MICROWORKERS VS. FACEBOOK: THE IMPACT OF CROWDSOURCING PLATFORM CHOICE ON EXPERIMENTAL RESULTS

B. Gardlo, M. Ries, T. Hoßfeld, R. Schatz

Dept. of Telecommunications and Multimedia, University of Zilina, Slovakia
Dept. of Radio Electronics, Brno University of Technology, Czech Republic
Institute of Computer Science, University of Würzburg, Germany
Telecommunications Research Center Vienna - FTW, 1220 Vienna, Austria

ABSTRACT

Subjective laboratory tests represent a proven, reliable approach towards multimedia quality assessment. Nonetheless, in certain cases novel progressive quality of experience (QoE) assessment methods can lead to better results or enable test execution in more cost-effective ways. In this respect, crowdsourcing can be considered as emerging method enabling researchers to better explore end-user quality perception when requiring a large panel of subjects, particularly for Web application usage scenarios. However, the crowdsourcing platform chosen for recruiting participants can have an impact on the experimental results. In this paper, we examine the platform's influence on QoE results by comparing MOS scores of two otherwise identical subjective HD video quality experiments executed on one paid and one non-paid crowdsourcing platform.

Index Terms— QoE Assessment, Crowdsourcing, Reliability, Facebook, Microworkers

1. INTRODUCTION

In the past few years, we could witness several important changes within the Internet ecosystem: increasing network speeds, large data stores, and significant improvements of data encoding efficiency and transmission. These all are responsible for raising popularity of the multimedia content and the resulting rapid increase of global audio and video data traffic volumes. In this context, customer satisfaction resulting from ensuring high quality audiovisual experiences is essential for content providers, however, at the same time it is also important to minimize costs of service provision as well. Therefore, QoE-aware end-to-end optimization of content provisioning and distribution is extremely important.

The general concept for evaluating the customer satisfaction with the quality of the provided content are subjective assessments performed in a controlled lab setting [1]. However, a controlled lab environment does not always fully match real world usage contexts. Therefore, alternative testing methods such as crowdsourcing receive growing attention. In the context of QoE assessment, crowdsourcing refers to recruiting selected test user audiences over the internet and letting participants also execute the whole experiment remotely on their PCs. This approach does not only match certain use cases and applications such as online video consumption very well. It also enables cost-effective testing and fast campaign execution by eliminating the expensive requirement of handling physically present participants in an instrumented lab [5].

Recently, several QoE crowdsourcing studies have been conducted. Whether they focus on a specific crowdsourcing platform [2, 3], or they deal with the different approaches to testing methodology or filtering [4, 5], all of them target their focus on evaluation of the QoE at the end user side. The results from these studies are obtained on behalf of different user audiences as a consequence of the different platforms used. Therefore, test results might be influenced by the actual crowdsourcing platform used. In this respected, the intention or willingness of the users to participate in the subjective assessments can be considered as a good point of division. Subjects can either be rewarded with the money for the finished task as it is in the case of Microworkers and Amazon Mechanical Turk, or they can participate voluntarily, perhaps for their own amusement, like it is in the case of Youtube or Facebook social network.

Based on such differentiation, we performed the very same QoE assessment on Facebook social network with volunteering users, but also with the payed workers hired on Microworkers.com crowdsourcing platform. We wanted to explore the influence of the crowdsourcing platform on the overall QoE. The results are compared and presented in this paper and they justify the importance for differentiating between platforms.

2. STUDY DESIGN AND RESULTS

The initial design of the study used for comparing crowdsourcing platforms is based on a previous HD video quality crowdsourcing test [2]. The underlying methodology was fur-

Platform	Users	Age	Female	Nationality
Lab	23	27	39%	UK
Facebook	46	23.5	22%	EU
Microworkers	34	25.5	26%	Asia

Table 1: Users demographic data for different platforms.

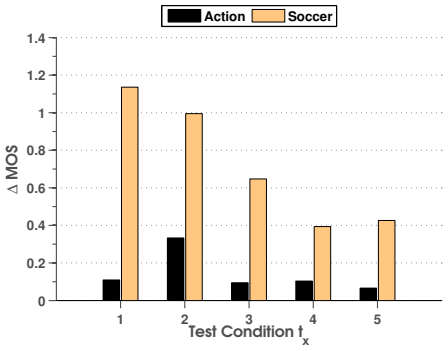

Fig. 1: Absolute differences between crowdsourcing platforms.

ther analyzed and improved [4]; and by introducing additional screening techniques into our former design we achieved reliable platform for performing remote quality surveys. In this work, we continue with the very same technical design and codec settings, using the action movie and soccer content classes. The selection of video bitrates in test conditions $t_1 \sim t_5$ was also justified by the results from the subjective assessment performed in the controlled lab environment [1].

On the Figure 1 are depicted absolute differences between mean opinion scores from Facebook and Microworkers crowdsourcing platforms. These differences were calculated from the MOS obtained by user's rating on the 5-grade MOS scale using ACR method. In all test cases ($t_1 \sim t_5$) paid users from **Microworkers** platform did rate the perceived quality with **higher opinion scores**. The absolute differences are higher for the lower quality test cases and even higher for the soccer content class, which usually is perceived more critically. Reasons for such differences will be examined in the next section.

3. FUTURE WORK

Testing the audiovisual content quality in the user-related environment means taking into account all possible variables influencing the user perception. These variables can either raise or reduce the overall quality perception, and this results into changes in MOS values. This fact is very well visible on Figure 1 and the possible reasons can be clarified by the influence of:

Demographics. A quick look into demographic details (Tab. 1) reveals differences in the tested groups of people. Concerning the Microworkers platform, it gathers very specific users with majority of them coming from the south-east part of Asia. To omit the influence of this "area-specific" behavior, in the future work we will target our focus into regions of Europe and North America, as users from this regions are more similar to the typical users of our interest.

Expectations. The expectations level closely relates to the demographics. Users from different regions may have different expectations about the provided content quality. Especially people who do not use video streaming services often are more tolerant to the worse quality. This might be the case of microworkers since they often access the Internet from the Internet cafès. To omit the influence of such users is also related to the proper environment monitoring.

Environment. Screen resolutions, screen brightness, luminance, or direct sunlight, these variables have strong impact on the perceived quality. Therefore it is important to develop simple test patterns (similar to those used for professional screens calibration), which will enable us to detect improper viewing conditions.

Training session. Even the testing in controlled environment often suffers from corrupted results, when the users were not properly trained for using the whole rating scale. The intention of our methodology was to keep the assessment duration as short as possible, therefore we kept training session very short. However, users from paid crowdsourcing platforms often do not fully understand the given task and are afraid to use the whole rating scale. Therefore it is important to develop new, short and simple training session, which will help to avoid low variances in the users ratings.

All these impacts are specific for user-related environment and for crowdsourcing based QoE assessments. They very well reflects users perception and better reflect the actual quality of experience. However, they also represent new problems, which are yet to be coped with, since they are unknown for testing in the controlled laboratory environment. Therefore new methods for proper monitoring of the environment, and for better understanding of users expectations need to be defined. The results presented in this paper serve as a reference point for improving results and in our future work we will focus on further development of presented crowdsourcing QoE methodology.

4. REFERENCES

[1] ITU-R, "Methodology for the subjective assessment of the quality of television pictures," Rec. BT.500-12, September 2009.

[2] B. Gardlo, M. Ries, M. Rupp, and R. Jarina, "A QoE evaluation methodology for HD video streaming using social networking," in *ISM 2011*, Dana Point, CA, Dec. 2011, pp. 222–227.

[3] M. Hirth, T. Hossfeld, and P. Tran-Gia, "Cost-optimal validation mechanisms and cheat-detection for crowdsourcing platforms," in *IMIS 2011*, Seoul, Korea, July 2011, pp. 316–321.

[4] B. Gardlo, M. Ries, and T. Hoßfeld, "Impact of screening technique on crowdsourcing QoE assessments," in *Radioelektronika 2012*, Brno, April 2012, pp. 55–58.

[5] T. Hoßfeld, M. Seufert, M. Hirth, T. Zinner, P. Tran-Gia, and R. Schatz, "Quantification of YouTube QoE via crowdsourcing," in *ISM 2011*, Dana Point, CA, Dec. 2011, pp. 494–499.