QualityCrowd – A Framework for Crowd-based Quality Evaluation

Christian Keimel, Julian Habigt, Clemens Horch and Klaus Diepold
Technische Universität München, Institute for Data Processing
Arcisstr. 21, 80333 Munich, Germany
christian.keimel@tum.de, jh@tum.de, ch@tum.de, kldi@tum.de

***Abstract*—Video quality assessment with subjective testing is both time consuming and expensive. An interesting new approach to traditional testing is the so-called crowdsourcing, moving the testing effort into the internet. We therefore propose in this contribution the QualityCrowd framework to effortlessly perform subjective quality assessment with crowdsourcing. QualityCrowd allows codec independent quality assessment with a simple web interface, usable with common web browsers. We compared the results from an online subjective test using this framework with the results from a test in a standardized environment. This comparison shows that QualityCrowd delivers equivalent results within the acceptable inter-lab correlation. While we only consider video quality in this contribution, QualityCrowd can also be used for multimodal quality assessment.**

I. Introduction

Quality assessment of video is usually done with subjective testing, as no universally accepted objective quality metrics exist, yet. Subjective testing, however, is both time consuming and expensive. On the one hand this is caused by the limited capacity of the laboratories due to both the hardware and the requirements of the relevant standards e.g. [1], on the other hand by the reimbursement of the test subjects, that needs to be competitive to the general wage level at the laboratories location, in order to be able to hire enough qualified subjects.

But do we really need to perform subjective tests in a laboratory? An alternative to the classical approach to subjective testing is crowdsourcing. Crowdsourcing is a relatively new concept, that uses the internet to assign simple tasks to a group of online workers and has recently become quite popular in social sciences [2]. Hence we no longer perform our tests in a standard conforming laboratory, but conduct them via the internet with participants from all over the world. This not only allows us to recruit the subjects from a larger, more diverse group, but also to reduce the financial burden significantly. Of course, we will loose some control over the test setup, but in turn we gain more subjects, leading to a more representative sample of the general population.

We therefore introduce in this contribution the *QualityCrowd* framework, a web-based platform for video quality evaluation with crowdsourcing. Our proposed framework is codec agnostic, as we deliver the videos under test via lossless compression to the participants, allowing us to assess the visual quality not only of existing coding technology, but also of future developments. Moreover, this framework requires no special software on the participants' side and works in common web browsers e.g. Firefox or Internet Explorer. Additionally, it implements standardized single stimulus and double stimulus testing methodologies. To demonstrate the framework's feasibility, we will perform an online subjective test in a local network environment with it.

In related works, Paolacci et al. examined in [3] if the results gained from crowdsourced experiments are comparable to results from traditional experiments in general and concluded that crowdsouring is a valid alternative. More related to subjective testing, Marge et al. have shown in [4] that crowdsourcing delivers similar results to traditional methods for audio transcription. Chen et al. conducted subjective audiovisual tests via crowdsourcing in [5], [6], but used a non-standardized testing methodology and MP3 and H.264/AVC for compression. Finally, Ribeiro et al. presented the *crowdMOS* framework in [7], [8], implementing standardized testing methodologies for both audio and still images, but do not provide lossless content delivery to the test subjects and thus are limited to current coding technologies.

This contribution is organized as follows: after a short introduction into the concept of crowdsourcing, we present our QualityCrowd framework, before continuing to a comparison of results gained with from QualityCrowd to the results from lab tests. Finally, we conclude with a short summary.

II. Crowdsourcing

The term *Crowdsourcing* has first been coined by Howe in [9]. It is a neologism from the words *crowd* and *outsourcing* and describes the transfer of services from professionals to the public via the internet. These services often consist of tasks which cannot or not efficiently be solved by computers but are simple enough to be performed by non-trained workers, e.g. tagging photos with meaningful key words. However, even rather complex services can be crowdsourced, like creative tasks such as the generation of new business ideas [10], all kinds of professional design work [10] or even financial services via crowd-funding [11]. There are many examples where such services are performed by volunteers, the most prominent one may be Wikipedia, but by now there also exist a number of professional platforms that connect businesses with workers willing to collaborate for a small payment. The first such platform was created in 2005 by Amazon Inc. under the name of *Mechanical Turk* where a *requester* can define and place so called *Human Intelligence Tasks* (*HITs*). These

 Digital Object Identifier: 10.1109/PCS.2012.6213338

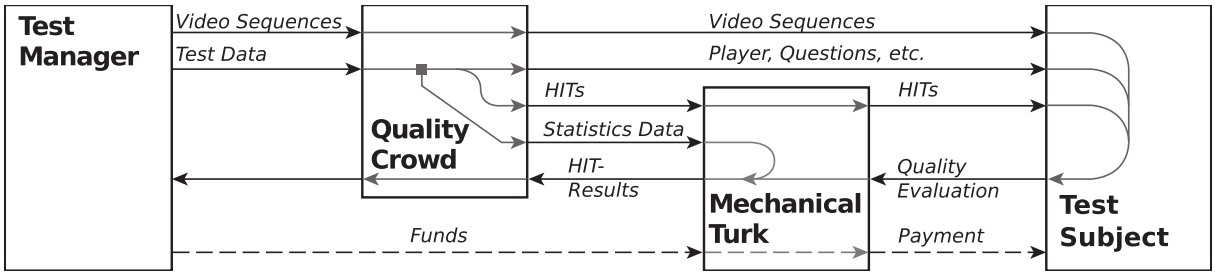

Fig. 1: Overview of the *QualityCrowd* framework.

HITs are small tasks which can be performed independently of each other. Any worker who is registered at the platform may choose to perform any *HIT* for the amount of payment which has been assigned to this *HIT* by the requester. There are, however, means to further limit the workforce based on age, nationality, or via a qualification test.

III. The QualityCrowd Framework

While Amazon does provide a web interface for the creation and management of *HITs*, this solution didn't provide the flexibility we needed to conduct video quality tests on Mechanical Turk. It is, however, possible to embed external web sites into a *HIT*, thereby rerouting the workers to another server where we were able to implement our framework to conduct the tests. As a separate HTTP-Server is needed in any case to transmit the videos to the worker, this approach has the added advantage that the test can be performed independently of the infrastructure of the crowdsourcing platform provider. While in the following we mainly focus on Amazon's Mechanical Turk, we maintain the possibility to use other providers or aggregators such as CrowdFlower or Microworkers. In the following paragraph, we will describe the implemenation of our framework in detail.

A. Software Architecture

We split our software framework into two parts; a front end that hosts the video test and is presented to the worker, and a back end where we can create new tests, upload new videos and manage existing tests. Both these interfaces are purely web based, meaning that both the worker and the operator will only need a reasonably up-to-date web browser to access the framework. This is particularly important for the front end, as most workers might not be willing to install new software on their system given the relatively small amount of payment for participating in the video test.

B. Video Delivery

At the center of an online video quality tests is the ability to play back videos to the worker. In traditional video testing, the video that is being presented to the test subjects is usually uncompressed raw video that has already been decoded from the original codec which is to be tested. This procedure is owing to the fact that the codecs which are to be tested are often still in development and the codec may not be available to the testing lab and may very well be much too complex to be decoded in real-time. While this is usually not a problem in a laboratory environment where data rate isn't a limiting factor, transmitting uncompressed video via the internet leads to prohibitively long waiting times for the worker, especially when considering wage rates. Lossy compression is also not an option as this will influence the test results. Therefore, the only solution is the application of lossless video coding. Also, as we want to reach the broadest worker base possible, we can't rely on additional plugins that the worker might have to install. We therefore evaluated exisiting solutions for embedding videos into web sites and came up with two options that we use in our front end.

The Adobe Flash Player is still the *de facto* standard for online video delivery. We evaluated the video formats and codecs that are supported by Flash Player and opted for the use of H.264/AVC with the High 4:4:4 Profile which supports lossless compression. The other option we chose to embed video in our web front end is to use the video tag that has been introduced by the World Wide Web Consortium (W3C) with HTML5. This element enables native browser support for video without any additional plugin, however, supported formats and codecs are not specified and therefore dependent on the browser. We check which option is available on the workers browser via JavaScript and choose the technology to embed the video accordingly.

C. Video Test Administration and Testing Procedure

In the back end, the operator can manage all video tests in a web interface. In the first step, he selects the video sequences that are to be tested and uploads them via the web interface onto the QualityCrowd server. In the next step, the operator chooses the test mode and test chain and creates the questions for the video tests and the qualification test. After the configuration has been finished, the operator may choose to start the video test. The framework then automatically generates corresponding *HITs* and puts them onto Amazon's Mechanical Turk platform. When a worker selects a *HIT* in his browser, the previously defined questions and video sequences are being loaded directly from the QualityCrowd server. QualityCrowd currently supports both single stimulus and double stimulus testing methodologies. After the worker submitted all the results for this *HIT*, the QualityCrowd server stores the results in a database and sends an estimate of the quality of the answers of the test subject to Mechanical

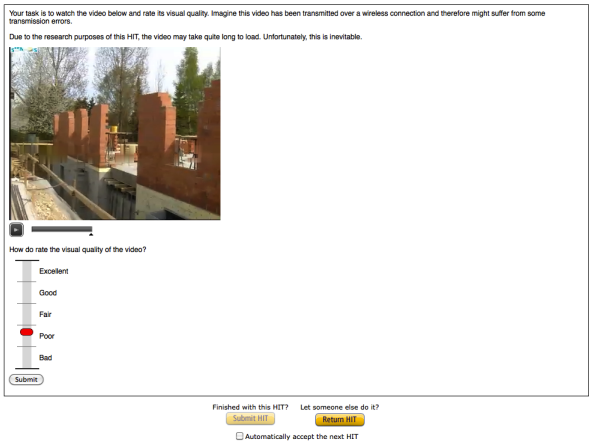

Fig. 2: *QualityCrowd* interface as seen by the test participants

Turk. Only when the test subject reaches a certain level of trustworthiness will he receive his payment from Mechanical Turk. This procedure is common with many crowdsourcing tasks to prevent fraud.

D. Server Architecture

QualityCrowd is a web application based on the Model View Controller (MVC) framework CakePHP. It therefore needs an HTTP server with PHP version 4.3.2 or higher and a database server, e.g. MySQL. The Amazon Mechanical Turk Command Line Tools for interfacing with the Mechanical Turk API require also a Java runtime environment on the server.

IV. Comparison to Lab Results

In order to confirm that QualityCrowd delivers valid results, we compare the results gained in a subjective test conducted with the QualityCrowd framework to the results from a test conducted in a standardized environment.

A. Comparison data set

We choose the data set presented by De Simone et al. in [12]. This data set contains the six CIF video sequences *Foreman*, *Hall*, *Mobile*, *News* and *Paris*, compressed with H.264/AVC, with two different realisations of 6 different packet loss rates, resulting in a total of 78 different processed videos including an error free version of the compressed video. The data set consists of both two subsets and a combined set with the mean opinion scores (MOS) from two different laboratories, *EPFL* and *PoliMi*, obtained in a single stimulus test. One motivation to use this data set was the availability of a very detailed description both of the test setup and processing of the votes in [12], allowing us to emulate the test environment and methodology of [12] in QualityCrowd as close as possible.

From the complete data set, we choose a subset consisting of the error free video and one packet loss realisation for the videos *Foreman*, *Hall*, *Mobile* and *Paris*, leading to 28 videos; *News* was used in the qualification test for the workers. We only selected one packet loss realisation as we are primarily interested in the different overall quality levels.

TABLE I: Correlation between QualityCrowd and the results from the different laboratories

	QualityCrowd			EPFL
	EPFL	PoliMi	EPFL+PoliMi	PoliMi
Foreman	0,9899	0,9929	0,9927	0,9949
Hall	0,9901	0,9922	0,9919	0,9955
Mobile	0,9966	0,9948	0,9972	0,9913
Paris	0,9963	0,9925	0,9963	0,9896
all	0,9920	0,9922	0,9937	0,9918

B. Test setup

The test was performed with in total 19 test subjects using the QualityCrowd framework and Mechanical Turk. In a first step, we decided to perform the test within the same local network as the server providing the video sequences in order to minimize possible internet connection problems. On average, the connection bitrate was 3.7 MBit/s and the evaluation of each video took 47 s. The web interface as seen by the test subjects in their browser is shown in Fig. 2.

All test subjects were required to take an online qualification test provided in the Mechanical Turk platform, before they were able to participate in the subjective test. Note that no further training or explanation was provided to the subjects, thus representing similar conditions to a test campaign using QualityCrowd in a more general setting. The processing of the votes and outlier detection was done according to [12].

C. Results

In Table I, we present the Pearson correlation coefficients between the results gained with QualityCrowd and both the complete data set and the two subsets. Additionally, we also provide the correlation between the two subsets EPFL and PoliMi themselves. We can see that the overall correlation between QualityCrowd and the results from the EPFL/PoliMi data set is comparable to the inter-lab correlation between both subsets. Note that in the recently finished Video Quality Experts Group (VQEG) HDTV Phase I project the lowest acceptable inter-lab correlation was 0.94 [13].

Fig. 3 additionally shows that for most video sequences and packet error rates the confidence intervals of the results from QualityCrowd overlap with results from the combined EPFL and PoliMi data set, indicating that there is no statistical significant difference between the results. The comparisons between QualityCrowd and each of the two subsets considered separately show similar results.

V. Conclusion

We presented the *QualityCrowd* framework, a web-based platform for video quality evaluation using crowdsourcing. The comparison with results from tests performed in a traditional lab setting shows that crowdsourced online testing delivers similar results. While we only considered video so far, the architecture of QualityCrowd is capable of delivering

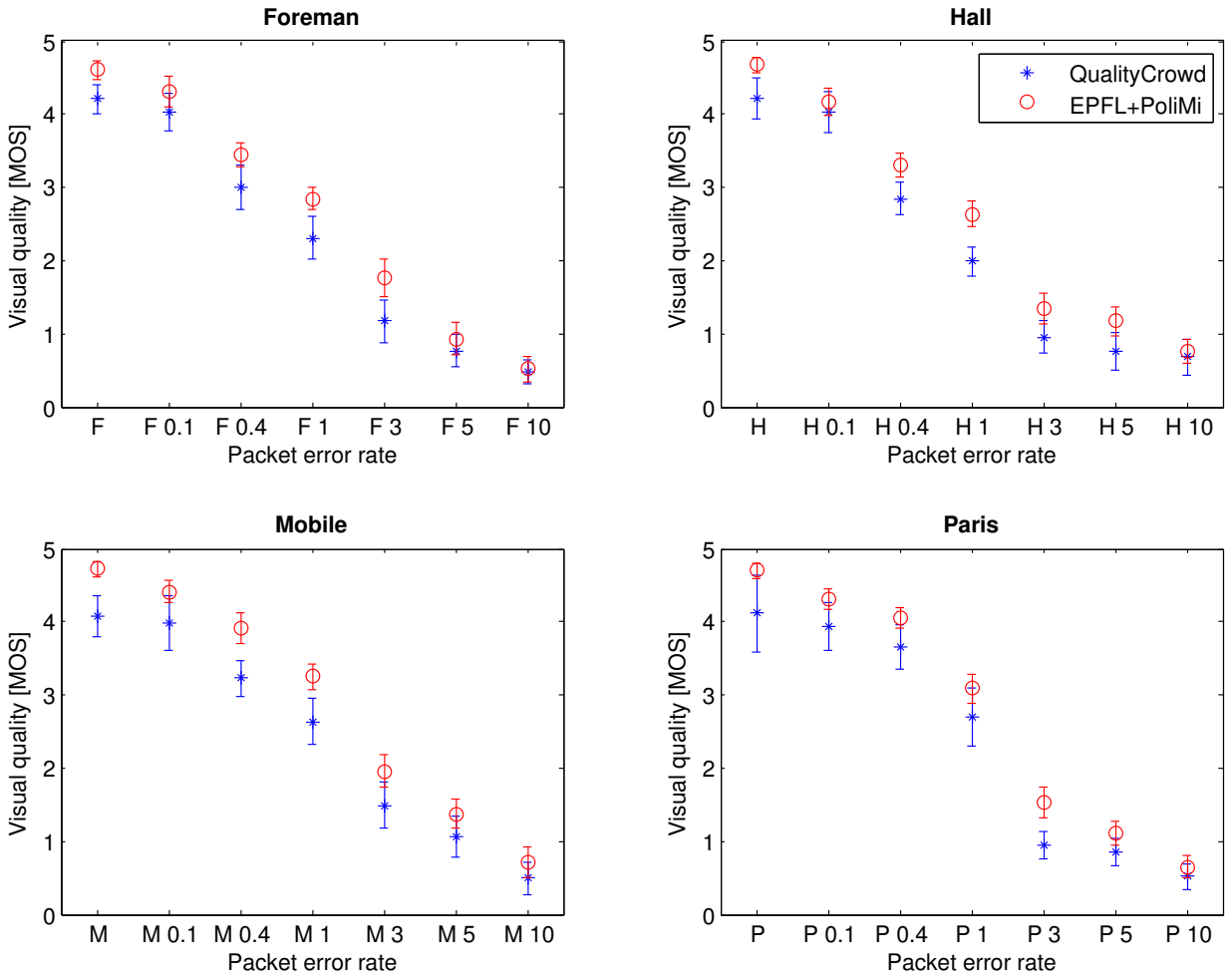

Fig. 3: *QualityCrowd* compared to the combined EPFL and PoliMi data sets: MOS and 95% confidence intervals for the video sequences *Foreman*, *Hall*, *Mobile* and *Paris*

losslessly compressed audio-visual content and thus can also support multimodal quality assessment. In future work, we plan to extend QualityCrowd to other testing methodologies and include an integrated outlier detection, but also to run tests with a wider audience and further study the reproducibility of the results.

The *QualityCrowd* framework is available for download at `www.ldv.ei.tum.de/videolab`

References

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

[2] J. Bohannon, "Social science for pennies," *Science*, vol. 334, no. 6054, p. 307, 2011.

[3] G. Paolacci, J. Chandler, and P. Ipeirotis, "Running experiments on Amazon Mechanical Turk," *Judgment and Decision Making*, vol. 5, no. 5, pp. 411–419, Jun. 2010.

[4] M. Marge, S. Banerjee, and A. Rudnicky, "Using the amazon mechanical turk for transcription of spoken language," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, Mar. 2010, pp. 5270 –5273.

[5] K.-T. Chen, C.-C. Wu, Y.-C. Chang, and C.-L. Lei, "A crowdsourceable QoE evaluation framework for multimedia content," in *Proceedings of the 17th ACM international conference on Multimedia*. ACM, 2009, pp. 491–500.

[6] 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, IEEE*, vol. 24, no. 2, pp. 28 –35, Mar. 2010.

[7] F. Ribeiro, D. Florencio, C. Zhang, and M. Seltzer, "Crowdmos: An approach for crowdsourcing mean opinion score studies," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, May 2011, pp. 2416 –2419.

[8] F. Ribeiro, D. Florencio, and V. Nascimento, "Crowdsourcing subjective image quality evaluation," in *Image Processing (ICIP), 2011 IEEE International Conference on*, Sep. 2011, pp. 3158–3161.

[9] J. Howe, "The rise of crowdsourcing," *Wired Magazine*, vol. 14, no. 06, 2006.

[10] D. C. Brabham, "Crowdsourcing as a model for problem solving," *Convergence: The International Journal of Research into New Media Technologies*, vol. 14, no. 1, pp. 75–90, 2008.

[11] A. Gaggioli and G. Riva, "Working the crowd," *Science*, vol. 321, no. 5895, p. 1443, 2008.

[12] F. De Simone, M. Naccari, M. Tagliasacchi, F. Dufaux, S. Tubaro, and T. Ebrahimi, "Subjective assessment of H.264/AVC video sequences transmitted over a noisy channel," in *Proceedings of the First International Workshop on Quality of Multimedia Experience (QoMEX 2009)*, Jul. 2009.

[13] Video Quality Experts Group (VQEG), "Report on the validation of video quality models for high definition video content," Tech. Rep., Jun. 2010.