Exploring Crowdsourcing for Subjective Quality Assessment of 3D Graphics

Yana Nehmé, Patrick Le Callet, Florent Dupont, Jean-Philippe Farrugia, Guillaume Lavoué

Exploring Crowdsourcing for Subjective Quality Assessment of 3D Graphics

Yana Nehmé
LIRIS CNRS, Univ Lyon, France
yana.nehme@liris.cnrs.fr

Patrick Le Callet
LS2N CNRS, Univ Nantes, France
patrick.lecallet@univ-nantes.fr

Florent Dupont
LIRIS CNRS, Univ Lyon, France
Florent.Dupont@liris.cnrs.fr

Jean-Philippe Farrugia
LIRIS CNRS, Univ Lyon, France
jean-philippe.farrugia@univ-lyon1.fr

Guillaume Lavoué
LIRIS CNRS, Univ Lyon, France
glavoue@liris.cnrs.fr

***Abstract*—Multimedia subjective quality assessment experiments are the most prominent and reliable way to evaluate the visual quality as perceived by human observers. Along with laboratory (lab) subjective experiments, crowdsourcing (CS) experiments have become very popular in recent years, e.g., during the COVID-19 pandemic these experiments provide an alternative to lab tests. However, conducting subjective quality assessment tests in CS raises many challenges: internet connection quality, lack of control on participants' environment, participants' consistency and reliability, etc. In this work, we evaluate the performance of CS studies for 3D graphics quality assessment. To this end, we conducted a CS experiment based on the double stimulus impairment scale method and using a dataset of 80 meshes with diffuse color information corrupted by various distortions. We compared its results with those previously obtained in a lab study conducted on the same dataset and in a virtual reality environment. Results show that under controlled conditions and with appropriate participant screening strategies, a CS experiment can be as accurate as a lab experiment.**

***Index Terms*—crowdsourcing, laboratory, subjective quality assessment, 3D graphics, virtual reality, accuracy**

I. Introduction

Subjective quality assessment experiments are of primary importance for assessing the Quality of user Experience (QoE), understanding human physiological and psychological behavior (perception of multimedia content), benchmarking and tuning objective measures and algorithms. With the growing trend of machine learning, large datasets are needed and are in high demand, which has prompted researchers to invest more in subjective experiments.

Subjective quality assessment experiments have been traditionally conducted in laboratories in a controlled environment and with high-end equipment. In recent years, crowdsourcing (CS) experiments have gained quite a lot of popularity, especially with the development of the internet, and have been an alternative to laboratory (lab) experiments in certain cases, particularly during the COVID-19 pandemic, where participants could not be physically present in the laboratory to carry out tests. CS has been exploited in several fields and for different types of media, such as the evaluation and annotation of images, videos, audio, speeches and documents. CS and lab studies differ considerably in several aspects. (1) A task is performed by an unspecific internet crowd in the former rather than a specific group of people in the latter. Thus, CS enable researchers to access a much larger and more diverse subject pool and build generalized datasets representative of real-life scenarios. (2) Experiments conducted in a lab environment typically last around 20-30 minutes [1], while CS experiments should be kept as short as possible. Indeed, previous works [2]–[4] pointed out that a CS task should last 5-10 minutes to avoid participants' boredom, frustration and decreased attention, leading to unreliable behavior and results. (3) Regarding time-effort, CS experiments are dramatically less time-consuming than lab tests, especially when evaluating large datasets. (4) Last but not least, lab experiments allow better control of the study setup, while CS experiments are carried out in uncontrolled test environments (different viewing conditions that affect participants' perception of quality, e.g. lighting, bandwidth constraints, display device, distance between the participant and the viewing screen, etc.).

As a result, CS imposes several challenges to overcome compared to similar lab tests, notably those related to the lack of control over the participants' environment and the trustworthiness of the participants since they are not supervised in these tests. A thorough overview of these concerns can be found in [5], [6]. To detect and deal with malicious/unreliable participants, several mechanisms have been proposed over the years [5], [7], [8]. Despite these challenges, CS studies are still capable of producing accurate and reliable results if the experiment framework has been properly designed [9], [10].

Regarding the experimental methodologies used in crowdsourcing, most CS studies have used the pairwise comparison (PC) method as it is straightforward: the task of choosing one of the two stimuli is simpler than rating them on a discrete or continuous scale [9], [11], [12]. Other works have adopted the Absolute Category Rating (ACR) method [3], [10].

This work was supported by French National Research Agency as part of ANR-PISCo project (ANR-17-CE33-0005).

978-1-6654-3288-7/21$31.00©2021 IEEE

Available crowdsourcing frameworks already implement these methods and offer the possibility of modifying them to fit the needs of the study. A detailed overview of the crowdsourcing frameworks and an evaluation of them is provided in [13].

In this work, we investigated the accuracy and reliability of CS studies to assess the quality of 3D graphics. Thus, we conducted a CS experiment based on the Double Stimulus Impairment Scale (DSIS) method, which is, to the best of our knowledge, one of its first uses in CS studies. We used a dataset of animated 3D models [14]. The results of this experiment were compared to those obtained from a previous lab experiment conducted in a Virtual Reality (VR) environment. The design of the CS experiment is detailed in section 2, while its results are presented in section 3. Note that, the only 3D representation used in this work is meshes, however, we believe that our results remain valid for other 3D representations, such as point clouds, hence the use of the term "3D graphics".

II. Crowdsourcing experiment

A crowdsourcing (CS) experiment has been carried out in which subjects were recruited to assess the quality of an animated 3D graphics dataset. Since our goal is to compare the results of this experiment to previously collected Lab results [14], we replicated the lab experiment as much as possible, using the same instructions, dataset and subjective evaluation methodology. This section provides details on the CS experiment.

A. Dataset

The 3D graphic stimuli used in this study were previously used in the laboratory based subjective experiment, reported in [14]. This dataset contains 80 animated meshes with diffuse color information. It was generated from 5 source models (*"Aix", "Ari", "Chameleon", "Fish", "Samurai"*), corrupted by 4 types of geometry and color distortion that represent common simplification and compression operations: Uniform geometric quantization (QGeo), Uniform LAB color quantization (QCol), simplifications that take into account either the geometry only (SGeo) or both geometry and color (SCol). Each distortion was applied with 4 different strengths adjusted manually in order to cover the whole range of visual quality from imperceptible to high levels of impairment. The source models as well as examples of distortions are illustrated in [14].

B. Experimental procedure

In order to be able to compare the results of the CS and lab studies, we opted for the same setup as implemented in the lab experiment [14]: the stimuli were animated with a slow rotation (of 15° around the vertical axis in clockwise and then in counterclockwise directions) and displayed in a neutral virtual room (light gray walls) under the same viewpoints as those used in the lab experiment. Their material type complied with the lambertian reflectance model, and they were visualized without shadows and under a directional light.

The subjective testing methodology used in CS is also the same as the lab experiment, namely the Double Stimulus Impairment Scale method (DSIS), in which observers see the reference model and the same model impaired, simultaneously, side by side, for 10 sec and rate the impairment of the second stimulus in relation to the reference using a five-level impairment scale, displayed after the presentation of each pair of stimuli. Thence, we generated videos of the final rendered dynamic stimuli. The videos were all in 650 x 550 resolution (so that the videos of the reference and degraded models fit simultaneously on a screen with a minimum resolution of 1920 x 1080) with a frame rate of 30 fps and encoded using H.264 encoder (mp4 container) at a bitrate of 5 Mbps to ensure imperceptibility of compression artifacts. The duration of the videos is 10 sec, which corresponds to the display time of the models in the lab experiment.

To conduct the crowdsourcing subjective experiment, a web-based platform was developed suitable for presenting videos according to the DSIS method. It does not require the participants to install any software on their device (except a web browser with an H.264/AVC decoder). The platform first checks the screen resolution of the participants (must be equal to or higher than 1920 x 1080) as well as the page zoom level and ask them to keep the full screen mode until the end of the experiment. Otherwise, they are not allowed to proceed in the test. Once the device's compatibility has been verified, the test instructions are displayed to the participants. At the bottom of this page is a progress bar showing the status of the loading process of all the video pairs that will be used in the test. When the loading is completed a start button appears leading to the test. In this way, the videos of the reference and distorted models are ensured to be played simultaneously without any latency or unintended interruptions.

The experiment starts with a training, during which observers familiarize themselves with the task [1]. We selected the same training model used in the training of the lab experiment as well as its distorted versions which cover the whole range of distortions. After displaying each pair of training videos for 10 sec, the rating interface is displayed for 5 sec and an example score assigned to this distortion is highlighted. Once the training is completed the actual test starts. The videos of the stimuli are displayed in a random order (3D models, distortion types and levels all mixed) to each participant. Each video/stimulus is presented once; participants are not able to replay the videos. Moreover, participants are not able to provide their score unless the videos have been played completely. There is no time limit for voting and videos of the stimuli are not shown during that time. At the end of the experience, participants will receive unique codes allowing them to get their remuneration. Figure 1 illustrates the graphical interface of our CS experiment.

C. Test sessions and participants

As stated in the introduction, CS experiments should be kept as short as possible to keep participants motivated and to avoid unreliable results. Therefore, we divided our dataset of 80

Fig. 1. The display interface based on the DSIS method (left) and the rating interface (right) of our CS experiment.

stimuli into 4 groups, called playlists (i.e. 20 stimuli/playlist), and each participant evaluates one playlist of the dataset. The stimuli were distributed evenly across the playlists so that each playlist contained the 5 source models and the 4 distortion types and strengths. Furthermore, as this database encloses the subjective scores obtained in the lab test, we opted to distribute the stimuli between the playlists so as to have the same Mean Opinion Scores (MOSs) distribution in these playlists.

Since participants in CS cannot be supervised, it is important to ensure the quality of annotations by detecting unreliable and malicious participants. To do so, there are 3 common/popular strategies [15]. The first one, known as gold standard, is the insertion of dummy stimuli or stimuli with trusted annotations. The second strategy (a.k.a consistency question) is to collect multiple scores for repeated stimuli, which allows to assess participant's consistency, while the third strategy is a grading task in which a participant looks at several annotated images and scores every annotation. Inspired by the first and second strategies and as in [6], we injected 3 trapping stimuli, that we called golden units, into the playlists of our test. They consisted of (1) a very poor quality stimulus (high level of impairment), (2) a very high quality stimulus (the reference is compared to itself) and (3) a stimulus displayed twice to assess the participant's consistency (coherence of his/her scores). Participants who fail to answer the golden units correctly are considered outliers and their scores are rejected (more details in section IV-A).

Thus, the test session of our CS experiment is constituted of 23 pairs of videos to rate (1 playlist) and lasts about 10 minutes: informed consent + loading videos + instructions + 6 training stimuli x (10s video length + 5s Rating) + 23 test stimuli x (10s video length + $\sim$ 4s Rating).

We ran our experiment until each playlist was fully rated by 60 participants, which required approximately 12 hours. A total of 240 participants took part in this study: 148 males and 92 females. They were from 33 different countries and aged between 18 and 68. All participants were naive about the purpose of the experiment. Note that, participants who started the experiment and did not complete it (101 participants) were discarded. The recruiting process of the participants was performed using Prolific[1], an internet marketplace that provides tens of thousands of trusted participants. Only participants having a high reliability score (score based on how well they did in past studies) and an adequate number of duly completed jobs (number that reflects their familiarity with the platform) on Prolific were admitted to the experiment.

[1]https://www.prolific.co/

III. Lab experiment

We have previously conducted a lab experiment using this dataset to assess the perceived quality of 3D graphics in virtual reality (VR) [14]. The experiment was based on the DSIS method and conducted in a VR environment, using the HTC Vive Pro headset. Stimuli, animated with the same slow rotation as the CS test, were rendered for 10 seconds in a virtual scene at a fixed distance from the observer and rotated in real time. The study involved 30 participants: students and professionals at the University of Lyon. Each participant evaluated the whole dataset (80 stimuli) in one session that lasted about 23 minutes. The rating scores collected in the lab experiment are compared to those obtained in the CS experiment in the following section.

IV. Results and comparison of experiments

This section presents the results of the crowdsourcing experiment described above and compares them to the lab results in terms of accuracy, confidence intervals, and participants' agreement. For the subsequent analyses, subjective scores ranging from very annoying to imperceptible are mapped on a discrete numerical scale from 1 to 5. Note that the scores assigned to the golden units are only taken into account in observers screening and are not considered in the rest of the analyzes.

A. Participants screening

Before starting any analysis, it is necessary to identify and remove outliers which could affect the accuracy of the results. The participants of the lab experiment were filtered according to the ITU-R BT.500-13 recommendation [1]. We found 1 outlier (i.e. 29/30 subjects remain).

For the CS study, as recommended in [5], participants were screened by combining: (1) the ITU-R BT.500-13 screening procedure, which revealed 7 outliers among the 240 participants and (2) the golden units analysis (trapping stimuli): we found that 4 participants incorrectly rated the very poor quality stimulus (i.e. its distortion is rated as imperceptible or

perceptible but not annoying), 2 participants misjudged the very high quality stimulus (i.e. its distortion is considered very annoying or annoying; however, this stimulus is not degraded.), and lastly, 1 participant gave inconsistent scores to the third golden unit, called G (i.e. $\left|s_i^{G_{rep1}} - s_i^{G_{rep2}}\right| \geq 3$, where s_i^G denotes the score assigned by participant i to stimulus G, shown twice $rep1$ and $rep2$). As a result, a total of 7 participants failed to rate the golden units, 3 of which were detected by the ITU-R BT.500-13 screening procedure. Thus, 11 participants were rejected (ITU-R outliers $\bigcup$ Golden units outliers), i.e. 229/240 subjects remain. After outliers removal, each stimulus is rated by 29 participants in the lab test and by at least 56 participants in CS test. Only the scores of these screened participants will be used in our subsequent analyzes.

B. Resulting MOSs

In order to analyze the results of the CS experiment, we computed for each stimulus the Mean Opinion Score (MOS), rating scores averaged over all participants. We compared them to those obtained from the lab experiment. Figures 2, 3 and 4 illustrate the results.

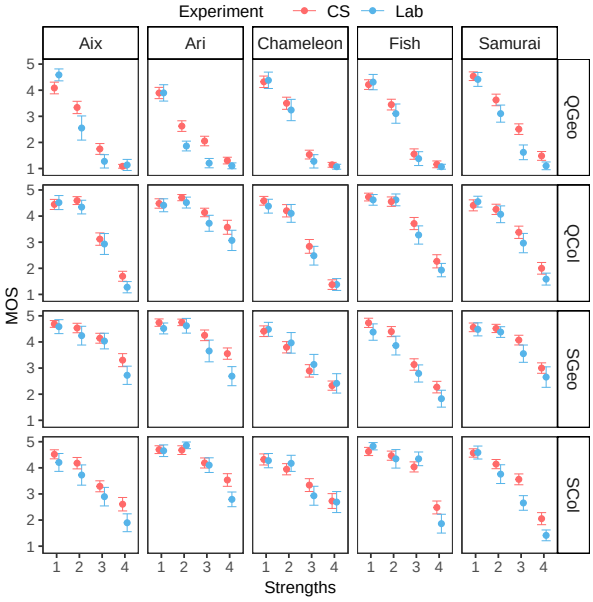

Fig. 2. Comparison of the mean opinion scores of the lab and CS experiments, for all the stimuli. Results are grouped by source models and types of distortion.

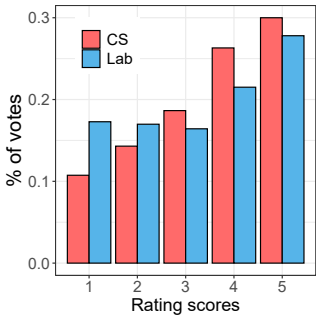

Fig. 3. Score distributions for the lab and CS experiments.

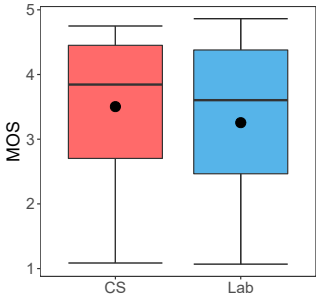

Fig. 4. Boxplots of MOSs obtained for the lab and CS experiments.

Figure 2 shows that the MOS results of the CS experiment strongly correlate with those of the lab test. Indeed, the (Pearson, Spearman's rank) correlation coefficients between CS's and lab's MOSs are (0.975, 0.954). Overall, no significant difference was found between the MOS means of the 2 studies (p-values = 0.191).

For both the lab and CS experiments, MOSs decrease as distortions strengths increase. However, we can notice some differences in the distribution of the scores and the use of the rating scale in the 2 experiments (see Figures 2 and 3): overall, the stimuli rated in the lab scored lower than those in CS. This effect is more visible for high strength distortions (strengths ≥ 3), meaning that the lab test participants were able to detect some distortions that the CS participants missed. Indeed, the lab test was conducted in a VR environment. Therefore, we believe that detecting visual quality losses of stimuli is easier in VR than on a 2D screen (CS), since VR headsets provide a bigger/wider field of view (FoV) than a desktop setup and so the size in terms of visual angle of objects in VR is considerably larger than on screen (the stimuli size is approximately 37 and 18 degrees of visual angle in VR and on-screen, respectively).

C. Confidence intervals

Since participants in a CS experiment are not supervised, the evaluation of confidence intervals (CIs), i.e. the dispersion of individual scores, is particularly important. We therefore calculated the 95% CIs of the MOSs for the CS study and compared them to those obtained in the lab test. We assessed the evolution of the CIs width according to the number of ratings collected per stimulus (which is indirectly related to the number of participants involved in the test). Thus for each stimulus, we considered all possible combinations (without repetition) of N ratings and averaged the width of the CIs over all these ratings combinations. $N \in [1, 29]$ and $[1, 56]$ for the lab and CS test, respectively. Results are shown in Figure 5.

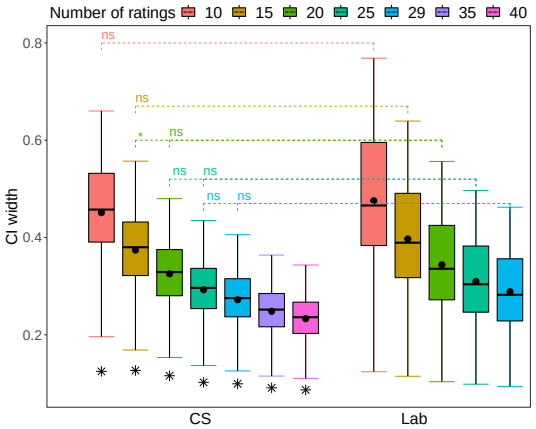

Fig. 5. Variation of CIs according to the number of ratings per stimulus in the CS and lab experiments. ns and (*) refer to not statistically significant (p-value ≥ 0.05) and $0.01 \leq$ p-value < 0.05, respectively.

Overall, for a given number of ratings, the CIs of the lab test are slightly larger than those of the CS test. This

difference is not significant, according to the results of the t-tests (at a level of significance of 5%). Thus, both studies showed similar agreement between the participants' scores. Participants' agreement is further explored in section IV-E.

D. Accuracy of subjective scores

We investigate, in this section, the accuracy (discrimination ability) of the CS test and compare it to that of the lab test. A more accurate test should yield in a larger number of pairs of stimuli whose quality can be considered different in a statistical test. As in [14], we conducted two-samples Wilcoxon tests between rating scores of each possible pairs of stimuli and computed the percentage of pairs of stimuli rated significantly different (p-value < 0.05) among all the possible pairs ($80 \times 79/2 = 3160$ pairs).
Similar to the previous subsection, we evaluated the evolution of accuracy as a function of the number of ratings per stimulus: for a given stimulus and number of ratings N, we averaged the number of pairs of stimuli rated significantly different over the possible combinations of ratings. Figure 6 shows the results.

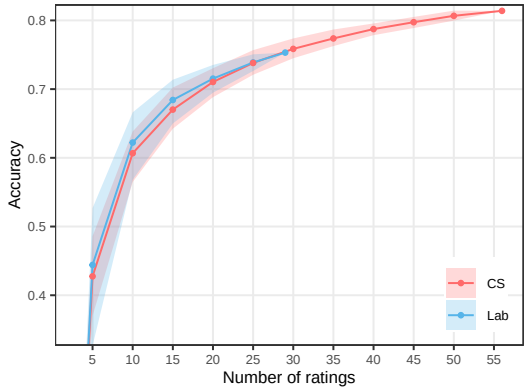

Fig. 6. Variation of the accuracy according to the number of ratings per stimulus for the CS and lab experiments. The accuracy (y-axis) is defined as the percentage of pairs of stimuli whose qualities were assessed as statistically different. Curves represent mean values of these percentages and areas around curves represent 2.5th - 97.5th percentiles.

Ratings of the lab test are slightly more accurate than those of the CS test. For instance, at least 17 rating scores per stimulus (equivalent to 17 participants) are needed in the lab experiment to achieve accuracy with an overall level of 70%, whereas this number increases to 19 ratings in the CS experiment (corresponding to at least 76 participants in our actual CS test setup). Although there was no significant difference between the CIs of the 2 experiments (see section IV-C), the lab study produces slightly more accurate results, as the lab participants used the rating scale better (as shown in section IV-B).

Regarding the overall trend of the curves and despite this slight difference in the accuracy, the DSIS method offers consistent performance in the lab and CS experiments.

E. Participants' agreement / consistency

In this section, we evaluate the agreements between the participants. To do this, we evaluated the internal consistency of participants' data, as proposed by [6]. For each stimulus, we randomly split the participants who rated it into two equal size groups and computed the Spearman's rank correlation between the MOSs of the 2 groups. After 500 splits, the range of correlations was between 0.94 and 0.978 (with a mean of 0.963) in the CS experiment, and between 0.898 and 0.965 (with a mean of 0.934) in the lab experiment. Results show a high degree of inter-subject agreement in both experiments. This agreement is slightly lower in the lab experiment, possibly due to its more complex immersive viewing environment. This is in line with the CIs being slightly larger in the lab test (see section IV-C).

F. Content ambiguity

As stated in [16], some contents tend to be more difficult to rate than others. This is known as "Content ambiguity" and can be estimated by the Maximum Likelihood Estimation (MLE) model [16]. Therefrom, we computed the ambiguity values of the 5 source models that constitute our dataset. Since the content ambiguity obtained by the MLE model [16] depends on the participants, we considered the same number of rating per stimulus ($N = 29$) for the lab and CS experiments in order to be able to compare their results. Thus, for the CS experiment, we randomly selected 100 combinations of 29 ratings for each stimulus. We averaged the ambiguity values over these combinations of ratings.

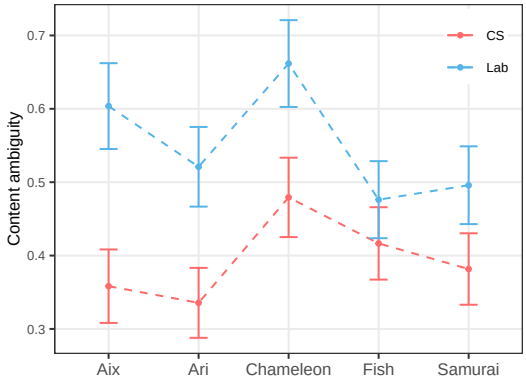

Fig. 7. Content ambiguity of each source model associated its CI (calculated as described in [16]), for the CS and lab experiments.

Figure 7 shows that all the source models are more ambiguous in the lab experiment than in the CS experiment. This may be because in VR, due to the larger FoV (as explained in section IV-B), participants can see more details and low-level features, which makes the task of assessing the differences between the reference and the distorted stimulus more difficult. The *Chameleon* model is associated with the highest content ambiguity in both experiments.

Content ambiguity is related to the dispersion of subjective scores (CIs). Therefore, we averaged the CIs of the stimuli over the models. For both experiments, we obtained the same shape of curves as in Figure 7, meaning that models with high CIs are also associated with high ambiguity values.

V. Discussion

We designed a subjective experiment for assessing the quality of 3D graphics in crowdsourcing (CS). The experiment was based on the DSIS method. Since in CS experiment, participants are unsupervised, we sought to impose controls on several aspects: (1) we used the Prolific platform which is more selective in recruiting and filtering participants than other similar platforms such as Mturk, Microworkers, etc. (2) We pre-screened participants during the recruiting process based on their reliability score on this platform. (3) We tried to control the viewer's environment as much as possible (e.g. minimum screen resolution required, maintain a full screen mode, limited interactions to rating). Finally, (4) we combined different screening strategies to identify outliers. The results of this experiment were compared to results collected from a previous lab experiment conducted in a VR environment.

Results showed that under controlled conditions and with a precise/proper participant screening/filtering approach, a CS experiment can be as accurate as a lab experiment. Indeed, we obtained a high correlation between the mean opinion scores of the 2 experiments as well as a good agreement between the participants' ratings (CIs). This agreement was slightly lower in the lab test which can be due to the VR environment. Our findings corroborate previous studies, by [6], [10], [12], which also achieved high correlation with the lab results, and furthermore, they highlight the quality of the Prolific participants involved in our CS study, their reliability and seriousness despite the fact that the participants were not supervised. It is worth mentioning that the golden units approach we implemented seems to work well. In fact, all the outliers detected by the inspection of golden units were found to have a high inconsistency according to MLE model [16]. Regarding the experimental methodology, DSIS seems accurate and adapted for evaluating the quality of 3D graphics in crowdsourcing. We believe that this due to the fact that this method presents explicit references which simplifies the task and makes it easier: according to a short questionnaire asked at the end of our CS test, 89% of the participants found the experiment's instructions clear and 94% rated the work as easy.

In term of time-effort, it took us 12 hours to collect all the data in the CS experiment (5520 quality judgments collected), compared to 36 hours in the lab experiment (2400 quality judgments collected). Crowdsourcing is quite faster. It is frequently used to evaluate large datasets, which requires weeks of laboratory evaluations. However, building and designing our CS experiment tool was a time-intensive task and required significant software development effort (both technical and of conceptual challenges). Furthermore, as stated earlier, the CS experiment must be short to keep participants accurate and consistent, making our CS test require additional programming effort to implement a tool that evenly divides the dataset into batches/playlists and assigns a different batch to each participant.

VI. Conclusion

In this work, we investigated how crowdsourcing could be used for quality assessment of 3D graphics. CS experiments based on the DSIS method appear to be a promising way not only to quickly collect a large amount of realistic quality judgments, but also to provide (when combined with appropriate participant screening strategies) accurate and reproducible results. We expect our findings to help the scientific community when designing subjective quality assessment studies in CS. Further experiments are still needed to find the best compromise between the number of stimuli, the duration of the test and the accuracy of the results. Also, it would be very interesting to repeat the same experiment in the lab, this time using a desktop setup, to clearly isolate the effects of VR.

References

[1] I.-R. BT.500-13, "Methodology for the subjective assessment of the quality of television pictures BT Series Broadcasting service," 2012.

[2] J. Redi, E. Siahaan, P. Korshunov, J. Habigt, and T. Hossfeld, "When the crowd challenges the lab: Lessons learnt from subjective studies on image aesthetic appeal," in *Proceedings of the Fourth International Workshop on Crowdsourcing for Multimedia*, ser. CrowdMM '15, 2015, p. 33–38. [Online]. Available: https://doi.org/10.1145/2810188.2810194

[3] R. Z. Jiménez, L. F. Gallardo, and S. Möller, "Influence of number of stimuli for subjective speech quality assessment in crowdsourcing," in *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, 2018, pp. 1–6.

[4] M. Reimann, O. Wegen, S. Pasewaldt, A. Semmo, J. Döllner, and M. Trapp, "Teaching Data-driven Video Processing via Crowdsourced Data Collection," in *Eurographics 2021 - Education Papers*, 2021.

[5] 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," *IEEE Transactions on Multimedia*, vol. 16, no. 2, pp. 541–558, 2014.

[6] D. Ghadiyaram and A. C. Bovik, "Massive online crowdsourced study of subjective and objective picture quality," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 372–387, 2016.

[7] J. Vuurens, A. P. de Vries, and C. Eickhoff, "How much spam can you take? an analysis of crowdsourcing results to increase accuracy," in *Proc. ACM SIGIR Workshop on Crowdsourcing for Information Retrieval (CIR'11)*, 2011, pp. 21–26.

[8] J. Li, S. Ling, J. Wang, and P. Le Callet, "A probabilistic graphical model for analyzing the subjective visual quality assessment data from crowdsourcing," in *Proceedings of the 28th ACM International Conference on Multimedia*, ser. MM '20, 2020, p. 3339–3347. [Online]. Available: https://doi.org/10.1145/3394171.3413619

[9] 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, 2009, p. 491–500. [Online]. Available: https://doi.org/10.1145/1631272.1631339

[10] F. Ribeiro, D. Florencio, and V. Nascimento, "Crowdsourcing subjective image quality evaluation," in *2011 18th IEEE International Conference on Image Processing*, 2011, pp. 3097–3100.

[11] T.-K. Huang, C.-J. Lin, and R. C. Weng, "Ranking individuals by group comparisons," in *Proceedings of the 23rd International Conference on Machine Learning*, ser. ICML '06, 2006, p. 425–432. [Online]. Available: https://doi.org/10.1145/1143844.1143898

[12] J. SØgaard, M. Shahid, J. Pokhrel, and K. Brunnström, "On subjective quality assessment of adaptive video streaming via crowdsourcing and laboratory based experiments," *Multimedia Tools Appl.*, vol. 76, no. 15, p. 16727–16748, Aug. 2017. [Online]. Available: https://doi.org/10.1007/s11042-016-3948-3

[13] T. Hoßfeld, M. Hirth, P. Korshunov, P. Hanhart, B. Gardlo, C. Keimel, and C. Timmerer, "Survey of web-based crowdsourcing frameworks for subjective quality assessment," in *2014 IEEE 16th International Workshop on Multimedia Signal Processing (MMSP)*, 2014, pp. 1–6.

[14] Y. Nehmé, J.-P. Farrugia, F. Dupont, P. L. Callet, and G. Lavoué, "Comparison of subjective methods for quality assessment of 3d graphics in virtual reality," *ACM Trans. Appl. Percept. (TAP)*, vol. 18, no. 1, Dec. 2021. [Online]. Available: https://doi.org/10.1145/3427931

[15] A. Sorokin and D. Forsyth, "Utility data annotation with amazon mechanical turk," in *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2008, pp. 1–8.

[16] Z. Li and C. G. Bampis, "Recover subjective quality scores from noisy measurements," in *2017 Data Compression Conference (DCC)*, 2017, pp. 52–61.