Video Quality Evaluation in the Cloud

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 evaluation 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. The QualityCrowd framework allows codec independent, crowd-based video quality assessment with a simple web interface, usable with common web browsers. However, due to its codec independent approach, the framework can pose high bandwidth requirements on the coordinating server. We therefore propose in this contribution a cloud-based extension of the QualityCrowd framework in order to perform subjective quality evaluation as a cloud application. Moreover, this allows us to access an even larger pool of potential participants due to the improved connectivity. We compare 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.**

I. Introduction

Video quality is usually evaluated 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' locations in order to be able to hire enough qualified subjects.

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 introduced the *QualityCrowd* framework in [3], a codec agnostic, web-based platform for video quality evaluation with crowdsourcing, allowing us to assess the visual quality not only of existing coding technology, but also of future developments in common web browsers e.g. Firefox or Internet Explorer. As the focus of this previous contribution was on the feasibility of the QualityCrowd framework itself, we conducted the verification of the framework in a local network environment due to the lossless compression of the videos and the resulting bandwith demands. Also the demographic of the online workers was therefore rather limited.

In this contribution we therefore extend QualityCrowd into the cloud: firstly by leveraging the global worker pool in the subjective testing and secondly by shifting the videos into the cloud, thus optimizing the distribution of the videos to the workers. We will show that crowdsourcing delivers comparable results to subjective testing in a standardised environment. This is to the best of our knowledge the first contribution that proposes video quality evaluation as a cloud application.

Paolacci et al. examined in [4] whether 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 [5] that crowdsourcing delivers similar results to traditional methods for audio transcription. Chen et al. conducted subjective audio-visual tests via crowdsourcing in [6], [7], but used a non-standardized testing methodology and MP3 and H.264/AVC for compression. Finally, Ribeiro et al. presented the *crowdMOS* framework in [8], [9], 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. Also none of these previous contributions utilized the cloud to deliver the content to the workers.

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

II. Crowdsourcing

A. The Crowdsourcing Principle

The term Crowdsourcing has first been coined by Howe in the article *The Rise of Crowdsourcing* in Wired Magazine in 2006 [10]. 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

 Digital Object Identifier: 10.1109/PV.2012.6229729

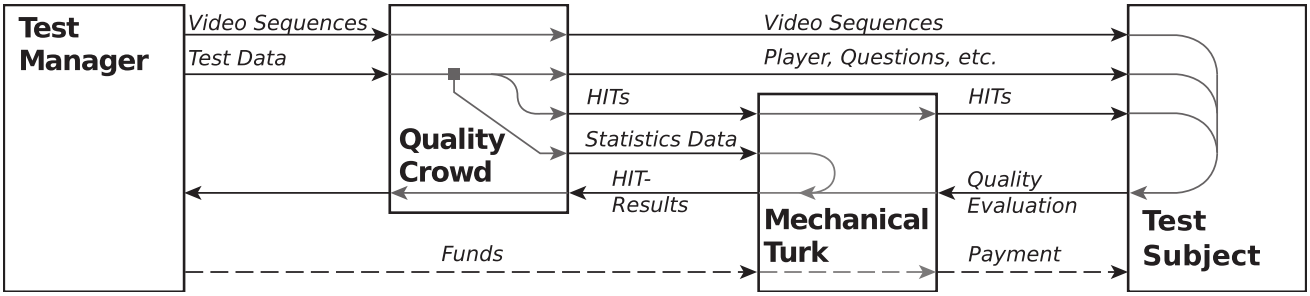

Fig. 1: Overview of the *QualityCrowd* framework.

business ideas [11], all kinds of professional design work [11] or financial services via crowd-funding [12]. 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.

B. Crowdsourcing Platforms

The first and still most prominent platform was created in 2005 by Amazon Inc. under the name *Mechanical Turk* where a *requester* can define and place so called *Human Intelligence Tasks* (*HITs*). These *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.

One limitation of Amazon's Mechanical Turk is the restriction in its terms of use that while workers may sign-up from all over the world, requesters must a either be US citizens or legal entities registered in the US. There are, however, alternative crowdsourcing platforms available e.g. *Microworkers* [13] that do not have this limitation. Unfortunately, these alternatives usually do not provide a worker pool as large as Amazon's service. Another option is the use of *aggregators*, that bundle access to multiple, different crowdsouring platforms in a common API, thus acting as an abstraction layer for the different platforms. One such aggregator is *CrowdFlower* [14], that provides access to multiple crowdsourcing platforms via a common API interface and was also used in this contribution as a *wrapper* for Mechnical Turk.

III. The QualityCrowd Framework

In this section, we will shortly describe the *QualityCrowd* framework as presented in [3]. While Amazon or aggregators do provide a web interface for the creation and management of *HITs*, these solutions 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. In the following section, we will describe the implementation 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 won't 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

We had to make several design choices regarding video delivery to the worker. In a traditional lab environment, the video that is being presented to the test subjects is usually uncompressed raw video that was already coded and decoded with the codec that is to be tested. This procedure owes to the fact that the codecs which are to be tested are often still in development so that they are not only unavailable to the testing lab, but usually have also a computational complexity not suitable for real-time decoding. However, in an internet based test, data rate limitations for transferring the videos to the user need to be considered. Transmitting uncompressed video would lead to unacceptably high waiting times, especially when considering the relatively small amount of payment that the worker receives for each video. On the other hand, we obviously can't use any form of lossy compression as this would influence the test results. Therefore, we need to employ lossless video compression. But we also want to reach the broadest worker base possible, so we can't rely on additional plugins that the worker might have to install. After evaluating different video embedding solutions for web browsers, we identified two suitable options for our front end.

The Adobe Flash Player is still the *de facto* standard for online video delivery with more than 95% penetration for PC

browsers [15]. 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. As second option we chose to embed video is 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 then 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 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. 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 Turk. An overview of the complete *QualityCrowd* framework is shown in Fig. 1.

IV. Moving QualityCrowd into the Cloud

One problem of the lossless compression scheme is that the videos delivered to the workers have a comparably large file size and therefore pose high bandwidth requirements on the server. If we consider videos in CIF format with a spatial resolution of 352×288 pixels as used in section V, the file size of a typical 10 s video test sequence after lossless compression with H.264/AVC is between 5 MByte and 16 MByte, depending on the content. In this contribution, this would result in nearly 260 MByte to be transferred, if a worker were to complete all HITs. The bandwidth needed to transmit these files is 10 to 20 times larger than lossy compression with H.264/AVC, where for good visual quality the file size for the same sequences is between 0.4 MByte to 0.8 MByte [16], [17]. Additionally, the workers are highly likely to be distributed all over the world. So we have to ensure that we can reach each worker with an acceptable bitrate.

We therefore decided to extend our QualityCrowd framework into the cloud by moving the videos under test into the cloud, thus not only avoiding bandwidth restrictions at the QualityCrowd server, but also providing better access to the data to workers regardless of their location. In this contribution, we use Amazon Web Services' (AWS) CloudFront content distribution network (CDN) for the delivery of the videos. A CDN consists of servers located all over the world at so called *edge locations*. This ensures that each user may be able to connect to a server which is close to his network, thus enabling file transfer with high data rate and minimum latency.

The QualityCrowd server will now automatically transfer these videos to a so called *S3 bucket*, AWS' cloud storage system after the videos have been uploaded via the QualityCrowd back end. From there, the files are transfered to the CloudFront edge location servers, as shown in Fig. 2.

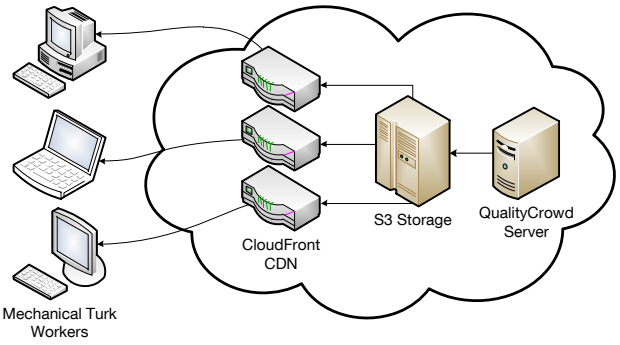

Fig. 2: Distribution of videos via CloudFront

Moreover, this shift into the cloud also allows us to access a larger worker pool with a more representative demographic, as the improved connectivity due to the CloudFront edge locations makes the *HITs* attractive to a larger group of potential workers.

V. Comparison to Lab Results

In order to confirm that the cloud-based extension of QualityCrowd delivers valid results, we compare the results gained in a subjective test conducted with the cloud-based QualityCrowd framework and Mechanical Turk with the results from a test conducted in standardized environment and the results from [3], where we introduced the QualityCrowd framework.

A. Comparison data set

We choose the data set provided presented by De Simone et al. in [16], [17]. This data set contains the six CIF video sequences *Foreman*, *Hall*, *Mobile*, *Mother*, *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 a error free version of the compressed video. The data set consists of two subsets with the mean opinion scores (MOS) from two different laboratories, *EPFL* and *PoliMi*, obtained in singe stimulus test with scale a from 0 to 5, from worst to best visual quality.

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 [16], [17], allowing us to emulate the test environment and methodology of [16], [17] 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

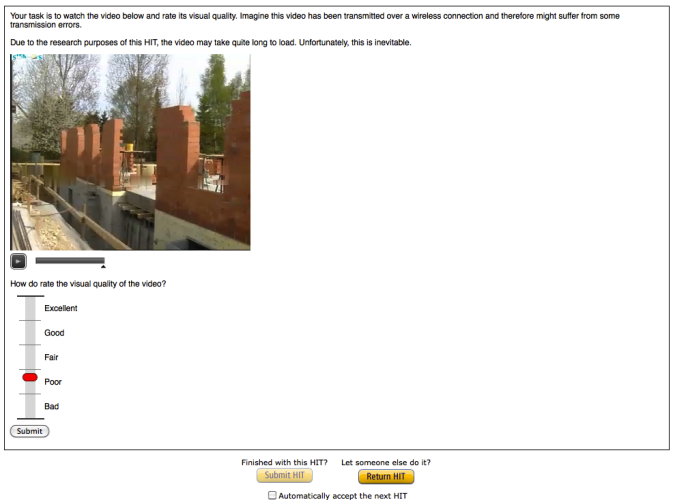

Fig. 3: *QualityCrowd* interface as seen by the test participants

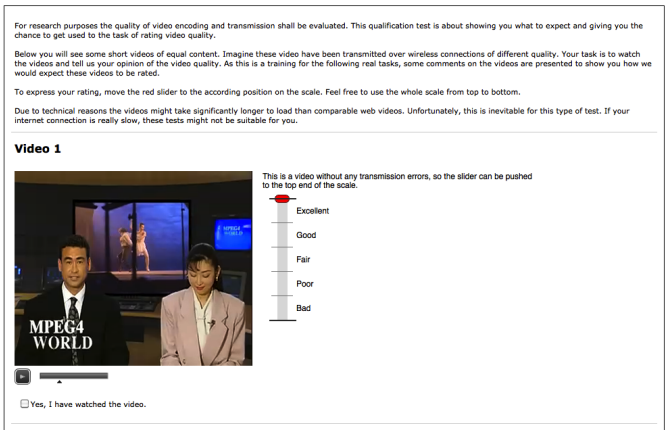

Fig. 4: *QualityCrowd* online training

videos *Foreman*, *Hall*, *Mobile* and *Paris*, leading to 28 videos; *News* was used in the training of the workers. We only selected one packet loss realisation as we are primarily interested in the different overall quality levels

B. Comparison to Local QualityCrowd

Additionally, we compare the results to our previous contribution in [3], where a subjective test using the QualityCrowd framework with same data set was performed within the same local network as the server providing the video sequences. This was done to minimize possible internet connection problems, as the focus was primarily on the feasibility of crowdbased quality assessment with uncompressed content via a web interface. In total 19 test subjects using the QualityCrowd framework and Mechanical Turk took part and on average the connection bitrate was 3.7 MBit/s, which suggests that a sufficient bandwidth is necessary, if a larger number of simultaneous workers are working on the same quality assessment *HIT*. Each worker rated all videos in the data set. Also the demographic of this test was rather limited, as all 19 subjects were students at Technische Universität München and thus able to ask for further assistance, if problems occurred. Also no monetary compensation was provided to the participating workers.

C. Test Setup of Cloud-based QualityCrowd

In contrast to the local QualityCrowd test described above, we now moved both the videos and the worker pool into the cloud, loosening our control of the overall test setup compared to [3]. The web interface as seen by the test subjects in their browser is shown in Fig. 3. Each video sequence under test was mapped to one *HIT*. All test subjects were asked to perform an online training provided at the QualityCrowd server before participating in the subjective test. In this training, the workers were shown the video sequence *News* at different quality levels and corresponding hints about the suitable quality rating as shown in Fig. 4. Note that no further training or explanation was provided to the subjects. Moreover, due to the design principle behind the Mechanical Turk and the used CrowdFlower aggregator service, no direct interaction between the workers and requester is possible. Hence, no individual support as in our local QualityCrowd test was possible.

The compensation per *HIT* was set to 0.08 USD. Each *HIT* corresponds to one of the 28 video sequence described in section V-A. In total, 96 different workers from 12 different countries participated in the test. Each worker rated on average 7 videos. Note, that due to the split of the complete test into *HITs*, the number of different videos rated by each worker can vary significantly, as it is not possible to enforce that each worker rates all video sequences. Hence, only roughly 8% of all workers rated all 28 videos and 85% of all workers rated less than half of all videos. Due to these incomplete data set per worker no outlier detection was performed on the results.

VI. Results and Discussion

In Table I, we present the Pearson correlation coefficients between the results gained with the cloud-based QualityCrowd, the comparison data sets and the local QualityCrowd. Additionally, we also provide the correlation between the two subsets EPFL and PoliMi themselves. We can see, that the overall correlation between the cloud-based QualityCrowd and the results from the combined EPFL+PoliMi data set is slightly worse compared to the local QualityCrowd with an overall correlation coefficient of 0.9514 versus 0.9937.

In the recently finished Video Quality Experts Group (VQEG) HDTV Phase I project the lowest acceptable inter-lab correlation was 0.94 [18]. Hence, the results from the cloud-based QualityCrowd are still within the acceptable inter-lab variation and can be considered to be valid.

Fig. 5 shows for the four video sequences and bit error rates the confidence intervals and MOS scores of the results from cloud-based QualityCrowd compared with the results from the combined EPFL+PoliMi data sets. We notice that while for average visual quality the confidence intervals overlap for most video sequences and thus indicating that no statistically significant difference between the cloud-based QualityCrowd

TABLE II: Standard deviation of the cloud-based QualityCrowd results and both the results from the comparison data sets and the local QualityCrowd evaluation.

	EPFL+PoliMi	local QualityCrowd	cloud-based QualityCrowd
Foreman	0.562	0.501	1.026
Hall	0.507	0.490	1.065
Mobile	0.547	0.583	1.170
Paris	0.476	0.603	1.036
average	0.523	0.544	1.074

and the EPFL+PoliMi data sets exist, the results on both ends of the visual quality scale differ significantly.

The observed sigmoid shape is an indication of a typical phenomena in subjective testing, occurring when test subjects are not utilizing the complete scale: they avoid both ends of the scale and thus the votes tend to saturate before reaching the end points. In our case, the votes were limited mostly to the range between 1 and 4. Hence only 60% of the MOS scale was utilized by the test subjects. Usually this phenomena is avoided in a lab environment by providing the test participants with an extensive training phase including individual feedback by the test supervisor if a participant seems to have problems. One the one hand, direct feedback was not possible, but on the other hand, we could also not ensure that all workers completed the online training properly.

Furthermore, the variation of the individual votes in the cloud-based QualityCrowd is significantly larger compared to both the local QualityCrowd and the EPFL+PoliMi data set as indicated by the standard deviation in Table II, but also the rather large confidence intervals in Fig. 5. This can be explained by the fact, that only 8% of all workers rated all videos, compared to both the local QualityCrowd and the EPFL+PoliMi data set, where all workers/subjects rated all videos. Hence, most workers were probably unable to gain sufficient experience and were thus not as proficient as possible in the quality assessment. Moreover, we could also not apply common outlier detection methods as e.g. in [1], as these methods assume that all subjects rated all videos.

VII. Conclusion

We extended the *QualityCrowd* framework for web-based video quality evaluation with crowdsourcing into the cloud and introduced video quality evaluation as a cloud application. The comparison with results from tests performed in a traditional lab setting, but also with a local QualityCrowd setting, shows that the cloud-based QualityCrowd delivers acceptable results and is thus an valid alternative.

There are, however, two issues that need to be addressed in future works to increase the reliability of the cloud-based results: firstly, the pre-test training of the participating workers needs to be improved, including an feedback mechanism, in order to ensure that the workers are properly prepared for their task. Secondly, the segmentation of the overall test in single *HITs* need to be reconsidered, as it allows workers to only participate in a small subset of the test, leading to problems in the processing of the votes and possibly to a lack of workers' experience. Additionally, more sophisticated validation methods from the crowdsourcing community need to be adopted to introduce a better screening of the workers with respect to their reliability.

The *QualityCrowd* framework including its cloud extension 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] C. Keimel, J. Habigt, C. Horch, and K. Diepold, "QualityCrowd - A Framework for Crowd-based Quality Evaluation," in *Picture Coding Symposium (PCS) 2012*, 2012, submitted.

[4] G. Paolacci, J. Chandler, and P. Ipeirotis, "Running experiments on amazon mechanical turk," *Judgment and Decision Making*, vol. Vol. 5, no. No. 5, pp. 411–419, Jun. 2010.

[5] 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.

[6] 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*, ser. MM '09. ACM, 2009, pp. 491–500.

[7] 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.

[8] 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.

[9] 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.

[10] J. Howe, "The rise of crowdsourcing," *Wired Magazine*, vol. 14, no. 06, 2006. [Online]. Available: http://www.wired.com/wired/archive/14.06/crowds.html

[11] 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.

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

[13] Weblabcenter, Inc. (2011, Dec.) Microworkers. [Online]. Available: http://www.microworkers.com/

[14] CrowdFlower, Inc. (2011, Dec.) Crowdflower. [Online]. Available: http://www.crowdflower.com

[15] Adobe Systems Inc., "Adobe flash platform runtimes: PC penetration statistics," 2011. [Online]. Available: http://www.adobe.com/products/flashplatformruntimes/statistics.html

[16] 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.

[17] F. De Simone, M. Tagliasacchi, M. Naccari, S. Tubaro, and T. Ebrahimi, "A h.264/avc video database for the evaluation of quality metrics," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, march 2010, pp. 2430 –2433.

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

TABLE I: Correlation between the cloud-based QualityCrowd results and both the results from the comparison data sets and the local QualityCrowd evaluation for each video sequence and the overall correaltion for all sequences. Additionally, the results between the local QualityCrowd and the comparison data sets presented in [3].

	cloud-based QualityCrowd				local QualityCrowd			EPFL
	local QualityCrowd	EPFL	PoliMi	EPFL+PoliMi	EPFL	PoliMi	EPFL+PoliMi	PoliMi
Foreman	0.9871	0.9619	0.9634	0.9651	0.9897	0.9929	0.9927	0.9949
Hall	0.9785	0.9573	0.9613	0.9622	0.9901	0.9921	0.9919	0.9955
Mobile	0.9543	0.9675	0.9315	0.9494	0.9966	0.9948	0.9972	0.9913
Paris	0.9840	0.9905	0.9715	0.9812	0.9962	0.9925	0.9963	0.9896
all	0.9672	0.9543	0.9450	0.9514	0.9926	0.9922	0.9937	0.9918

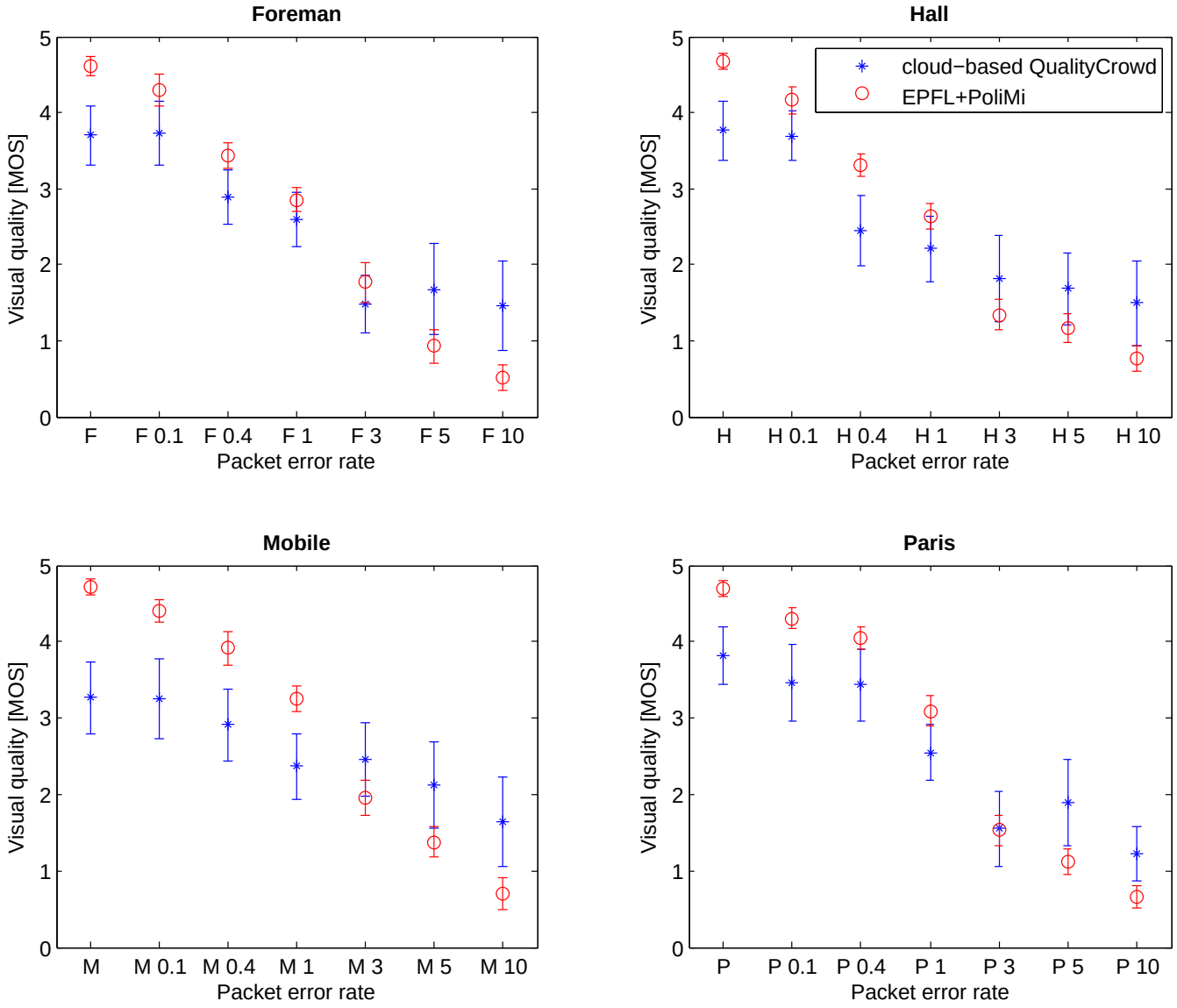

Fig. 5: Cloud-based *QualityCrowd* compared to the combined EPFL+PoliMi data set: MOS and 95% confidence intervals for the video sequences *Foreman*, *Hall*, *Mobile* and *Paris*