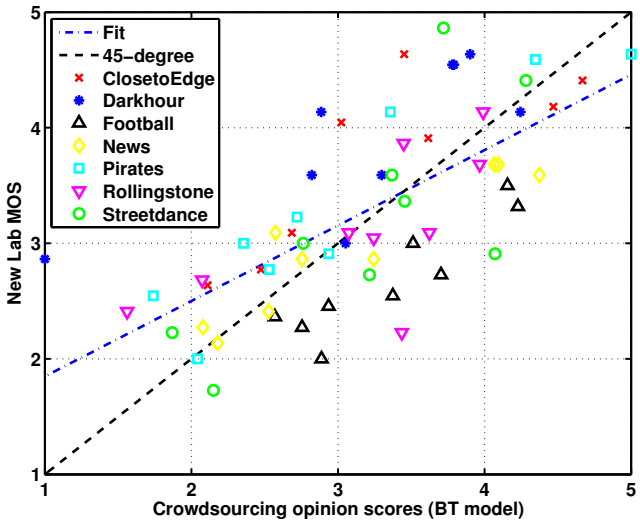

Fig. 8 Scatter plot for Crowdsourcing and Lab 2 experiment

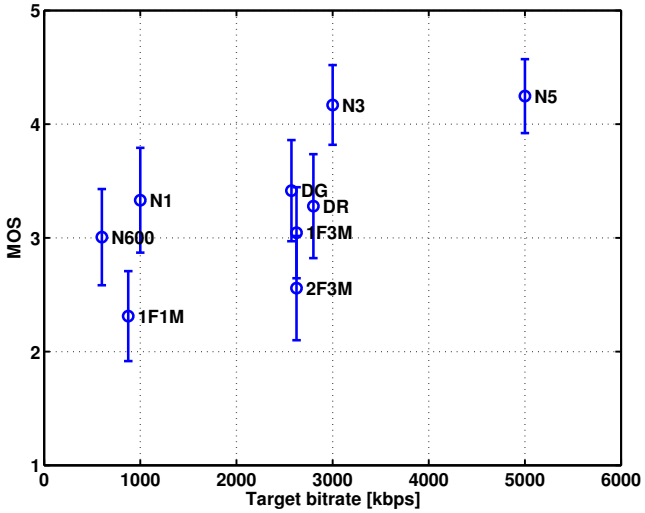

Fig. 9 MOS versus the average bitrate for Laboratory Experiment 2.

videos are used as in Fig. 8, while the linear correlation is 0.70 if the full set of videos are used to calculate it. Therefore, it seems that the PC methodology is not suited as a substitution for the ACR methodology in this case where the video pairs can be from different periods of time in the original source video. Even so, the trend of the overall ranking of the degradation which can be seen in Fig. 9 seems to be well aligned with results from the crowdsourcing experiment.

The MOS for each PVS can be seen in Fig. 10 and again it is evident that the content plays a major role in the perception of quality, however in some cases the conclusions seems to differ somewhat from the crowdsourcing experiment. Generally, the content originating from blockbusters (Closetoedge, Darkhour, Pirates

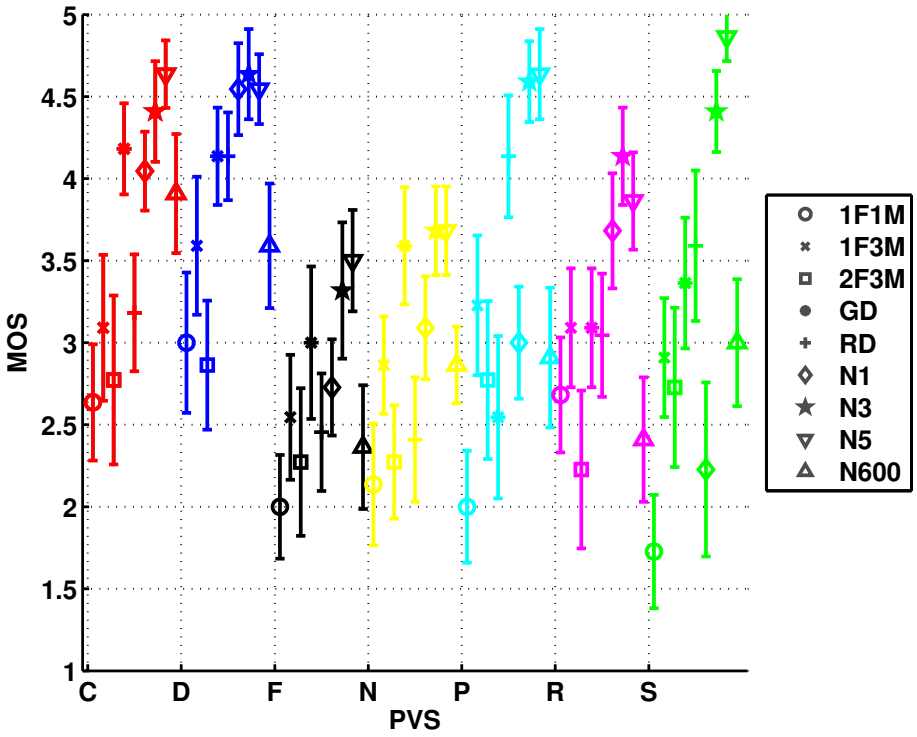

Fig. 10 MOS with 0.95 confidence intervals for every PVS for Laboratory Experiment 2. The PVSs are indexed by the first letter in their names (see Section 4).

and Streetdance) have higher MOS for high constant bitrates than the rest of the content. This difference in the experiment could be caused by the assumed difference in screen sizes and screen quality from the crowdsourcing experiment to the lab experiment. The content with lowest standard deviation of the MOS across the different degradation types is still the Football content (now with a standard deviation of 0.50 compared to values from 0.61 to 0.99 for other contents). However, the content with the highest standard deviation of the MOS across degradation types is the Streetdance content (with a standard deviation of 0.99 with Pirates begin second highest with a value of 0.93). In this experiment the PVS with gradual decreasing bitrate still has higher MOS for ClosetoEdge, but now also for Darkhour instead of News compared to other content (statistical significant with an overall confidence level of 0.90 to all other contents except News). For Darkhour the 2F3M score is still the lowest for that specific content, but not the lowest score when compared to other content. The 600 kbps PVS for the Rollingstone content where the still has a very low score compared to the other degradations in that content, but this trend is not nearly as drastically for the Football content. The value of the MOS of the 600 kbps PVS subtracted from the mean of the other PVSs for the Rollingstone content is 0.82 compared to other contents where this value is in the range of -0.29 to 0.45.

Table 3 Summary of the results.

Description	Experiment	Result
Correlation:	Crowdsourcing and laboratory 1	0.77
	Laboratory 1 and 2	0.96
	Crowdsourcing and laboratory 2	0.69
Lowest MOS variation:	Crowdsourcing	Football
	Laboratory 2	Football
Highest MOS variation:	Crowdsourcing	Pirates
	Laboratory 2	Streetdance
Highest MOS for GD:	Crowdsourcing	Closetoedge
	Laboratory 2	Closetoedge
Buffering with highest MOS (across content):	Crowdsourcing	1F3M
	Laboratory 2	1F3M
Buffering with lowest MOS (across content):	Crowdsourcing	1F1M
	Laboratory 2	1F1M

For Laboratory Experiment 2, we also performed statistical tests to see if the means of the PVSs were significant different. In this case, we found that at an overall confidence level of 90% four PVSs had significant different means. This corresponds to 6.3% of the total number of PVSs. The four cases were: the movie clip from ClosetoEdge with constant 5 Mbps bitrate, the movie clips from Darkhour with gradually decreasing bitrate and with constant 5 Mbps bitrate, and the movie clip from Streetdance with constant 5 Mbps bitrate. In all four cases the scores from the crowdsourcing experiment was significantly lower than the scores from Laboratory Experiment 2.

7 Discussion

The lab-based experiments are more reliable than crowdsourcing experiments however, they have limitations such as 1) high cost in terms of time and labor, 2) limited participant's diversity. In addition, users need to be physically present in the laboratory to perform the test [46]. On the other hand, crowdsourcing experiments allow an investigator to get opinions from a vast variety of subjects; in a time-flexible, test-data size scalable, and swift manner [35].

A summary of the results of both experiments is presented in Table 3. The obtained results shows acceptable correlation between the laboratory test and the crowdsourcing test. One of the trade-offs when using crowdsourcing compared to laboratory studies is the increase in uncertainty that comes from lack of control of the viewer, the environment of the viewer when doing the test and the equipment used by the viewer. This could manifest itself into increased variation or standard deviation in the experiment. On the other hand it is relatively cheap to increase the sample size in crowdsourcing test compared to a corresponding lab test. Let us assume a fairly common set-up for a video quality test with about 100 video clips in a within subject design and we assume Normal distribution. If we would like to be able to find difference in MOS that is e.g. about one, i.e. change one level on the ACR 5 point scale, and compensating for multiple comparisons using Bonferroni, giving an overall 95% confidence of significance for the whole test, which would give an alpha of 0.05/(100*(100-1)/2 = 0.00001 per comparison. In the Fig. 11 we plot the required sample size to reach a power of the test to be 0.8 as function

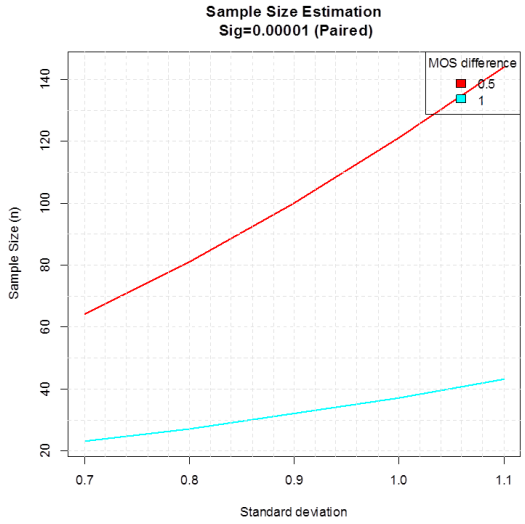

Fig. 11 The required sample size to show a significant difference between MOS that have differences of 0.5 and 1.0, as a function of standard deviation for a test of 100 video clips and giving a power of 0.8.

of the standard deviation for MOS differences of 0.5 and 1.0. We can then see that if for instance the standard deviation is increased from 0.8 to 1.0 for MOS difference 1.0 (blue curve), we would need to go from less than 30 viewers to about 40 viewers to keep the power of the test at the 0.8 level. Half a MOS is also an interesting case, then we would like to resolve a 0.5 MOS difference, which is also shown in the Fig. 11 (red curve), the same increase will require that the number viewers are increased from about 75 to about 125. The point of this discussion is that adding 50 or 100 viewers in crowdsourcing is fairly easy and could very well compensate the increase in uncertainty based on the lack of control. Due to this advantage and the promising correlation, crowdsourcing experiments have gained enough popularity and can be a alternative to lab-based experiments.

8 Conclusion

We presented a study on the investigation of crowdsourcing based subjective assessment of video quality as a potential alternative for laboratory based experiments. Our novel approach includes the application of Paired Comparison based subjective assessment in a crowdsourcing environment. In our experiments, we employed a test stimuli representative of various adaptive video streaming scenarios that are adopted in practice by most service providers. The subjective experiment conducted in a crowdsourcing environment verifies the results of earlier studies of adaptation scenarios, including the effect of buffering events. Our study suggests that in a network environment with fluctuations in the bandwidth, a medium or low video bitrate which can be kept constant is the best approach. Moreover, if there are only a few drops in bandwidth, one can choose a medium or high bi-

trate with a single or few buffering events. In this case the duration of the buffering events should be long enough to minimize the risk of another buffering event in the near-future. Additionally, we reported on the content-dependency of the subjective QoE.

Lastly, we conducted a laboratory based subjective experiment to further investigate the results of the crowdsourcing based experiment. The obtained results suggest that correlation of the crowdsourcing based results with the laboratory based might have been affected by the use of paired-comparison (PC) technique of presentation of test stimuli to the viewers combined with the intermix of content and degradations. More experiments can be performed to verify this indication to weigh this disadvantage of PC as compared to its advantage in simplifying the test procedure. This is especially important in a crowdsourcing environment where an investigator has lesser control on test setup adopted by a remote user.

An interesting extension of this work would be to analyze the demographics of the viewers from the experiments, especially the crowdsourcing experiment, and investigate any correlations between the demographics and the perceptual preference of video quality.

References

1. Information technology – dynamic adaptive streaming over HTTP (DASH) – part 1: Media presentation description and segment formats. `http://www.iso.org/iso/home/store/catalogue_ics/catalogue_detail_ics.htm?csnumber=65274`.
2. ITU-T Rec. P.910: Subjective video quality assessment methods for multimedia applications (2008).
3. ITU-R Rec. BT.500-13: Methodology for the subjective assessment of the quality of television pictures (2012).
4. ITU-T Rec. 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 (2014).
5. Blender Foundation: URL `http://www.sintel.org`
6. Bradley, R.A., Terry, M.E.: Rank analysis of incomplete block designs: I. The method of paired comparisons. Biometrika **39**, 324–345 (1952)
7. Brunnström, K., Cousseau, R., Jonsson, J., Koudota, Y., Bagazov, V., Barkowsky, M.: VQEGPlayer: open source software for subjective video quality experiments in windows. `http://vqegjeg.intec.ugent.be/wiki/index.php/VQEGplayer-main` (2014).
8. Chen, K.T., Chang, C.J., Wu, C.C., Chang, Y.C., Lei, C.L.: Quadrant of euphoria: a crowdsourcing platform for QoE assessment. Network, IEEE **24**(2), 28–35 (2010)
9. Cisco Visual Networking Index: Global mobile data traffic forecast update, 2014-2019.
10. Coolican, H.: Research methods and statistics in psychology. Psychology Press, (2014)
11. Cranley, N., Murphy, L.: Incorporating user perception in adaptive video streaming systems. Digital Multimedia Perception and Design pp. 244–265 (2006)
12. Garcia, M.N., De Simone, F., Tavakoli, S., Staelens, N., Egger, S., Brunnstrom, K., Raake, A.: Quality of experience and HTTP adaptive streaming: A review of subjective studies. In: Sixth Int'l Workshop on Quality of Multimedia Experience (QoMEX), pp. 141–146. IEEE (2014)
13. Gardlo, B., Egger, S., Seufert, M., Schatz, R.: Crowdsourcing 2.0: Enhancing execution speed and reliability of web-based QoE testing. In: Proceedings of the International Conference on Communications, pp. 1070 – 1075 (2014)
14. Grafl, M., Timmerer, C.: Video quality in next generation mobile networks - perception of time-varying transmission. In: Proceedings of the 4th International Workshop on Perceptual Quality of Systems, pp. 178–183 (2013)
15. Hossfeld, T., Egger, S., Schatz, R., Fiedler, M., Masuch, K., Lorentzen, C.: Initial delay vs. interruptions: Between the devil and the deep blue sea. In: Proceedings of the Fourth International Workshop on Quality of Multimedia Experience (QoMEX), pp. 1–6 (2012)

16. Hoßfeld, T., Keimel, C.: Quality of Experience. Springer (2014). Crowdsourcing in QoE Evalution.
17. Hossfeld, T., Keimel, C., Hirth, M., Gardlo, B., Habigt, J., Diepold, K., Tran-Gia, P.: Best practices for QoE crowdtesting: QoE assessment with crowdsourcing. IEEE Trans. on Multimedia **16**(2), 541–558 (2014)
18. Hossfeld, T., Seufert, M., Sieber, C., Zinner, T.: Assessing effect sizes of influence factors towards a QoE model for HTTP adaptive streaming. In: Proceedings of the Sixth International Workshop on Quality of Multimedia Experience (QoMEX), pp. 111 – 116 (2014)
19. Keimel, C., Habigt, J., Horch, C., Diepold, K.: Qualitycrowd - a framework for crowdbased quality evaluation. In: Proceedings of Picture Coding Symposium, pp. 245–248 (2012)
20. Korhonen, J., Reiter, U., Ukhanova, A.: Frame Rate versus Spatial Quality: Which Video Characteristics Do Matter? In: Proceedings of IEEE International Conference on Visual Communication and Image Processing (VCIP13), pp. 1–6 (2013)
21. Lee, J.S., De Simone, F., Ebrahimi, T.: Subjective quality evaluation via paired comparison: Application to scalable video coding. IEEE Transactions on Multimedia **13**(5), 882–893 (2011)
22. Lee, J.S., De Simone, F., Ramzan, N., Zhao, Z., Kurutepe, E., Sikora, T., Ostermann, J., Izquierdo, E., Ebrahimi, T.: Subjective evaluation of scalable video coding for content distribution. In: Proceedings of the International Conference on Multimedia, pp. 65–72 (2010)
23. Lewcio, B., Belmudez, B., Mehmood, A., Waltermann, M., Moller, S.: Video quality in next generation mobile networks - perception of time-varying transmission. In: Proceedings of the IEEE International Workshop Technical Committee on Communications Quality and Reliability (CQR), pp. 1–6 (2011)
24. Li, J., Barkowsky, M., Le Callet, P.: Analysis and improvement of a paired comparison method in the application of 3DTV subjective experiment. In: 19th IEEE International Conference on Image Processing (ICIP), pp. 629–632 (2012)
25. Li, J., Barkowsky, M., Le Callet, P.: Subjective assessment methodology for preference of experience in 3DTV. In: IEEE 11th IVMSP Workshop. (2013)
26. Mok, R., Chan, E., Chang, R.: Measuring the quality of experience of http video streaming. In: IFIP/IEEE Int. Symposium on Integrated Network Management (IM), pp. 485–492 (2011)
27. Nosek, B.A. and Open Science Collaboration: Estimating the reproducibility of psychological science. Science, **349**(6251), 1–8 (2015)
28. Rainer, B., Timmerer, C.: Quality of experience of web-based adaptive HTTP streaming clients in real-world environments using crowdsourcing. In: Proc. of the 2014 Workshop on Design, Quality and Deployment of Adaptive Video Streaming, pp. 19–24. ACM (2014)
29. Rainer, B., Waltl, M., Timmerer, C.: A web based subjective evaluation platform. In: Proceedings of the Fifth International Workshop on Quality of Multimedia Experience (QoMEX), pp. 24–25. IEEE (2013)
30. Robinson, D.C., Jutras, Y., Craciun, V.: Subjective video quality assessment of http adaptive streaming technologies. Bell Labs Technical Journal **16**(4), 5–23 (2012)
31. Rodríguez, D.Z., Wang, Z., Rosa, R.L., Bressan, G.: The impact of video-quality-level switching on user quality of experience in dynamic adaptive streaming over HTTP. EURASIP Journal on Wireless Communications and Networking **2014**(1), 1–15 (2014)
32. Rossholm A., Shahid M., Lövström B.: Analysis of the impact of temporal, spatial, and quantization variations on perceptual video quality. In: Proceedings of IEEE Network Operations and Management Symposium (NOMS), pp. 1–5 (2014)
33. Seufert, M., Egger, S., Slanina, M., Zinner, T., Hossfeld, T., Tran-Gia, P.: A survey on quality of experience of HTTP adaptive streaming. IEEE Communications Surveys Tutorials **17**(1), 469–492 (2015)
34. Shahid, M., Rossholm, A., Lövström, B., Zepernick, H.J.: No-reference image and video quality assessment: a classification and review of recent approaches. EURASIP Journal on Image and Video Processing 2014 **2014**(40)
35. Shahid, M., Søgaard, J., Pokhrel, J., Brunnström, K., Wang, K., Tavakoli, S., Gracia, N.: Crowdsourcing based subjective quality assessment of adaptive video streaming. In: Proceedings of the Sixth International Workshop on Quality of Multimedia Experience (QoMEX), pp. 53–54 (2014)

36. Søgaard, J., Pokhrel, J.: Interface template for subjective video experiments using paired comparison. `https://github.com/J-Soegaard/PC-Video-Test-Interface` (2014). [Online; accessed: 20-March-2015]
37. Staelens, N., De Meulenaere, J., Claeys, M., Van Wallendael, G., Van den Broeck, W., De Cock, J., Van de Walle, R., Demeester, P., De Turck, F.: Subjective quality assessment of longer duration video sequences delivered over http adaptive streaming to tablet devices. IEEE Transactions on Broadcasting **60**(4), 707–714 (2014)
38. Tavakoli, S., Brunnström, K., Garcia, N.: About subjective evaluation of adaptive video streaming. In: Proceedings of the Human Vision and Electronic Imaging XX, SPIE Vol: 9394, B. Rogowitz, T. N. Pappas, and H. de Ridder Eds., paper 4(2015)
39. Tavakoli, S., Brunnström, K., Gutiérrez, J., Garcia, N.: Quality of Experience of adaptive video streaming: Investigation in service parameters and subjective quality assessment methodology. Elsevier Signal Processing: Image Communication Vol: 39, part B, pp. 432–443 (2015)
40. Tavakoli, S., Brunnström, K., Wang, K., Andrén, B., Shahid, M., Garcia, N.: Subjective quality assessment of an adaptive video streaming model. In: Proceedings of Image Quality and System Performance XI, SPIE Vol: 9016, S. Triantaphillidou and L. Mohamed-Chaker Eds., paper 20 (2014)
41. Tavakoli, S., Shahid, M., Brunnstrom, K., Lovstrom, B., Garcia, N.: Effect of content characteristics on quality of experience of adaptive streaming. In: Proceedings of the Sixth International Workshop on Quality of Multimedia Experience (QoMEX), pp. 63–64 (2014)
42. Thang, T.C., Nguyen, H., Pham, A., Ngoc, N.P.: Perceptual difference evaluation of video alternatives in adaptive streaming. In: Proceedings of the Fourth International Conference on Communications and Electronics (ICCE), pp. 322–326 (2012)
43. Van Kester, S., Xiao, T., Kooij, R., Brunnström, K., Ahmed, O.: Estimating the impact of single and multiple freezes on video quality. In: Proc. of SPIE-IS&T Human Vision and Electronic Imaging XVI, vol. 7865. B. Rogowitz and T. N. Pappas Eds., paper 25 (2011)
44. Wickelmaier, F., Schmid, C.: A matlab function to estimate choice model parameters from paired-comparison data. Behavior Research Methods, Instruments, & Computers **36**(1), 29–40 (2004)
45. Yao, J., Kanhere, S.S., Hossain, I., Hassan, M.: Empirical evaluation of HTTP adaptive streaming under vehicular mobility. NETWORKING, Springer Berlin Heidelberg, 92–105 (2011)
46. Yen, Yu-Chuan., Chu, Cing-Yu., Yeh, Su-Ling .,Chu, Hao-Hua.,Huang, Polly.:Lab Experiment vs. Crowdsourcing: A Comparative User Study on Skype Call Quality. In: Proceedings of the 9th Asian Internet Engineering Conference, AINTEC, 65–72 (2013)

Author Biographies

Jacob Søgaard received the B.S. degree in engineering, in 2010, and the M.S. degree in engineering, in 2012, from the Technical University of Denmark, Lyngby, where he is currently pursuing his Ph.D. degree with the Coding and Visual Communication group at the Department of Photonics. His research interests include image and video coding, image and video quality assessment, visual communication, and machine learning for Quality of Experience purposes.

Dr. Muhammad Shahid received his PhD in Applied Signal Processing and MSc in Electrical Engineering from Blekinge Institute of Technology, Sweden in 2014 and in 2010 respectively. His research interests include video processing, video quality assessment, and objective and subjective methods of video quality assessment. Dr. Shahid has published over 22 peer reviewed journal and conference research papers.

Jeevan Pokhrel is currently working as a research engineer at Montimage, France. He completed his PhD from Institute Mines Telecom, France in 2014. He received a Dual Master degree from Institute Mines Telecom, France and Asian Institute of technology, Thailand in Communication and Network services (ComNETS) and Information Communication Technology (ICT) in 2011. He has been contributing in some of the European and French research projects. His research is focused on network performance evaluation and security issues. His topics of interest cover performance evaluation, multimedia Quality of Experience (QoE), wireless networks, machine learning etc.

Kjell Brunnström, Ph.D., is a Senior Scientist at Acreo Swedish ICT AB and Adjunct Professor at Mid Sweden University. He is an expert in image processing, computer vision, image and video quality assessment having worked in the area for more than 25 years, including work in Sweden, Japan and UK. He has written a number of articles in international peer-reviewed scientific journals and conference papers, as well as having reviewed a number of scientific articles for international peer-reviewed journals. He has supervised Ph.D. and M.Sc students. Currently, he is leading standardisation activities for video quality measurements as Co-chair of the Video Quality Experts Group (VQEG). His current research interests are in Quality of Experience for visual media in particular video quality assessment both for 2D and 3D, as well as display quality related to the TCO requirements.

Jacob Søgaard

Click here to download Author Photographs Jacob.tif

Muhammad Shahid

Click here to download Author Photographs Muhammad.tif

Jeevan Pokhrel

Click here to download Author Photographs Jeevan.tif

Kjell Brunnström

Click here to download Author Photographs Kjell.tif