©2012 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. The definitive version of this paper has been published in 22nd International Conference Radioelektronika 2012, Special Session on Quality in multimedia systems, 2012.

Impact of Screening Technique on Crowdsourcing QoE Assessments

Bruno GARDLO[1], *Michal RIES*[2], *Tobias HOSSFELD*[3]

[1] Dept. of Telecommunications and Multimedia, University of Zilina, Zilina, Slovak Republic
[2] Dept. of Radio Electronics, Brno University of Technology, Brno, Czech Republic
[3] University of Würzburg, Institute of Computer Science, Würzburg, Germany

gardlo@fel.uniza.sk, ries@feec.vutbr.cz, hossfeld@informatik.uni-wuerzburg.de

Abstract. *Evaluation of quality as perceived by the user in his natural environment is a difficult and strenuous task. Simulation of real world conditions in the laboratory is often inefficient and expensive. Recently, crowdsourcing as novel methodology for testing Quality of Experience (QoE) at the end user side has been proposed. In this paper we discuss (a) the challenges of performing subjective assessments in the crowdsourcing domain and (b) highlight the importance of proper filtering of unreliable users from the overall results. In particular, we introduce various ways for detecting unreliable users and compare results from two similar QoE studies applying different screening techniques.*

Keywords

Subjective Tests, QoE Methodology, Crowdsourcing, Social Networks, Reliability, Screening Techniques.

1. Introduction

The fundamental basis of Quality of Experience (QoE) research are subjective user studies, in which users rate the perceived quality of a service or an application. Typically, such studies are carried out in a laboratory. However, such lab studies are time-consuming and expensive, if the test subjects are paid for participating in the survey. In this context, crowdsourcing emerges as an interesting concept for conducting subjective user tests in a real world environment. The crowdsourcing users conduct the tests remotely at their own computers in a familiar environment. In particular, they launch a web-based application in their browser and click through the subjective test. The basic idea of crowdsourcing is to utilize the huge number of Internet users, such that tasks are completed within short time and at low costs.

Crowdsourcing QoE assessment means to outsource subjective studies to a crowd in the Internet. Since the users are conducting the test remotely, *reliability* of test users is not given which is the major challenge for crowdsourcing QoE assessment. Furthermore, moving the process of quality investigation from the laboratory to the the real-world end-user environment causes a loss of perfect controlled environment in the laboratory domain. As a consequence, crowdsourcing-based QoE tests requires proper *screening techniques* on QoE assessments. This has been already noted before [1], and will be further analyzed in this paper.

The concept of crowdsourcing as novel QoE testing methodology has received strong attention in the QoE research community. Several studies using crowdsourcing or social networking for QoE assessments started to appear recently. Joint effort of University of Würzburg and FTW Vienna resulted e.g. in quantification of QoE for Youtube video streaming [1]. For the subjective experiments, the Microworkers.com platform was used for acquiring of users. In our former study [2], we investigated QoE assessment methodology using Facebook.com social network.

The contribution of this paper is two-fold. First, challenges of performing subjective assessments in the crowdsourcing domain are discussed. Second, the importance of proper screening techniques is shown based on two subjective QoE tests via Facebook. Furthermore, if QoE tests are to be performed in real world conditions, methods have to be defined for controlling these conditions and for controlling the results from such a survey. The detection of unreliable users is very essential, and the lack of proper methodology for screening the users can cause difficulty in evaluation process of the results. In particular, we introduce various ways for detecting unreliable users and compare the results from the two similar QoE studies.

2. QoE and Crowdsourcing

Crowdsourcing represents an efficient tool for performing subjective tests, where representative samples of population are required. Everyday people forms pools of available labour and they are using their spare cycles to create content, solve problems or even do corporate R&D. In 2006 Jeff Howe has defined crowdsourcing as "the act of taking a job traditionally performed by a designated agent (usually an employee) and outsourcing it to an undefined, generally large group of people in the form of an open call".

It has to be differentiated between a) paid crowdsourcing platforms like Microworkers.com or Amazon Mechanical Turk and b) non-paid crowdsourcing platforms where users have different incentives than financial reward, e.g. fun while playing games. Here, social networks can be considered as non-paid crowdsourcing platforms, since they allow creating a large pool of volunteers, who may perform various tasks for free. However, from a methodological viewpoint both types of platforms, i.e. paid and non-paid, lead to the same challenges for conducting QoE tests.

The investigation of quality of experience is a very complex task, since many variables influencing the overall perceived quality have to be taken into account. The influence factors can be grouped into the following three aspects for video streaming exemplarily [3]. I. Context: personal characteristics, environment, content, social and cultural background, etc. II. Expectations: application type, image and brand, usage history, etc. III. Technical system: video codec, end user devices, video delivery mechanisms.

Performing QoE assessments in the laboratory is not only time consuming and expensive, but the design of such test to match "real world" conditions is very difficult. Social networks like Facebook.com allow – with the user permission – to extract various data describing the user's social and cultural backgrounds, or even estimate his environment. Furthermore, technical influence factors can also be derived in the crowdsourcing domain. In summary, crowdsourcing platforms are promising for QoE assessment, since they gather various types of users from every possible environment and with many different habits in computer and Internet usage.

According to latest statistics, there were over 800 million active users on Facebook in February 2012 [4]. This number covers people from all over the world. It has been already established by different studies, that it represents a unique domain for performing miscellaneous research tasks and studies [5, 6, 7]. Recently we also proved, that it is possible to perform QoE assessments within Facebook.com application interface [2]. But, no matter how many advantages crowdsourcing could have, there exists a crucial difference between performing a QoE assessment in the laboratory and performing it in the social network environment. In the laboratory survey we could easily manage to have 100 % of reliable users, whereas this is not realistic in crowdsourcing. Possible reasons for unreliable results include problems in understanding the test correctly, language problems, technical problems to conduct the test (due to insufficient hardware at the end user site or inefficient Internet connection), inattention or tiredness during the test, but also cheating.

In this context, a *reliable user* means, that (a) she is willing to express her true feeling about the perceived quality on a given rating scale, (b) she is neither bored nor distracted with the assessment, and (c) she is rating the quality upon her best conscience. On the other hand, unreliable users are often rating the quality with random or constant grades, not watching the clips or even not doing the survey at all. More important, if they are payed according to time spent in the survey, they are often hurrying to finish it quickly and not paying attention to the test itself. The crucial point of performing QoE studies on crowdsourcing domains is therefore a filtering of these unreliable users (see Section 3), which will be further analyzed in Section 4.

3. Screening Features and Techniques in Social Network QoE Tests

When designing a technical system for QoE surveys in the Facebook.com application interface, we needed a transparent system, which is completely independent from the third party platforms. In [1] they relied on YouTube platform and tested the effect of stalling length on the user's perception. However, third party platforms are limited and if one wants to investigate the influence of various video quality settings on QoE, an independent platform has to be developed. Development of independent platform also enables adjustments to other research needs. By implementing social network functionalities, it is possible to better control the testing subjects and gain various user's informations related to the given testing scenario.

To omit the influence of the transport network, we switched to using the default property of an MP4 format container, where the "moov-atom" needed for playback start is placed at the end of the file. This ensures that the whole file is fully downloaded to the user's computer and only then the playback could be started locally. This enables us a) to use an ordinary web server for providing the video files and b) to use the simple Flash player, which is widely available in modern web browsers. With the Flash player we are also able to monitor the user's behavior and control the playback and playback status. Detection of browser events is essential part for detecting unreliable users and is described below.

Although the methodology design has evolved a lot compared to our previous test [2], the technical system for encoding and serving the videos to the users is almost identical. In both test cases we use 5 different video quality levels with constant audio quality present. The video encoder is using H.264 in High profile and level 4.0 settings. The video codec parameters were adjusted so they comply with the internet streaming demands and in that sense are very similar to settings which also YouTube uses. The video quality in both cases varies from 600 kbit/s to 2400 kbit/s. The test conditions t_x correspond to the video quality levels in both test cases T_1 and T_2, see Table 1. Audio is encoded with AAC codec and 96 kbit/s and 48 kHz. Two video contents are used that is a soccer and an action movie clip of 15 s duration.

Simple Screening. From beginning of this Facebook QoE testing application development we had to cope with the problem, that not all volunteers were evaluating the videos in expected way. The initial filtering was quite simple, but still plays a major role even in the current ongoing campaign. The volunteers were considered as unreliable, if they evaluated all videos with the same grade, or if they did not finish the entire survey. This is referred to as simple screening in this paper which was applied in Test T_1. This simple filtering however leads to rather wide confidence intervals, see the mean opinion scores (MOS) for soccer (S_1) and action (A_1) for the first test T_1 in Figure 1. The results for the test T_2 are denoted as S_2 and A_2, respectively.

Advanced Screening. Hoßfeld et al. introduce various filtering options in [1] which enables to screen unreliable users more efficiently. We adopted some of them and adjusted according to our testing scenario. The filtering of unreliable users can be divided into several groups as listed here:

1. *Application layer monitoring:* Monitoring video player events like fullscreen mode enter and exit time, start time of the playback, end time of the playback, number of pauses in video, pauses positions, etc. Monitoring browser focus time, and time spent on the page.
2. *Control questions concerning the playback:* Questions about the playback itself, whether the movement in the video was jerky, choppy or somehow corrupted. Additional questions about the acceptability of the video quality or the video content.
3. *Consistency questions:* Questions about the content, which user was able to see in the presented clip. We ask the user if she saw objects not presented in the videos, or ask about the sport which was presented. Questions are designed so that user is forced to give precise answer. These consistency questions provide another detection, if the user did even watch the presented clip.
4. *"Gold" data:* Performing two test with the same conditions. Although this doubles the assessment time, it improves the quality by significant amount. By doing double presentation of the same clips, we can also detect and eliminate changes in evaluation in the consecutive test rounds. If the grade in first and second round differs by two or more degrees, than this evaluation is also eliminated from the results.

Additionally, we can compare answers with data extracted from the user's Facebook profile, e.g. age. The important issue here is, not to overshoot the number of the control questions, otherwise the assessment will become to long and boring, and users may leave the survey before its end.

4. Subjective User Results and the Importance of Screening Techniques

The scope of this section is to highlight the importance of proper screening techniques for crowdsourcing QoE assessment. For this purpose, we compare the results from the first survey [2] applying simple screening with results from the ongoing second test, where we applied advanced screening, see Table 1. In the first study T_1, 67 users participated in the assessment and after elimination process 36 of them remained. In the second study T_2, 101 users conducted the survey at the time of writing this paper.

test	screening	test condition t_x, $x \in \{1, \ldots, 5\}$
T_1	simple	$t_x \in \{0.8, 1.1, 1.4, 1.7, 2.0\}$ Mbit/s
T_2	advanced	$t_x \in \{0.6, 1.1, 1.4, 1.8, 2.4\}$ Mbit/s

Tab. 1. Two tests T_1 and T_2 were conducted with similiar test conditions t_x, but different screening techniques.

Screening of the users was performed in several steps. In the first step, simple screening was applied. Thus, we eliminated the users who did not finish the task completely, or did evaluate with constant grades. This resulted into 47 remaining users. In the second step, advanced screening was performed. Depending on the player events, time spent on the page, answers given to consistency questions or other methods described earlier, we considered the user as reliable or not. After this process, only 35 users from the initial 101 users remained for test T_2 according to our strict advanced screening.

Figure 1 shows the mean opinion scores (MOS) of reliable users with corresponding 95% confidence intervals for both test studies. Two different video contents are considered, that are soccer ('S') and action movie ('A'). The label C_i in Figure 1 denotes the user ratings for content $C \in \{A, S\}$ in test T_i. The results show a clear dependance between MOS and video quality level T_x. As expected, MOS raises with increasing video bitrate. The video quality for soccer content class is perceived more critically as compared to the action movie content class. Furthermore, it can be clearly seen that there are significant differences of the MOS for both tests – independent of the type of video content. Furthermore, the (mostly) larger confidence intervals observed in the first test for the different test conditions t_x indicate that not all unreliable users were filtered out. We conclude that the simple screening technique is not sufficient for proper crowdsourcing QoE assessment.

The efficiency of the advanced screening technique is visualized in Figure 2. We consider now the test T_2 and investigate the impact of the screening technique on the MOS. In particular, we average the user ratings over all users $\mathrm{MOS}_{\mathrm{all}}$ and over reliable users only $\mathrm{MOS}_{\mathrm{rel}}$ according to the advanced screening technique. Then, we compute the

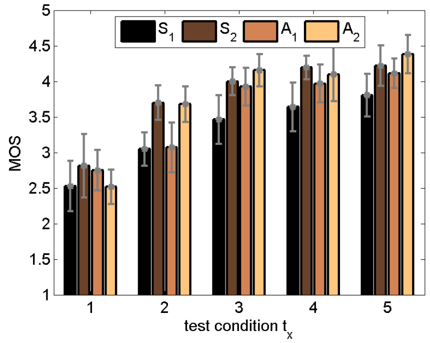

Fig. 1. MOS values of reliable users with corresponding 95% confidence intervals for QoE tests T_1 and T_2.

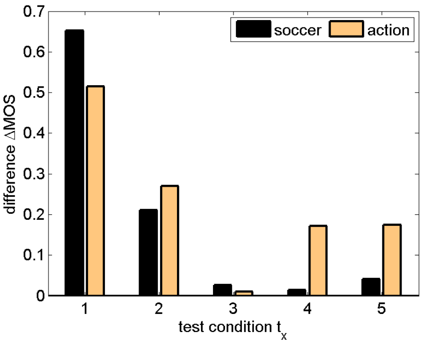

Fig. 2. Difference ΔMOS between average subjective scores with and without advanced screening for test study T_2.

absolute difference, i.e. $\Delta\text{MOS} = \text{MOS}_{\text{all}} - \text{MOS}_{\text{rel}}$. It can be clearly seen that the filtering of unreliable users is an important issue and essential for crowdsourcing QoE assessment. Although for some test conditions the differences are rather low (which may be explained due to saturation effects in rating scales), the strong deviations for other conditions demand for proper filtering of unreliable users.

Finally, we analyse standard deviations of opinion scores (SOS) depending on MOS which follow a square relationship according to [8]. Figure 3 shows the SOS depending on MOS for the second test. Each dot represents the results for a certain test condition t_x; the solid and the dashed lines show the corresponding square function [8] for filtered and not filtered ratings, respectively. It can be seen that the results without screening ('R=0') show a higher standard deviation than the properly filter user ratings ('R=1'). This can be observed for both content types in test T_2, i.e. action ('A2') and soccer ('S2'). Not filtering the data leads to SOS, since the users are rating more or less randomly. This can also been measured in terms of inter-rater reliability which increases by a factor of 1.75 when applying the advanced screening technique.

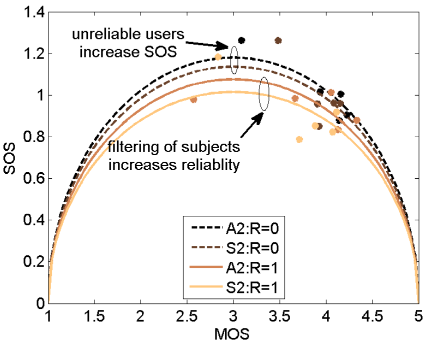

Fig. 3. SOS analysis for test T_2 without (R=0) and with advanced screening (R=1) for soccer ('S') and action ('A').

5. Conclusions

Crowdsourcing is a novel QoE methodology for conducting QoE tests in short time and at low costs. The crowdsourcing setting allows for realistic testing in the natural environment of users and for identification of QoE influence factors beyond technical ones, like the user's social and cultural background – almost impossible in laboratory environments. Due to the remoteness of the subjective test, however, mechanisms and tools are needed to detect the conditions and the environment at the end-user side for proper analysis of QoE influence factors. We have proposed various possibilities on the example of video streaming. The major challenge of crowdsourcing QoE assessment addresses the reliability of users. Screening techniques are essential for proper crowdsourcing QoE assessment. In this work, we conducted two different QoE studies by acquiring test subjects from a social networking application. The analysis of the test results clearly showed the importance of filtering unreliable users out of the data set. Furthermore, we demonstrated that advanced screening techniques are required for proper crowdsourcing QoE assessment.

Acknowledgement

The results introduced in the paper are supported by the project CZ.1.07/2.3.00/20.0007 WICOMT of the operational program Education for competitiveness. This work was partially supported by the COST IC1003 European Network on Quality of Experience in Multimedia Systems and Services - QUALINET (http://www.qualinet.eu/). The research that led to this paper was partially supported by the European Commission under the contract FP7-247688 3DLife.

References

[1] HOSSFELD, T., SEUFERT, M., HIRTH, M., ZINNER, T., TRAN-GIA, P., SCHATZ, R. Quantification of YouTube QoE via Crowdsourcing, in 2011 IEEE Int. Symposium on *Multimedia (ISM)*, dec. 2011, pp. 494 –499.

[2] GARDLO, B., RIES, M., RUPP, M., JARINA, R. A QoE evaluation methodology for HD video streaming using social networking, in 2011 IEEE Int. Symposium on *Multimedia (ISM)*, dec. 2011, pp. 222 –227.

[3] OPTIBAND, Criteria specification for the QoE research, *Optiband Proj. Del. D2.1*, Vienna, June 2010.

[4] SOCIALBAKERS.COM. (2012, February) World continents facebook statistics. [Online]. Available: http://www.socialbakers.com/countries/continents

[5] CHING CHEN, Y. Learning styles and adopting facebook technology, in *Technology Management in the Energy Smart World (PICMET), 2011 Proceedings of PICMET '11:*, 31 2011-aug. 4 2011, pp. 1 –9.

[6] CHU, H.-C., DENG, D.-J., PARK,J. H. Live data mining concerning social networking forensics based on a Facebook session through aggregation of social data, *Selected Areas in Communications, IEEE Journal on*, vol. 29, no. 7, pp. 1368 –1376, august 2011.

[7] FAN, W., YEUNG, K. Virus propagation modeling in Facebook, in *Advances in Social Networks Analysis and Mining (ASONAM), 2010 International Conference on*, aug. 2010, pp. 331 –335.

[8] HOSSFELD, T., SCHATZ, R., EEGGER, S. SOS: The MOS is not enough! in 3rd Int. Workshop on *Quality of Multimedia Experience (QoMEX)*, sept. 2011, pp. 131 –136.