CHALLENGES IN CROWD-BASED VIDEO QUALITY ASSESSMENT

Christian Keimel, Julian Habigt and Klaus Diepold

Technische Universität München, Institute for Data Processing
Arcisstr. 21, 80333 Munich, Germany
christian.keimel@tum.de, jh@tum.de, kldi@tum.de

ABSTRACT

Video quality evaluation with subjective testing is both time consuming and expensive. A promising new approach to traditional testing is the so-called crowdsourcing, moving the testing effort into the Internet. The advantages of this approach are not only the access to a larger and more diverse pool of test subjects, but also the significant reduction of the financial burden. Recent contributions have also shown that crowd-based video quality assessment can deliver results comparable to traditional testing in some cases. In general, however, new problems arise, as no longer every test detail can be controlled, resulting in less reliable results. Therefore we will discuss in this contribution the conceptual, technical, motivational and reliability challenges that need to be addressed, before this promising approach to subjective testing can become a valid alternative to the testing in standardized environments.

Index Terms— Crowdsourcing, subjective testing, video quality assessment, cloud applications

1. 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.

Crowdsourcing is an alternative to the classical approach to subjective testing that has received increased attention recently. It uses the Internet to assign simple tasks to a group of online workers. Hence tests are no longer performed in a standard conforming laboratory, but conducted 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 expenditures significantly.

Comparisons between the results from classic and crowdsourced subjective testing show a good correlation for some methodologies [2,3] similar to usual inter-lab correlations. In general, however, new problems arise, as we can no longer control every detail both in the test setup, but also in the testing itself, resulting in less reliable results.

In this contribution we therefore provide an overview of the challenges faced in the context of crowd-based video quality assessment that need to be addressed, before this promising new approach to subjective testing can become a valid alternative to the assessment in the standard lab environment. In the field of social sciences, crowdsourcing has become quite popular and many contributions, e.g. [4, 5] discuss its advantages, but also its challenges. More closely related to this contribution, Chen et al. present in [6] a framework for crowd-based quality assessment and discuss influences of crowdsourcing on the results, but use a non-standardized methodology especially fitted to its use in a web application. To the best of our knowledge, this is therefore the first contribution focusing on the overall challenges faced in general by subjective video quality assessment with crowdsourcing.

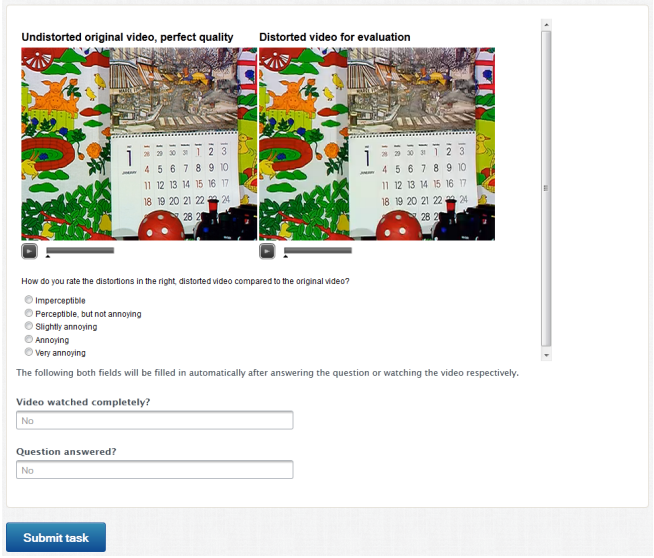

Fig. 1: Crowd-based video quality assessment: Web interface as seen by test participants [2]

This contribution is organized as follows: after a short introduction into the concept of crowdsourcing and crowd-based video quality assessment, we discuss the challenges faced on the conceptual, technical, motivational and reliability levels before we conclude with a short summary.

2. CROWDSOURCING

The term Crowdsourcing has first been coined by Howe in the article *The Rise of Crowdsourcing* in Wired Magazine in 2006 [7]. 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 per-

 Digital Object Identifier: 10.1109/QoMEX.2012.6263865

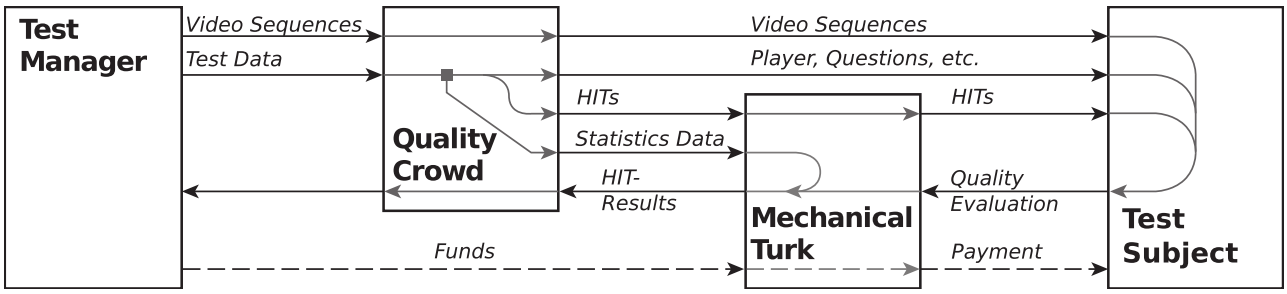

Fig. 2: Overview of the *QualityCrowd* framework [2]

formed by non-trained workers, e.g. tagging photos with meaningful key words. However, even rather complex services can be crowd-sourced, like creative tasks such as the generation of new business ideas [8], all kinds of professional design work [8] or financial services via crowd-funding [9]. 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 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 [2].

3. CROWD-BASED VIDEO QUALITY ASSESSMENT

In crowd-based video quality assessment we utilize these crowdsourcing platforms to perform subjective testing with a global worker pool, usually with a web-based application, that can be accessed via common web browsers, e.g. Firefox or Internet Explorer. Examples of web-based audio-visual quality assessment applications include [2, 6, 10–12].

We will illustrate the basic principles shortly on the *QualityCrowd* framework as proposed in [2]. In the overview of the framework in Fig. 2 we can see how the crowdsourcing platform acts as an extra layer between test manager and test subject, handling the recruiting and payment of the test participants. The videos under test are losslessly compressed and then provided to the test subjects via a web interface in their browser, as shown in Fig. 1 for a double-stimulus testing methodology with a discrete impairment scale. Subjects then assess the visual quality and the corresponding judgements are provided to the test manager. The aim is to keep the methodology as close as possible to the methodology used for subjective tests in a lab environment. Additionally, an online training with explanations similar to those in a lab environment is provided to the participants, see Fig. 7.

Note that the test manager has neither direct influence on the selection of the participants, i.e. who is performing a *HIT*, nor has he immediate access to the evaluation results. Only the videos with the corresponding presentation interface are provided directly to the test subjects. The videos may either be delievered directly from the test manager's site or hosted in the cloud via a content delivery network, e.g. Amazon's Cloudfront as in [3].

4. CHALLENGES

Even though the results for single-stimulus methodologies in [2, 3] have been very promising so far, with correlations between traditional subjective testing and crowdbased video quality assessment exceeding the minimum inter-lab correlation as proposed by VQEG in [13], there still remain many open issues. Compared to subjective testing in a standard environment, the results are therefore often less reliable.

In order to overcome the current limitations, we need to address the unique challenges faced by moving video quality assessment online with a crowd-based approach. We shall therefore discuss in this section these challenges from a conceptual, technical, motivational and reliability point of view.

4.1. Conceptual Challenges

The first challenges we face derive from some of the differences in the basic concepts of crowdsourcing compared to the structure and procedures of subjective testing.

HITs are supposed to be small tasks that can be done by the workers both fast and easily. While there are no technical limitations on the *HITs'* complexity or the time requirements on the predominant crowdsourcing platforms, the larger and more time consuming a *HIT* is, the longer it takes to find workers doing this *HIT* or the workers may even ignore such *HITs* mostly, as many workers prefer a high reward per hour [14]. Hence, it is not possible to just run a test designed for a lab environment without modifications; it rather needs to be partitioned into smaller chunks, e.g. its basic test cells (BTCs) or at least a rather small subset of BTCs of the overall test as illustrated schematically for a single-stimulus test in Fig. 8. These separate *HITs* are then grouped in one overall *job* or *batch*, representing the complete test. Moreover, in [15] results suggest that the more *HITs* for given job exist, i.e. in our case the overall test, the more likely it is that the *HITs* will be chosen by the workers, thus implying that the attractiveness for a certain job is influenced by the granularity of the work.

An additional problem arises from the fact that usually each *HIT* is only considered on its own and thus the *HITs* in one job or batch are presented in random order to the workers. This, however, also means that certain design rules for subjective testing can no longer be adhered to. In particular, it is no longer assured that specific design considerations with respect to a stabilization phase at the beginning of a test or a particular sequence order to avoid contextual effects will be followed [16].

Moreover, we now can also no longer guarantee that every test subjects judges every video in the test, but often the majority of the workers only assess a subset of the video sequences under test. In [3]

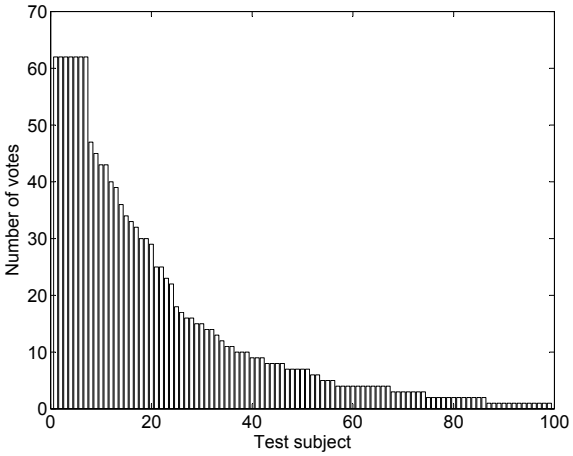

Fig. 3: In total 63 different videos, but only 7 out of 99 participants completed the whole test

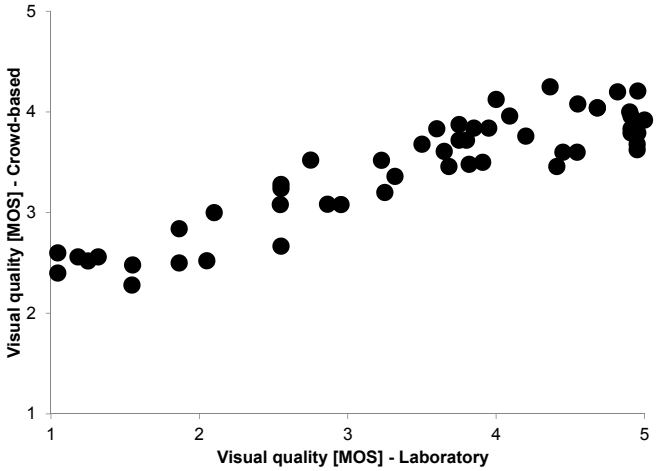

Fig. 4: Untrained subjects: results from crowd-based video quality assessment vs. results from a lab environment

for example, only 7% of all workers assessed the complete test set and 83% of all workers finished less than half of all videos as shown in Fig. 3. This has in turn also consequences for the processing of the votes, as the statistical methods and assumptions of most outlier detection methods, e.g. in [1] assume that each subject assessed all videos. Thus these methods can no longer be applied in crowd-based video quality assessment and new methods for processing and outlier detection need to be devised.

Another limitation is, that it is not possible to perform a training of the subjects in the same way as in a traditional lab environment. Even though so-called *qualification tests* as shown in Fig. 7 can be provided online, giving the workers an introduction into the test methodology and context similar to a training in the lab, there is no feedback between the test supervisor and subject in the training phase. In particular, we can neither determine if the test subject really understood the task at hand even after completing the online training, nor can we guarantee with some crowd-sourcing plattforms, which do not include mandatory *qualification tests*, e.g. [17], that the training was done at all. The consequences of this lack of training are illustrated in Fig. 4, where we compare the results from a crowd-based double-stimulus assessment with the results from a lab environment [18]. The observed sigmoid shape is an indication of a typical phenomenon 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. This effect is also noticeable when comparing the results in [2], where the video quality assessment was web-based, but the subjects had direct contact to the test managers, and [3], where the same test was run globally without contact to the test managers: the results from [3] exhibit also a noticeable sigmoid shape compared to [2]. Usually this phenomena can be 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.

4.2. Technical Challenges

The second dimension in which we face challenges compared to the standardized lab environment are the technical aspects of the crowd-based video assessment.

Obviously, we are neither able to control the setup of the testing environment itself, e.g. room illumination, nor the used equipment, e.g. displays. Although a standard conforming display might not be that important [19], we should still ensure that at least some minimum requirements are fulfilled, e.g. when a worker is supposed to assess the visual quality of high definition material, the worker's display should be capable to display the demanded resolution. Once again, however, this criterion cannot be enforced.

The choice of purely web-based quality assessment applications is not so much driven by the convenience for the workers, but rather by the limitations due to the crowdsourcing platforms' policies. Amazon's Mechanical Turk for example, explicitly prohibits in its policies that workers are required to download and install additional software in order to complete a *HIT* [20]. Hence, we are limited to the functionality provided by current web browsers and their commonly available plug-ins, in particular Adobe Flash. Without these limitation, one could deploy specific applications, e.g. the interactive SAMVIQ test [21], better suited to overcome some of the pitfalls described in this contribution.

Another point is that we also need to deliver the video sequences under test to the workers via the Internet. This seems to be an easy task, as video streaming has become quite common, e.g. with video hosting services such as YouTube, in the last few years, but once again we face unique challenges due to the very nature of video quality assessment. Firstly, we need to consider that in general the worker's web-browser and plug-ins cannot be assumed to support the original encoding format of our videos, as this would necessarily limit our research to already widely adopted coding standards and their profiles. Therefore we have to deliver the videos either uncompressed or using lossless compression to the workers, if we want to be able test also new coding technologies or other processing algorithms. One could of course re-encode the videos for the delivery with common lossy coding techniques, but then we would move even further from the ideal lab setup, as we then also implicitly assess the artefacts introduced by this additional compression. Lossless compression, however, leads to comparably large files and thus to higher bitrate demands. In [2, 3], lossless compression with H.264/AVC results in file sizes between 5 MByte and 16 MByte for 10 s test sequences in CIF resolution. This is, depending on sequence, 10 to 20 times larger than lossy compression with H.264/AVC. Hand in

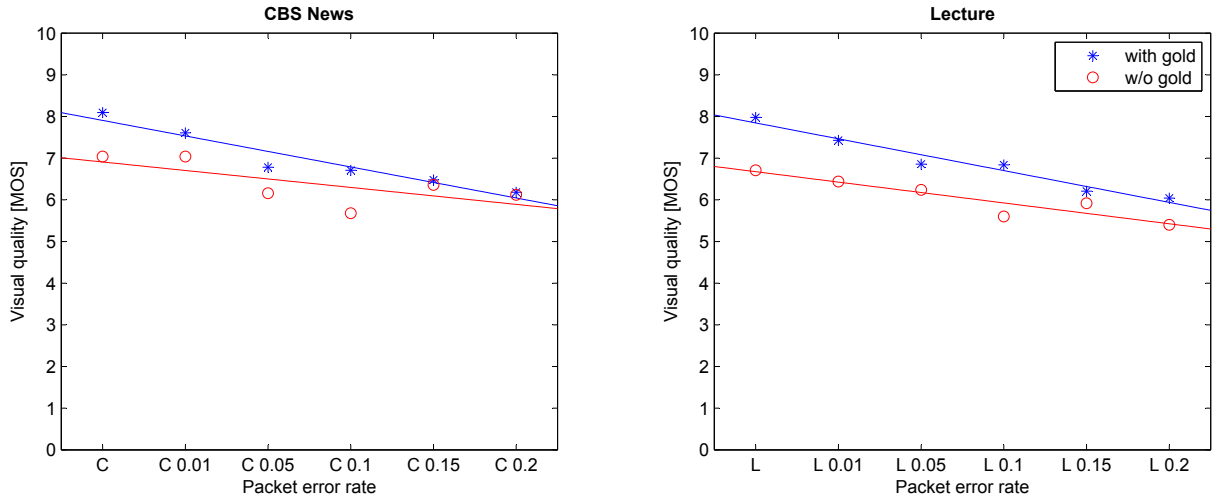

Fig. 5: The introduction of reference video data as a gold standard to monitor the accuracy of the votes can lead to a bias in the test result

hand with this goes the necessity to transfer this large amount of data to workers worldwide and that the bitrate available at the workers' premises is sufficient for a speedy transfer of the videos. For high definition video sequences, these demands on the worker's Internet connection may very well become too prohibitive.

4.3. Motivational Challenges

The third area we have to consider when transferring subjective testing tasks into the crowd, is the motivation of the test subjects to participate in the test.

One factor is, of course, the financial incentive leading to the question of how much to pay the test subjects, as on the one hand we need to pay enough to attract qualified workers and also a quick turnaround of the *HITs*, on the other hand we don't want to overspend as the cost savings are one of the main benefits of crowdsourcing. Hence, it is important to find an appropriate equilibrium between cost, timeliness and the quality of results. Horton and Chilton showed in [22] that the median reservation wage, i.e. the minimum wage a worker is willing to accept for a given task, of workers in Amazon's Mechanical Turk was $1.38/h. Although some contributions indicate that the reward itself does not influence the quality of the results, but rather the number of *HITs* done by each worker [23], other contributions also indicate that there might be some influence in some applications [24]. Moreover, experimental results in [23, 25] suggest that the compensation scheme has more overall influence on the quality of the results. With crowd-based video quality assessment, we face the problem that already a simple 10 s BTC in a video test requires a much larger time investment from the workers than most of the other *HITs*. Thus it is quite difficult to find the reward sweet-spot between cost and result quality.

Compared to a lab environment it is also difficult to exploit a subject's intrinsic motivation or provide a social reward, as workers participate in the online experiment mostly for financial gain. Although many participants in lab studies are also mainly motivated by financial incentives, the *peer pressure* either by their fellow participants or by the test manager tends to be an incentive to do the work more properly. Using non-financial awards, the quality of the result can be similar or even better than purely financial awards [23].

4.4. Reliability Challenges

Lastly, we will shortly discuss the challenges with respect to the reliability of the results gained in crowd-based subjective testing compared to more traditional setups.

A common method in many crowdsourcing platforms to control the accuracy of the results is the defintion of *gold standard* data [26, 27]. The requester poses a task with a known answer and monitors the results from the test. Users that fail a certain number of such gold tasks because they have too high a deviation from the expected result can be disqualified from the test without any payment, recommitted to the instructions or excluded retroactively from the data set. In subjective video testing, however, usually the only data point that is known a priori is the reference video, that is assumed to have the highest quality level. Note, that this is also limited to scenarios where a reference is available for testing. To investigate a possible influence of using the reference material as a gold standard on the results, we conducted two tests, each with the same QCIF videos sequences from [28]. In one test, we had the interface tell the user to redo the training when their votes on the reference material were too low, in this case on the lower half of the quality scale. The other test was done without any gold. Fig. 5 shows the results of both tests. We can see that the introduction of a gold standard leads to a significant bias in the test results. As we only have reference data with high quality and therefore only discipline the user when his votes are too low, the videos with higher quality get rated too high. Videos with low quality, however, got quite similar results in both tests. We therefore conclude that it might be better to do the outlier processing after the test at the cost of a higher number of test subjects. More in general, the problem in crowd-based video quality assessment is that due to its intrinsically subjective nature, *gold standard* or ground truth data is not available. An alternative, however, can be *gold* not based on the visual quality itself, but rather on other known properties of the videos under test. For example, we can ask the subjects to describe the content of a video sequence. The same sequence will then be repeated later on in the test. By comparing the answers of both presentations of the video sequence, we can then assess if the subject did pay attention and answers consistently. The assumption is that the results provided by attentive and consistent subjects are more reliable [29].

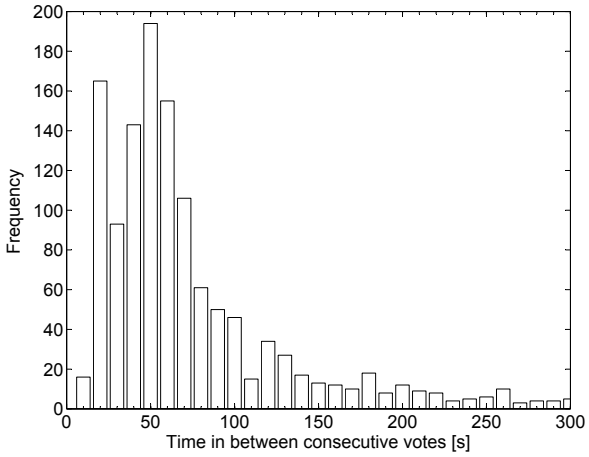

Fig. 6: The distribution of the answer times for the judgements shows that there were a significant number of test subjects that didn't watch the videos completely

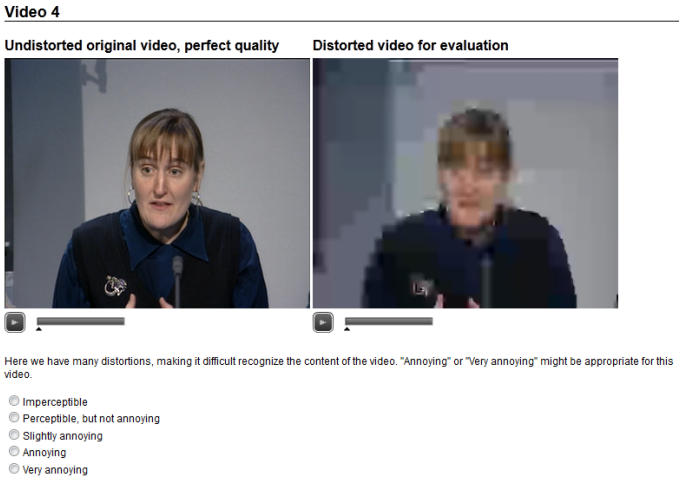

Fig. 7: Crowd-based video quality assessment: Training interface as seen by test participants (introduction text not shown) [2]

Another quite simple criterion to judge the reliability of the votes of one user is the time the user needs to process one video. While the QualityCrowd framework implements measures to ensure that votes can only be submitted when a video has been watched completely, we still had in [3] a significant number of votes where this protection was circumvented by the user as seen in Fig. 6. The data for this figure was gained from a double-stimulus video test where each video had a minimal length of 10 s. We processed 1,435 votes and measured the time between two consecutive votes of each test subject. Out of these intervalls there were 16 that were below 10 s, meaning that this user watched neither of the two videos completely and 165 votes that took between 10 s and 20 s, which probably stemmed from test subjects watching both the reference and the test video in parallel.

5. CONCLUSION

Crowd-based video quality assessment is a promising approach to reduce the financial and organisational burden of subjective video quality assessment. Although results so far have shown good correlation with traditional lab tests, for some methodologies there still remain many open issues.

We discussed in this contribution the challenges faced by crowd-based quality assessment from a conceptual, technical, motivational and reliability perspective compared to a traditional lab environment and subjective testing setup. While some of these challenges, e.g. non-standard test equipment, are inherent to the crowdsourcing principle, other challenges, e.g. reliability issues, may be overcome by improving the test design, especially taking into account the particularities of crowdsourcing.

Even though there still remains significant work to be done before crowd-based video quality assessment can be a viable alternative to subjective testing in the lab, we believe that the overall benefits to be gained by crowdsourcing will make it an important tool in subjective testing.

6. 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] C. Keimel, J. Habigt, C. Horch, and K. Diepold, "QualityCrowd - A Framework for Crowd-based Quality Evaluation," in *Picture Coding Symposium (PCS) 2012*, Krakow, Poland, May 2012, pp. 245–248.

[3] ——, "Video Quality Evaluation in the Cloud," in *Proceedings International Packet Video Workshop (PV) 2012*, Munich, Germany, May 2012.

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

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

[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] J. Howe, "The rise of crowdsourcing," *Wired Magazine*, vol. 14, no. 6, 2006. [Online]. Available: http://www.wired.com/wired/archive/14.06/crowds.html

[8] 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.

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

[10] 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.

[11] 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),*

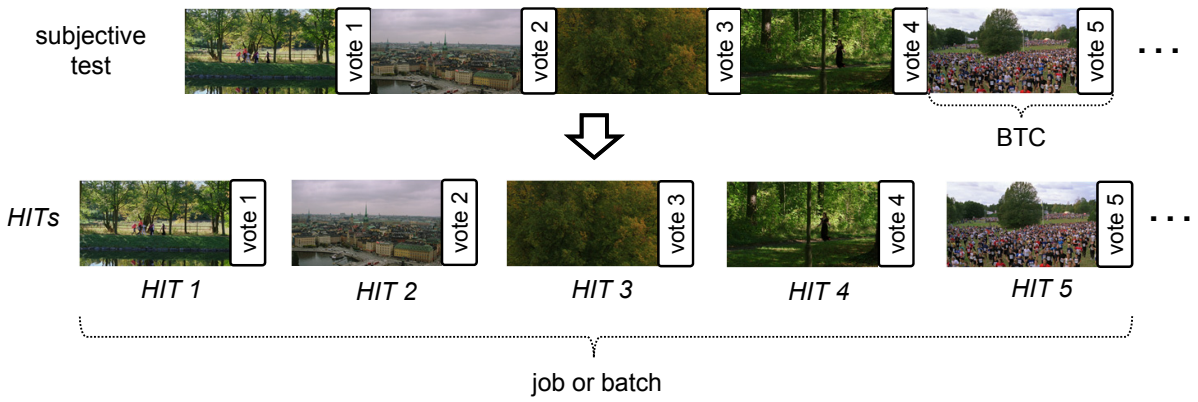

Fig. 8: Partitioning into *HITs*: a subjective test is split into its BTCs and each BTC is considered as a separate *HIT*. While all *HITs* are part of the same job or batch, neither the order of the *HITs* is fixed, nor is a subject required to complete all *HITs*

2011 IEEE International Conference on, May 2011, pp. 2416–2419.

[12] 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.

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

[14] T. Schulze, S. Seedorf, D. Geiger, N. Kaufmann, and M. Schader, "Exploring task properties in crowdsourcing - an empirical study on mechanical turk," in *ECIS 2011 Proceedings*, 2011.

[15] S. Faradani, B. Hartmann, and P. G. Ipeirotis, "What's the right price? pricing tasks for finishing on time," in *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*, Aug. 2011.

[16] P. Corriveau, C. Gojmerac, B. Hughes, and L. Stelmach, "All subjective scales are not created equal: The effects of context on different scales," *Signal Processing*, vol. 77, no. 1, pp. 1–9, 1999.

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

[18] T. Brandão and M. Queluz, "No-reference quality assessment of H.264/AVC encoded video," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 20, no. 11, pp. 1437–1447, Nov. 2010.

[19] C. Keimel and K. Diepold, "On the use of reference monitors in subjective testing for HDTV," in *Quality of Multimedia Experience (QoMEX), 2010 Second International Workshop on*, Jun. 2010, pp. 35–40.

[20] Amazon.com, Inc. (2012, Mar.) Amazon mechanical turk requester policies. [Online]. Available: https://requester.mturk.com/policies

[21] *ITU-R BT.1788 Methodology for the subjective assessment of video quality in multimedia applications*, ITU-R Std., Jan. 2007.

[22] J. J. Horton and L. B. Chilton, "The labor economics of paid crowdsourcing," in *Proceedings of the 11th ACM conference on Electronic commerce*, 2010, pp. 209–218.

[23] W. Mason and D. J. Watts, "Financial incentives and the "performance of crowds"," in *Proceedings of the ACM SIGKDD Workshop on Human Computation*, ser. HCOMP '09. New York, NY, USA: ACM, 2009, pp. 77–85.

[24] 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.

[25] S. Suri, D. Goldstein, and W. Mason, "Honesty in an online labor market," in *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*, Aug. 2011.

[26] E. Huang, H. Zhang, D. C. Parkes, K. Z. Gajos, and Y. Chen, "Toward automatic task design: A progress report," in *Proceedings of KDD-HCOMP'10*. ACM, 2010.

[27] A. Kittur, E. H. Chi, and B. Suh, "Crowdsourcing user studies with mechanical turk," in *Proceedings of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, ser. CHI '08. New York, NY, USA: ACM, 2008, pp. 453–456.

[28] M. Goudarzi, L. Sun, and E. Ifeachor, "Audiovisual quality estimation for video calls in wireless applications," in *Global Telecommunications Conference (GLOBECOM 2010), 2010 IEEE*, Dec. 2010, pp. 1–5.

[29] T. Hoßfeld, R. Schatz, M. Seufert, M. Hirth, T. Zinner, and P. Tran-Gia, "Quantification of YouTube QoE via Crowdsourcing," in *IEEE International Workshop on Multimedia Quality of Experience - Modeling, Evaluation, and Directions (MQoE 2011)*, Dana Point, CA, USA, Dec. 2011, pp. 494–499.