Survey of Web-based Crowdsourcing Frameworks for Subjective Quality Assessment

Tobias Hoßfeld [1*], Matthias Hirth [1], Pavel Korshunov [2], Philippe Hanhart [2], Bruno Gardlo [3]
Christian Keimel [4], Christian Timmerer [5]

[1] *University of Würzburg, Institute of Computer Science, Chair of Communication Networks, Germany*
[*] tobias.hossfeld@uni-wuerzburg.de

[2] *Multimedia Signal Processing Group, Ecole Polytechnique Fédérale de Lausanne - EPFL, Switzerland*

[3] *Telecommunications Research Center Vienna - FTW, Austria*

[4] *Technische Universität München, Institute for Data Processing, Germany*

[5] *Multimedia Communication, Alpen-Adria-Universität Klagenfurt, Austria*

***Abstract*—The popularity of the crowdsourcing for performing various tasks online increased significantly in the past few years. The low cost and flexibility of crowdsourcing, in particular, attracted researchers in the field of subjective multimedia evaluations and Quality of Experience (QoE). Since online assessment of multimedia content is challenging, several dedicated frameworks were created to aid in the designing of the tests, including the support of the testing methodologies like ACR, DCR, and PC, setting up the tasks, training sessions, screening of the subjects, and storage of the resulted data. In this paper, we focus on the web-based frameworks for multimedia quality assessments that support commonly used crowdsourcing platforms such as Amazon Mechanical Turk and Microworkers. We provide a detailed overview of the crowdsourcing frameworks and evaluate them to aid researchers in the field of QoE assessment in the selection of frameworks and crowdsourcing platforms that are adequate for their experiments.**

I. Introduction

Crowdsourcing is an increasingly popular approach for employing large numbers of people for performing short and simple online tasks. Several commercial crowdsourcing platforms provide online workers with varying cultural and social backgrounds from around the world. Since typical payment for a crowdsourcing job is small and, often, is less than a dollar, crowdsourcing can be a powerful and cost effective tool for performing work that can be easily divided into a set of short and simple tasks, such as surveys, image tagging, text recognition, and viral campaigns.

Subjective quality assessment or QoE assessment of multimedia is another task suitable for crowdsourcing. A typical subjective test consists of a set of repetitive tasks and, hence, can be easily implemented using the crowdsourcing principle. In particular, the cost effectiveness and access to a large pool of test subjects makes crowdsourcing an attractive alternative to lab-based evaluations. Therefore, researchers in quality assessment increasingly use crowdsourcing in various research areas, including rebuffering in streaming video [1], aesthetics of images [2], emotional reaction caused by image content [3], quality assessment of 3D video [4], privacy issues in HDR images [5], or audio quality [6].

Performing multimedia-based assessments online, however, is challenging. Such assessments often include video or images, sometimes even uncompressed, resulting in a large amount of data that is not only difficult to transmit to workers with slow network connections, but also to display on low resolution screens used by some workers. Moreover, crowdsourcing provides little control over the environments of the workers compared to the dedicated test labs, which are usually equipped in compliance with recommendations like ITU-R BT.500 [7]. Unreliable workers can also affect the repeatability of results [8]. Therefore, for a practical use, crowdsourcing-based subjective assessment of multimedia requires an additional set of tools and utilities.

In this paper, we provide an overview of the existing crowdsourcing platforms and web-based frameworks that aid in multimedia quality assessments. We briefly describe well-known crowd providers, such as Mechanical Turk and Microworkers, as well as aggregator platforms (Crowdflower and Crowdsource) and specialized platforms (Microtask and Taskrabbit). We then focus the discussion on crowdsourcing frameworks, assuming that they provide tools for developing and running subjective quality assessment experiments in a web browser, considering only those frameworks that support typical web-based crowdsourcing platforms, which means mobile agents designed to help with subjective evaluations [9] or OS-specific desktop implementations [10] are excluded.

The rest of the paper is organized as follows. Section II summarizes crowdsourcing in general and discusses existing crowdsourcing platforms. Section III gives a detailed overview of the existing crowdsourcing frameworks designed for multimedia quality assessments. Section IV compares these frameworks and concludes the paper.

2014 IEEE 16th International Workshop on Multimedia Signal Processing (MMSP), Sep. 22-24, 2014, Jakarta, Indonesia.
978-1-4799-5896-2/14/$31.00 ©*2014 IEEE.*

II. Background on Crowdsourcing

Before discussing crowdsourcing frameworks, this section briefly introduces the general principle of crowdsourcing, crowdsourcing related terminology, and crowdsourcing platforms. It also discusses the possible benefits and challenges of using crowdsourcing for subjective assessments, followed by a motivation why dedicated frameworks are needed.

A. Principle of Crowdsourcing

In contrast to traditional recruiting processes, where dedicated employees are selected and assigned to tasks by an employer, in crowdsourcing, the *employer* submits the task as an open call to a large anonymous crowd of *workers*. The workers can then freely decide which available task they want to work on. Usually these tasks have a smaller granularity than traditional forms of work organization and are highly repetitive, such as labeling large number of images. The tasks are usually grouped in larger units, referred to as *campaigns*. Maintaining a dedicated worker crowd, including the required infrastructure, is usually not feasible for most employers and therefore mediators are used to access the crowd, termed *crowdsourcing platforms*. These platforms abstract the crowd to a certain extent, but sometimes also provide additional services, e.g., quality control or worker selection mechanism.

B. Possibles Benefits and emerging challenges

Subjective quality assessments require human participation as the judgments are based on a personal opinion. Traditionally, these assessments are performed in a controlled lab environment with selected paid or voluntary participants.

Crowdsourcing offers the possibility to conduct web-based tests with participants from all over the world. Such flexibility enables a faster completion compared to traditional forms of assessment as more potential participants are available. It can help to reduce the costs of the experiments, since no dedicated test lab is required. It also helps to create a realistic test environment, as the assessment is done directly on the participants' device. The diversity of the test participants helps to avoid possible biases caused by the limited number of participants in traditional lab tests.

However, crowdsourced quality assessment introduces new challenges. The test implementation has to be more robust to cope with different end user devices used by the participants and surrounding conditions. For example, a noise during audio tests has to be detected and, if possible, considered in the evaluation of the results. Besides these technical aspects, interaction with the test participants in a crowdsourcing campaign is different compared to a lab-based test. The lab environment allows personal interactions, such as giving detailed training, clarifications of unclear instructions, using visual aids, etc., while in the crowdsourcing-based quality assessment, the participants usually remain anonymous and instructions are provided in a written form only. Additionally, some workers may perform the tasks sloppily or even cheat if it helps to maximize their income.

C. Existing Crowdsourcing Platforms

Currently, a number of commercial crowdsourcing platforms are available, and Fig. 1 illustrates that different platforms target different use cases.

Aggregator platforms like Crowdflower[1] or Crowdsource[2] can be considered as the most high-level type of crowdsourcing platforms. They often do not maintain their own workforce, but delegate the tasks to different channels that provide the actual workers. Aggregator platforms use a very high abstraction layer, i.e., they provide means to upload input data for the crowd task, adjust the quality level of the results, and download the post-processed results. Therefore, they usually focus on a limited set of predefined tasks only, such as image tagging. Some of the platforms also offer the implementation of custom solutions, mainly targeting business customers aiming to integrate crowd-based solutions into existing work flows. The high abstraction is the major drawback of these platforms with respect to subjective quality assessment, as some aspects of the experiment might not be directly controllable [11]. Due to internal, platform-specific recruiting mechanisms, the available workers might already be pre-filtered, also limiting their diversity. Furthermore, common quality assurance methods are usually not applicable for the quality control of subjective assessments.

Specialized platforms only focus on a limited set of tasks or on a certain worker type. Two examples of this class are Microtask[3], specializing in document processing and data entry, and Taskrabbit[4], focusing on location based crowdsourcing services. In contrast to aggregator platforms, specialized platforms maintain their own workforce, but also offer quality assurance and specific management tools tailored for the platforms main use case. Such focus means that these platforms, in general, are not suitable for subjective assessments.

Crowd providers are the most flexible type of crowdsourcing platforms. Representatives of this class are Amazon Mechanical Turk (MTurk)[5], Microworkers[6], and TaskCN[7]. These platforms focus mainly on self-service and maintain a large worker crowd. To simplify the access to the crowd workers, the platforms often provide filter mechanisms to select workers base on location or predefined skills, and also implement APIs to allow the automated generation and management of tasks. Crowd provider platforms are the most suitable option for subjective quality assessments, because they offer the most direct and unfiltered access to the recruited participants. This flexibility enables a fine granular filtering and the identification of possible biases, e.g., due to the cultural background of the workers. Also, crowd providers usually offer flexible interfaces to design individual or experimental tasks, required for the individual assessment tests. However, due to the vast variety

[1] http://crowdflower.com
[2] http://www.crowdsource.com
[3] http://www.microtask.com
[4] http://www.taskrabbit.com
[5] http://mturk.com
[6] http://microworkers.com
[7] http://www.taskcn.com

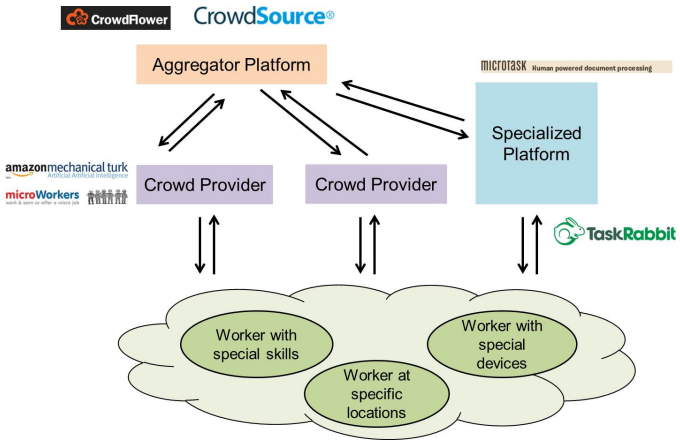

Fig. 1. Overview on types of crowdsourcing platforms.

of tasks on these platforms, implementation of global quality assurance mechanisms is generally not possible. It is usually left to the employer to implement an appropriate task design and to maintain the required quality standards.

Considering crowdsourcing as a more general concept, social networks like Facebook can also be considered as crowd provider. In contrast to paid crowdsourcing platforms, the users from social networks are not motivated primarily by monetary incentives. Therefore, the task design has to be different compared to commercially recruited user, e.g., less emphasis has to be put on cheat detection but more on issues like gamification. Due to the different requirements of the task design, social networks are not considered in this paper as crowd providers for QoE assessment.

Crowdsourcing frameworks for subjective quality assessment help to overcome the drawbacks of the crowd provider platforms by providing tested task design, post-processing, and evaluation mechanisms. The frameworks remove the burden of re-implementing solutions for simple tasks in a quality assessments test, for example, implementing a rating scale, or possibly even more complex problems, like playing YouTube videos with predefined impairment independent of the participants network connection.

III. Existing Crowdsourcing Frameworks

Web-based crowdsourcing frameworks for multimedia quality assessment represent a conceptual approach with programming tools to develop subjective experiments that can be executed in a web browser. In particular, such frameworks allow multimedia content to be displayed in a browser for subjects to evaluate the quality using web forms. The test logic may be implemented at the client-side, e.g., javascript or at the server-side, e.g., PHP. Also frameworks allow to enable the execution of the experiments utilizing typical crowd-provider platforms. The basic functionality of a framework includes (a) the creation of the test (by supporting common testing methodologies like ACR, DCR, PC), (b) the execution of the test (by supporting training, task design, task order, screening), and (c) the storage and access to the result data.

A. Quadrant of Euphoria

The *Quadrant of Euphoria* is a web-based online crowdsourcing platform for the QoE evaluation of audio, visual, and audio-visual stimuli proposed by Chen *et al.* in [12], [13] and extended by Wu *et al.* in [14] .

It allows for a pairwise comparison of two different stimuli in an interactive web-interface, where the worker can judge which of the two stimuli has a higher QoE. Additionally, the platform provides some rudimentary reliability assessment based on the actual user ratings under the assumption that the preferences of users is a transitive relation, expressed by the Transitivity Satisfaction Rate (TSR): If a user prefers the test condition A to B and B to C, the user will also prefer A to C. If this condition is not met for a certain number of triplets and the TSR is below a given threshold e.g. 0.8 as suggested in [12], the user is rejected. Expressed more formally, the TSR is defined as the number of judgment triplets satisfying transitivity divided by the total number of triplets where transitivity may apply.

The performance of the reliability assessment is similar to the crowdMOS [6] approach, described in Section III-B, as a large ratio of fake users is accepted while some reliable users are rejected [8]. Similarly to crowdMOS, the threshold value can be fine-tuned to reduce the acceptance of fake users, but at the cost of an increased rejection of reliable users.

B. CrowdMOS

The *crowdMOS* framework for subjective user studies was proposed by Ribeiro *et al.* [6] and is an open-source project that can be installed and modified with relatively low effort on any suitable web server.

Originally, *crowdMOS* was focused only on subjective audio testing and implemented two subjective audio quality assessment methodologies: Absolute Category Rating (ACR) from ITU-T P.800 [15] and MUSHRA from ITU-R BS.1534-1 [16]. Later it was extended in [17] to image quality assessment with the absolute category rating (ACR) for video from ITU-T P.910 [18]. For assessing the reliability of users, the sample correlation coefficient between the average user rating of a worker and the global average rating is used. Users are rejected if the correlation coefficient is below a certain threshold e.g. 0.25 as suggested in [6]. After users are rejected, the global average rating is recomputed for the remaining users and the correlation coefficient is determined again. Users are ranked in decreasing order of the correlation coefficient and the user screening starts again. A large fraction of fake users, however, is accepted, which can be reduced by increasing the threshold. But an increased threshold would result in an even larger ratio of reliable users rejected. This issue with this rejection algorithm is discussed in detail in [8].

C. QualityCrowd

The *QualityCrowd* [19] framework is a complete platform designed especially for QoE evaluation with crowdsourcing by Keimel et al. It is an open-source project that can be installed and modified with relatively low effort on any suitable web

server. It aims at providing the necessary tools to conduct subjective quality assessment tests, so that the focus can be set on the test design and not on the implementation of the test environment.

QualityCrowd consists of two parts: A front-end, which is presented to the test subject and where the actual test takes place and a back-end which the supervisor can use to create new test campaigns and collect the test results. The framework provides a multitude of different options for the test design and a test can consist of any number of questions and can contain videos, sounds or images or any combination. Moreover, it allows the use of different testing methodologies, e.g., single stimulus or double stimulus, and different scales, e.g., discrete or continuous quality or impairment scales. In its latest iteration QualityCrowd2[8], a simple scripting language has been introduced that allows for the creation of test campaigns with high flexibility, by not only enabling the combination of different stimuli and testing methodologies, but also by the possibility to specify training sessions and/or introduce control questions for the identification of reliable user ratings in order to ensure high data quality.

D. Web-based Subjective Evaluation Platform (WESP)

Rainer et al. describe a Web-based subjective evaluation platform (WESP) in [20] which is based on the ITU recommendations for subjective quality evaluations of multimedia and television [18], [21]. The platform was initially developed for subjective quality assessments of sensory experience but can also be used for general-purpose QoE assessments. WESP can be integrated into any crowd provider platform as long as there is support for embedding external web sites within a crowdsourced task e.g. Mechanical Turk and Microworkers. The evaluation framework is open source and available for download[9].

WESP provides a management and presentation layer: the former is used to configure the subjective quality assessment according to its requirements and goals, and the latter is responsible for the presentation of the actual user study and provides a specific view based on the configuration defined within the management layer. The management layer allows the configuration of each component e.g. pre-questionnaire, voting mechanism, rating scale, and control questions, independently and thus provides enough flexibility for a wide range of different methodologies e.g., single stimulus, double stimulus, pair comparison or continuous quality evaluation. Additionally, any new methodology can be implemented through the management layer. The presentation layer presents the content to the participants and is based on standard HTML elements. In particular, it allows the collection of explicit and implicit user input: the former is data entered by the user via explicit user input elements e.g. voting using a slider for a given rating scale, compared to the latter describing implicit input represented by data from the browser window e.g. window focus or duration of the test. Video content is presented using HMTL5 or Flash, either explicitly enforced via the management layer or determined automatically by the user agent. Javascript can be added if needed and plugins can be added for specific input/output hardware requirements in a lab environment e.g. 3D, haptics, sensory effects or electroencephalography.

E. BeaqleJS

BeaqleJS framework is developed for subjective audio studies by Kraft and Zölzer [22]. It is written in Javascript and PHP, and HTML5 is used to playback the audio clips[10]. Several audio formats are supported, including an uncompressed WAV PCM format, which is important for subjective audio tests. The framework allows implementation of different testing methodologies via some simple code extensions, with two evaluation methodologies already implemented: a simple ABX methodology[11] and MUSHRA defined in ITU-R BS.1534-1 [16]. Currently, there is no support for workers reliability detection and evaluations results are emailed to the organizer of an audio evaluation in a text file.

F. In-momento Crowdsourcing

Current approaches for crowdsourcing-based quality assessment often aim to ensure reliability of remote participants by introducing reliability screening questions throughout the test and these questions are usually analyzed a-posteriori [1]. This leads to relatively reliable ratings, but due to the strict a-posteriori filtering also produces a large amount of unusable ratings by participants labelled as unreliable and also results in additional administrative work.

Gardlo *et al.* [23] therefore introduced the *in-momento* crowdsourcing framework, combining careful user-interface design together with the best known practices for QoE crowdsourcing tests [8]. Instead of a-posteriori data analysis and subsequent removal of unreliable data, this framework aims at live or *in-momento* evaluation of the user's behaviour: as the user proceeds with the assessment, the reliability of the user is continuously updated and a reliability profile is built.

This reliability assessment utilizes a two stage design. It avoids questions targeted towards the shown content as well as repetitive questions for cheating detection. In contrast, for each stage of the test suspicious behavior is defined that is subsequently monitored in a background process e.g. focus time, video playback time, full-screen mode or switching to different browser window, etc. If suspicious behaviour is detected, the user is assigned penalty points, used to compute the overall reliability of the respective user. To balance between campaign speed, reliability of the results and users' enjoyment during the test, the assessment time is kept as short as possible and users are able to quit the assessment at any point unlike in other frameworks. The aim is to avoid forcing to continue with the test even though they are bored or lost the interest, as these two issues are closely related to unreliable behavior.

[8]https://github.com/ldvpublic/QualityCrowd2
[9]http://selab.itec.aau.at/
[10]https://github.com/HSU-ANT/beaqlejs
[11]http://home.provide.net/~djcarlst/abx.htm

TABLE I
COMPARISON OF CROWDSOURCING FRAMEWORKS FOR QOE ASSESSMENT.

Framework / Feature	Euphoria [12]	CrowdMOS [6]	QualityCrowd2 [19]	WESP [20]	BeaqleJS [22]	*in-momento* [23]
Media types	Image, video & audio	Image, audio	Image, video & audio	Image, video, audio, sensory effects	Audio	Image, video
Methodology	PC (binary scale)	ACR, DCR, MUSHRA	ACR, flexible: single & double stimulus; discrete & continuous scales	All (flexible), e.g., ACR, ACR-HR, DSCQE, Double stimulus for sensory effects	ABX, MUSHRA	ACR
Questionnaires	None	Embedded in evaluation	Separated tasks	Embedded in evaluation	None	None
Tasks design	Fixed template	Custom template All tasks have the same template	Custom template Tasks configured in script file	All tasks have the same template	Fixed template	Fixed template
Tasks order	Random All pairs	Random Full set or subset of all stimuli	Fixed	Flexible	Fixed	Random Based on actual number of ratings
Screening	Transitivity index	95% CIs	None	None	None	Reliability profile
Data storage	Text files	Text files	Text files CSV format	Database	Text files	Database
Open source	No[12]	Yes[13]	Yes[14]	Yes[15]	Yes [16]	Yes[17]
Programming language	N/A	Ruby	PHP + own script language	Javascript + PHP	Javascript + PHP	PHP

If the reliability score of a user drops for whatever reason, the user can claim his current reward, but is not able to continue with the test. On the other hand, users who finish the complete assessment with high reliability score receive a bonus payment for good work.

Therefore, the *in-momento* approach utilizes the potential of the huge worker pools nowadays available on crowdsourcing platforms with the in-momento verification of the user's reliability and dynamic task offer. Since the reliability profile is known at each stage of the assessment, it is possible to offer reliable users during the test additional tasks for an increased reward.

G. Other Tools and Approaches

Besides the frameworks mentioned above that aim to provide complete solutions for subjective assessment, also a number of other tools and approaches exist that are focused on specific problems encountered during assessment tests. They can be used in the development of a new test setup or might be added as features in existing frameworks.

A specialized approach to detect workers clicking randomly in crowdsourcing-based studies is presented by Kim *et al.* in [24]. They use Pearson's χ^2 test with the null hypotheses that the users are clicking randomly. The resulting p-value is used for excluding users with a p-value above a certain threshold. Hoßfeld et al. [8] showed that this methodology clearly reveals random clickers, but also rejects many reliable user ratings. This may be caused by the fact that users cannot differentiate the impact of the test conditions or perceive some test conditions equally, e.g. due to the used end-user device. Here, tools for automatically detecting, e.g., the screen quality of the test participant[18], can help to shed light on the reasons for the inconsistent ratings.

Another issue can arise from the rating mechanisms used during the assessment. While the commonly used five point MOS scale requires a training of the test participants to correctly evaluate the upper and lower baseline of the available samples, a pair-wise comparison is usually easier to understand. However, a pair-wise comparison of huge data sets is usually not possible, but approaches like the HodgeRank on Random Graphs [25] can be used to derive results from incomplete and imbalances comparison sets, extended in [26] for crowdsourcing-based assessment during which comparisons are produced consecutively. This version of the HodgeRank on Randoms Graphs is able to update the results online instead of working batch wise on the complete sample test.

IV. DISCUSSION AND CONCLUSIONS

Using crowdsourcing to conduct subjective quality assessments appears to be a promising way to quickly collect a large number of realistic test results. However it imposes new and different challenges compared to similar tests in a lab environment.

The first challenge is to find an appropriate pool of participants for the test and a crowd provider providing a flexible enough interface to run the experimental tasks. The second major challenge is the delivery of the test to the participants. It is often necessary to redesign the test to a web-based version which allows the access for the globally distributed workers

[12] http://mmnet.iis.sinica.edu.tw/proj/qoe/
[13] http://crowdmos.codeplex.com
[14] https://github.com/ldvpublic/QualityCrowd2
[15] http://selab.itec.aau.at/
[16] https://github.com/HSU-ANT/beaqlejs
[17] http://im.dataworkers.eu/
[18] https://github.com/St1c/screentest/

and does not require the workers to install any software on their device. During this process a significant software development effort is needed that can be reduced significantly by using an existing framework. The most suitable framework is determined based on specific criteria e.g. the type of media to evaluate, the experiment design, and the server environment. Table I presents a comparison of the available frameworks according to specific features serving as a guideline for this selection process.

Several methodologies have been designed for subjective quality assessment. Some of them are quite generic e.g. absolute category rating (ACR), but specific methodologies have also been designed in particular for listening tests. However, frameworks often only implement a limited set of methodologies. For example, Quadrant of Euphoria only implements paired comparison (PC) and cannot be modified to support other methodologies.

As crowdsourcing provides little control over the environment, best practices recommend to include control questions and screening strategies to detect cheaters and outliers. Many frameworks allow inserting additional questionnaires. Nevertheless, only QualityCrowd provides means to ask questions as separated tasks, as recommended by best practices, whereas the other frameworks only allow including questions inside the template used for the evaluation task.

Crowdsourcing is often used to evaluate large datasets, requiring weeks of lab evaluations. In this case, a random subset of all stimuli is presented to each worker. Some frameworks provide means to cope with large dataset and randomly distribute the workload, but most frameworks only consider a fixed list of tasks. Therefore, in this case, the workaround is to design several campaigns, each with different tasks, and to implement a tool to assign a different campaign to each worker.

As each crowdsourcing experiment is somewhat unique, it is very difficult to find a framework that can be used directly without any modification. Of course, a new framework can be developed from scratch, but this requires a lot of programming effort, especially to include all the necessary data checks and anti-cheating mechanisms. Therefore, using an existing framework as a starting base and modifying it to fit the requirements of the experiment design is a more sensible alternative.

ACKNOWLEDGMENT

This work is supported by the COST Action IC1003 Qualinet and by the Deutsche Forschungsgemeinschaft (DFG) under Grants HO4770/2-1 and TR257/38-1. The authors alone are responsible for the content.

REFERENCES

[1] T. Hoßfeld, M. Seufert, M. Hirth, T. Zinner, P. Tran-Gia, and R. Schatz, "Quantification of YouTube QoE via crowdsourcing," in *Symposium on Multimedia*, Dana Point, USA, Dec. 2011.

[2] J. Redi, T. Hoßfeld, P. Korshunov, F. Mazza, I. Povoa, and C. Keimel, "Crowdsourcing-based multimedia subjective evaluations: A case study on image recognizability and aesthetic appeal," in *Workshop on Crowdsourcing for Multimedia*, Barcelona, ES, Oct. 2013.

[3] I. Hupont, P. Lebreton, T. Mäki, E. Skodras, and M. Hirth, "Is it possible to crowdsource emotions?" in *International Conference on Communications and Electronics*, Da Nang, VN, Jul. 2014.

[4] P. Hanhart, P. Korshunov, and T. Ebrahimi, "Crowd-based quality assessment of multiview video plus depth coding," in *IEEE International Conference on Image Processing, ICIP'2014*, Paris France, Apr. 2014.

[5] P. Korshunov, H. Nemoto, A. Skodras, and T. Ebrahimi, "Crowdsourcing-based evaluation of privacy in HDR images," in *SPIE Photonics Europe 2014, Optics, Photonics and Digital Technologies for Multimedia Applications*, Brussels, Belgium, Apr. 2014.

[6] F. Ribeiro, D. Florencio, C. Zhang, and M. Seltzer, "CrowdMOS: An approach for crowdsourcing mean opinion score studies," in *International Conference on Acoustics, Speech and Signal Processing*, Prague, CZ, May 2011.

[7] *ITU-R BT.500 Methodology for the Subjective Assessment of the Quality for Television Pictures*, ITU-R Std., Rev. 12, Sep. 2009.

[8] T. Hossfeld, C. Keimel, M. Hirth, B. Gardlo, J. Habigt, K. Diepold, and P. Tran-Gia, "Best practices for qoe crowdtesting: QoE assessment with crowdsourcing," *Transactions on Multimedia*, vol. 16, no. 2, Feb 2014.

[9] F. Wei, J. Xu, T. Wen, X. Liu, and H. Yan, "Smart phone based online QoE assessment for end-to-end multimedia services on 3G mobile Internet," in *Consumer Electronics, Communications and Networks*, Yichang, CN, Apr. 2012.

[10] A. K. Jain, C. Bal, and T. Q. Nguyen, "Tally: A web-based subjective testing tool," in *Workshop on Quality of Multimedia Experience*, Klagenfurth, AU, Jul. 2013.

[11] Ó. Figuerola Salas, V. Adzic, A. Shah, and H. Kalva, "Assessing internet video quality using crowdsourcing," in *Workshop on Crowdsourcing for multimedia*, Barcelona, ES, Oct. 2013.

[12] K.-T. Chen, C.-C. Wu, Y.-C. Chang, and C.-L. Lei, "A crowdsourceable QoE evaluation framework for multimedia content," in *Multimedia*, Beijing, CN, Oct. 2009.

[13] K.-T. Chen, C.-J. Chang, C.-C. Wu, Y.-C. Chang, and C.-L. Lei, "Quadrant of euphoria: A crowdsourcing platform for QoE assessment," *Network*, vol. 24, no. 2, Mar. 2010.

[14] C. Wu, K. Chen, Y. Chang, and C. Lei, "Crowdsourcing multimedia QoE evaluation: A trusted framework," *Transactions on Multimedia*, vol. 15, no. 99, Jul. 2013.

[15] *ITU-T P.800 Methods for subjective determination of transmission quality*, ITU-T Std., Aug. 1996.

[16] *ITU-R BS.1534-1 Method for the subjective assessment of intermediate quality levels of coding systems*, ITU-R Std., Oct. 2003.

[17] F. Ribeiro, D. Florencio, and V. Nascimento, "Crowdsourcing subjective image quality evaluation," in *Image Processing*, Brussels, BE, Sep. 2011.

[18] *ITU-T P.910 Subjective video quality assessment methods for multimedia applications*, ITU-T Std., Apr. 2008.

[19] C. Keimel, J. Habigt, C. Horch, and K. Diepold, "QualityCrowd - a framework for crowd-based quality evaluation," in *Picture Coding Symposium*, Krakow, PL, May 2012.

[20] B. Rainer, M. Waltl, and C. Timmerer, "A web based subjective evaluation platform," in *Workshop on Quality of Multimedia Experience*, Klagenfurth, AT, Jul. 2013.

[21] ITU Radiocommunication Assembly, "ITU-R BT.500-12 Methodology for the subjective assessment of the quality of television pictures ," 2009.

[22] S. Kraft and U. Zölzer, "BeaqleJS: HTML5 and JavaScript based framework for the subjective evaluation of audio quality," in *Linux Audio Conference*, Karlsruhe, DE, May 2014.

[23] B. Gardlo, S. Egger, M. Seufert, and R. Schatz, "Crowdsourcing 2.0: Enhancing execution speed and reliability of web-based QoE testing," in *International Conference on Communications*, Sydney, AU, Jun. 2014.

[24] S.-H. Kim, H. Yun, and J. S. Yi, "How to filter out random clickers in a crowdsourcing-based study?" in *BELIV Workshop: Beyond Time and Errors - Novel Evaluation Methods for Visualization*, Seattle, USA, Oct. 2012.

[25] Q. Xu, Q. Huang, T. Jiang, B. Yan, W. Lin, and Y. Yao, "HodgeRank on random graphs for subjective video quality assessment," *Transactions on Multimedia*, vol. 14, no. 3, Jun. 2012.

[26] Q. Xu, Q. Huang, and Y. Yao, "Online crowdsourcing subjective image