Reliability of Crowdsourcing for Subjective Quality Evaluation of Tone Mapping Operators

Abhishek Goswami, Ali Ak, Wolf Hauser, Frédéric Dufaux, Patrick Le Callet

Reliability of Crowdsourcing for Subjective Quality Evaluation of Tone Mapping Operators

Abhishek Goswami[†‡], Ali Ak[*], Wolf Hauser[†], Patrick Le Callet[*], and Frederic Dufaux[‡]
[†]DxO Labs, France
Emails: {agoswami, whauser}@dxo.com
[*]Image Perception Interaction Team, LS2N, University of Nantes, France
Emails: {ali.ak, patrick.lecallet}@univ-nantes.fr
[‡]Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des Signaux et Systèmes, France
Emails:{abhishek.goswami, frederic.dufaux}@l2s.centralesupelec.fr

***Abstract*—Tone mapping operators (TMO) are functions which map high dynamic range (HDR) images to limited dynamic media while aiming to preserve the perceptual cues of the scene that govern its aesthetic quality. Evaluating aesthetic quality of TMOs is non-trivial due to the high subjectivity of preference involved. Traditionally, TMO aesthetic quality has been evaluated via subjective experiments in a controlled laboratory environment. However, the last decade has brought a surge in popularity of crowdsourcing as an alternative methodology to conduct subjective experiments. However, uncontrolled experiment conditions and unreliability of participant behaviour puts doubts on the trustworthiness of the collected data. In this study, we explore the possibility of using crowdsourcing platforms for subjective quality evaluation of TMOs. We have conducted three experiments with systematic changes to investigate the effect of experiment conditions and participant recruitment methods on the collected subjective data. Our results show that subjective evaluation of TMO aesthetic quality can be conducted on Prolific crowdsourcing platform with negligible differences in comparison to laboratory experiments. Furthermore, we provide objective conclusions about the effect of number of observers on the certainty of the pairwise comparison results.**

***Index Terms*—Subjective image quality, HDR, Tone mapping, Crowdsourcing**

I. Introduction

The objective of tone mapping is not just to reduce the dynamic range for better representation of the scene but also to preserve the perceptual cues for the human visual system to maintain the aesthetic quality of the scene. The advent of hyper-realistic multimedia has expedited the consumption of HDR content. Hence evaluating quality of tone mapped images has been a pertinent topic of research.

Evaluating quality of tone mapped images is subjective because the process affects the visual cues of a scene. Research on HDR imaging has produced many TMOs, however, due to the aforementioned subjectivity of the results, evaluating TMOs remains a non-trivial problem. Researchers have identified several objective factors like *brightness, contrast, colourfulness, structural fidelity* etc. to come up with objective metrics for image quality assessment (IQA) [1], [2]. However, the results of the metrics are often difficult to generalise and remain far from the subjective opinion, to be treated as the ground truth.

Subjective quality assessment is a more accurate method for evaluating TMOs. It is also essential to develop and optimize objective quality metrics. Over the last decade, crowdsourcing has gained popularity as a cost, time and resource efficient way to conduct subjective experiments. Crowdsourcing allows us to collect large amount of data in a short amount of time with minimal human interaction and is highly scalable. However, uncontrolled experimental conditions and unreliability of the participants has often put a barrier to a mass adoption of such platforms. Perceived image quality heavily depends on the visibility of distortions and which may be enhanced or masked depending on the viewing conditions such as display device, viewing distance, background luminance [3]. In aesthetic image quality evaluation scenario, distortion visibility plays a minimal role. Thus, subjective preferences may have desirable similarity between different viewing conditions. This has led us to investigate the possibility of using crowdsourcing platforms as a resource efficient platform to collect subjective preferences. Therefore, in this paper, we seek answers to the following questions: *"Can crowdsourcing platforms be used for TMO evaluation without compromising on the gathered data? What are the effects of experimental conditions and participant recruitment methods on the subjective preferences? What are the effects of number of observers on the certainty of the results?"*

II. Related Works

We have handpicked four TMOs for our subjective evaluation. A fairly recent comparative subjective study by Cerda-Company [4] suggested that TMOs by Kim et al. [5], Krawczyk et al. [6] and Reinhard et al. [7] have performed better in comparison to several other TMOs under varied scenarios. The final TMO is based on the recent work of Goswami et al. [8], Semantic TMO which presents a new approach of semantic-aware tone mapping.

In literature, there are several studies on subjective evaluation of TMOs [4], [9]. Based on the use-case of the study, TMO quality evaluation can be designed with the presence

This work has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie Grant Agreement No. 765911 (RealVision)

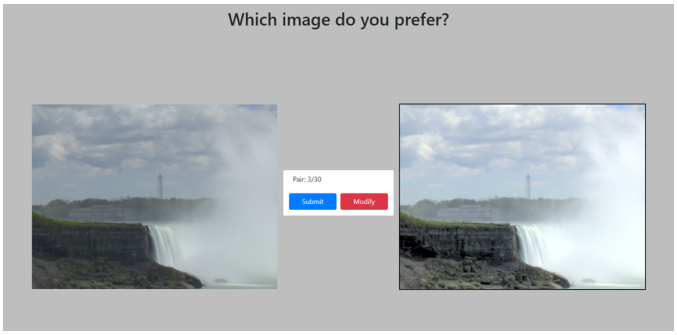

Fig. 1. Example test screen.

of the reference HDR image *i.e., Full-reference*, or without it *i.e., No-reference*. For an aesthetic quality evaluation use-case such as ours, a no reference methodology is preferred.

Crowdsourcing platforms such as Prolific [10], AMT [11] and Microworkers [12] have not been extensively utilised for subjective evaluation of TMOs. There is a recent large-scale study where crowdsourcing has been adopted to collect subjective preferences on aesthetic evaluation of HDR processing [13]. However, the dataset contains different types of algorithms to process HDR images and does not provide a TMO comparison for the same source image. As also pointed out in the aforementioned study, subjective data collected via crowdsourcing may be noisy due to the lack of control in the experiment conditions. Several methods have been developed to filter unreliable observers and noisy data [14], [15]. Ak et al. [14] provided comprehensive analysis on the methods to detect unreliable observers in a crowdsourcing experiment for aesthetic quality evaluation of TMOs. In literature, there are also traditional methods, such as reliability checks [16] and *gold standards* [17] where previously obtained results from reliable participants have been compared to crowdsourcers' responses to detect unreliable behaviours. Aesthetic evaluation may depend on several human factors which govern the observer preference for images. Such factors, which can be regulated in the controlled laboratory environment, are absent in the crowdsourcing setup [16]. Gadiraju et al. [18] focused on such contributing factors when dealing with human-centric experiments via crowdsourcing.

III. Experimental Design

We conducted 3 different experiments with systematic changes in order to investigate the accuracy of the data collected from Prolific crowdsourcing platform [10] for TMO evaluation. In the following subsections, we describe the experiment setup, the dataset and the platforms used for each experiment.

A. Experiment Setup & Procedure

Subjective evaluation of TMOs can be conducted with a full-reference or a no-reference methodology. Our study aims to compare TMOs among each other on the basis of observed aesthetic quality rather than comparing their naturalness, fidelity or proximity to the original scene. Therefore

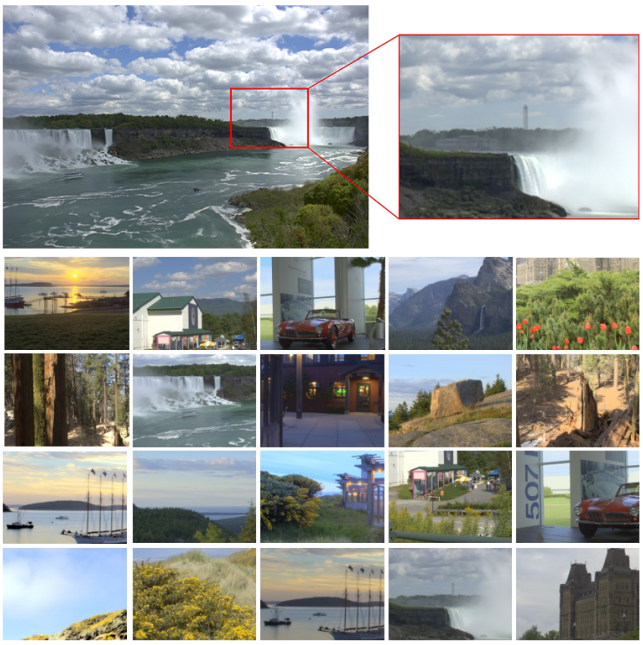

Fig. 2. Cropping high resolution images to create 20 SRCs.

no-reference methodology is more suitable for the task [19]. Furthermore, on crowdsourcing platforms it is practically difficult to conduct a full-reference experiment as it would require an HDR screen for each observer. Therefore, we follow a no-reference design to collect subjective preferences.

We adopt the forced-choice pairwise comparison (PC) method in the conducted experiments. It simplifies the evaluation task for the observers, therefore increasing the reliability of the collected preferences. Compared to alternatives, such as absolute category rating, the number of comparisons for the same number of content in PC is exponentially higher. Adaptive designs can be adopted to reduce the number of comparisons [20], resulting in unbalanced number of observations. Although, it may not be efficient to use adaptive designs in online platforms such as Prolific [10]. An application programming interface (API) is necessary to be able to benefit from such designs which is not available on every crowdsourcing platform. Additionally, since the aim of our study is to compare different platforms, unbalanced number of observations may result in unfair judgement. Therefore, we follow a full PC design in our experiments. An example test screen used in the experiments is shown in Fig. 1 where the observers are tasked to choose the image they prefer.

B. Stimuli & Database

For the creation of the TMO evaluation dataset, we selected 20 source contents (SRC) from Fairchild's HDR dataset [21]. The spatial resolution of the images in Fairchild dataset is fairly high. Thus we scale down and systematically crop the images to a resolution of 640×480. It allows us to display the stimuli side by side on display devices with 1080p resolution. We further shortlist a selection of crops using

TABLE I
OBSERVER STATISTICS

	Number of Unique Obs.	Mean Age (Years)	Gender Female / Male	Avg. Time Per Comparison (Seconds)
Exp-Lab	40	33.5	22 / 18	7.49
Exp-Online	50	22.6	28 / 22	4.33
Prolific	100	28.5	116 / 284	3.64

their absolute dynamic range and the entropy of their salient features in order to promote challenging content. Afterwards, the shortlisted crops are clustered based on TMQI [2] scores of the tone mapped images. Finally, we select a total of 20 SRC among the clusters. Fig. 2 contains a cropping example in addition to 20 SRCs used in the experiment. SRCs in the figure are tonemapped for visualization purposes using the ReinhardTMO [7] implementation from Banterle's HDR Matlab toolbox [22].

Four tone mapping operators, ReinhardTMO [7], Krawczyk-TMO [6], KimKautzTMO [5] and SemanticTMO [8], have been selected from literature. We have tone mapped 20 SRC using the selected TMOs using ReinhardTMO, Krawczyk-TMO and KimKautzTMO, implementations from the Matlab HDR Toolbox [22]. Adjustable parameters of each TMO have been optimized to maximise their respective TMQI scores [2].

In the end, we have compiled a dataset with 20 SRCs tone mapped using 4 TMOs, resulting in a dataset[1] with 80 tone mapped images and 120 unique pair of comparisons without cross content inclusion. We understand that creating a global ranking scale via cross content comparisons is not intuitive for us as we evaluate content specific aesthetic quality across TMOs.

C. Subjective Experiment Platforms

The primary experiment, *Exp-Lab*, was conducted within a controlled laboratory facility. The experiment conditions were set as recommended in ITU-R BT.500-14 [23]. Grundig Fine Arts 55 FLX 9492 SL is used to display the image pair side by side. 40 observers, 22 female and 18 male, who are not experts in image quality domain, were recruited through the university mailing list. The average age of the participants was 33.5 years. Each participant was checked for visual acuity with Monoyer test and color perception with Ishihara tests. Each observer provided their preferences for all of the 120 pairs in the dataset with a break after the 60^{th} pair. The average time taken per comparison was 7.49 seconds for an observer.

The second experiment, *Exp-Online*, was conducted with the same stimuli and experiment design. Participants were recruited through the same mailing list used for the *Exp-Lab* experiment. Each observer conducted the experiment on their own devices in their desired uncontrolled environment. 50 observers, 28 female and 22 male with 22.6 years average age, were recruited in total. Due to lower attention span of participants in crowdsourcing experiments, we split the initial dataset into 4 playlists of 5 SRCs with 30 comparisons in each [24] [14]. Each participant was asked to complete all 4 playlists without any constraint on the break between playlists. The average time taken for an observer was 4.33 seconds per comparison.

The third experiment, *Prolific*, was conducted with the same stimuli and experiment design on the Prolific [10] crowdsourcing platform. Unlike the first two experiments, observers were recruited through Prolific participants pool. 400 participants, 116 female and 284 male, were recruited from more than 20 different countries, majority being from Europe. Mean age of the participants was 28.5 years. Similar to the *Exp-Online* experiment, we split the initial dataset into 4 playlists of 5 SRCs with 30 comparisons in each. 100 unique participants evaluated each playlist. The average time spent per comparison was 3.64 seconds. Table I summarises the demographic information and statistics regarding to observers.

[1]Dataset and implementation codes for analysis can be found at ftp://ftp.polytech.univ-nantes.fr/TMOEval_PilotStudies

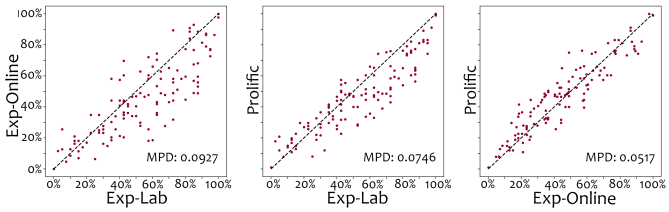

Fig. 3. Scatter plot comparison for the conducted experiments. Each point represents an image pair from the dataset. Axis values represent the percentage of votes for the same image in a pair. MPD is the mean of the perpendicular distances of the points from the diagonal.

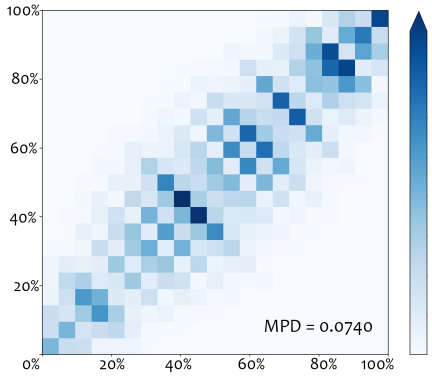

Fig. 4. Subjective preference comparison distributions of randomly split halves over 1000 permutations for Exp-Lab experiment. MPD value represents the mean perpendicular distance across all permutations.

IV. EVALUATION AND RESULTS

As described in the earlier sections, we conducted the same experiment on three different platforms with minimal difference in the setup. Using the collected data we continue to observe whether crowdsourcing methodology can be reliably used for aesthetic evaluation of TMOs. In the following subsections, we first evaluate the similarity between the results collected across the experiments. Furthermore, we investigate

TABLE II
COMPARING STATISTICALLY SIGNIFICANT DIFFERENCES BETWEEN IMAGE PAIRS ACROSS 3 EXPERIMENTS .

Comparison	Agreement	Disagreement	Contradiction
Exp-Lab vs Exp-Online	73	38	9
Exp-Lab vs Prolific	89	27	4
Exp-Online vs Prolific	89	31	0

the agreement among observers in the conducted experiments. Finally, we use the permutation test to quantify the effect of the number of observers on the certainty of the pairwise preferences.

A. Pairwise Preference Similarity Between Experiments

Fig. 3 presents a qualitative comparison of the conducted experiments. The plots compare the preference behavior of observers between the experiments. Each point, corresponding to an image pair in the dataset, plots the percentage of times Image-A in A-B comparison was preferred. Each axis represents the labeled experiment in comparison. Additionally, we compute the Mean Perpendicular Distance (MPD) to quantify the similarity between the experiments. MPD is calculated as the mean value of the perpendicular distance of each point from the diagonal. In case of a perfect agreement between the experiments, each image pair should lie on the diagonal. Therefore, a smaller MPD indicates a higher similarity between the compared experiments. Based on this, we observe that the distribution between *Exp-Lab* and *Prolific* experiment results are more linear and less scattered compared to *Exp-Lab* and *Exp-Online*, indicating a higher similarity for the former. We observe even higher similarity between *Exp-Online* and *Prolific* experiments despite the uncontrolled environmental conditions in both experiments. After comparing experiment results relative to each other, we use permutation test to compute an expected MPD value. We split the observers from *Exp-Lab* experiment into two disjoint groups and compare their cumulative preferences for 1000 iterations. We use this as a baseline to evaluate the cross-experiment agreements. Distribution of the permutation results is plotted as a two dimensional histogram in Fig. 4. Average MPD across 1000 iterations is calculated to be 0.0740. As reported in Fig. 3, MPD between *Exp*-Lab and *Prolific* is computed to be 0.0746, suggesting a desirable similarity between *Exp-Lab* and *Prolific* results.

In addition, we analyze the similarity of how much the acquired results agree on whether there is a statistically significant difference between the image pairs. Table II presents the result of this analysis. Each row in the table corresponds to a comparison between the conducted experiments. We use Barnard's test on the pairwise comparison results to determine the statistical significance between pairs [25]. *Agreement* value represents the number of pairs where both experiments agree on the outcome of Barnard's test of statistically significant difference. *Disagreement* value represents the number of pairs where only one of the experiments indicate a statistically significant difference while it is insignificant for the other. *Contradiction* value corresponds to pairs where the sign of the statistical difference is reversed i.e., one of the experiments indicates that for pair A-B, A is significantly better than B whereas the other experiment suggests preference of B over A. Similar to the MPD values, we observe a higher agreement between *Exp-Lab* and *Prolific* experiments compared to *Exp-Lab* and *Exp-Online* experiments in terms of statistically significant difference among the pairs.

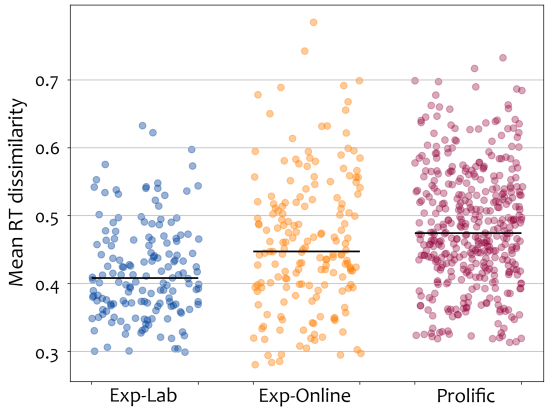

Fig. 5. Mean observer dissimilarity distributions. Each sample represents the mean RT dissimilarity of an observer to rest of the observers in the corresponding experiment

B. Observer Agreement

Due to subjectivity of the aesthetic preferences, it is expected to have a high variance between the observer preferences in TMO evaluation. Therefore, observer agreement does not necessarily correlate with the observer reliability. Nevertheless, it provides valuable insight regarding the effect of experimental conditions on the observer agreement.

Since pairwise preferences are acquired in a binary form, *i.e., Image A is better/worse than Image B*, traditional correlation analysis fail to capture the agreement among observers. In order to measure the agreement of each observer with the rest of the peer population in the experiment, we utilized the Rogers-Tanomoto (RT) distance [26]. RT metric measures the dissimilarity between two binary vectors. It is robust to sample size differences and can use a weight vector to prioritize each observation. Fig. 5 shows the distribution of the mean observer RT dissimilarities within their corresponding experiments. Black horizontal lines indicate the median observer dissimilarity for each experiment. RT values are bound between 0 and 1, and lower values indicate a higher agreement. Cumulative preferences of all observers are used as a weight in calculation. Therefore, distance calculation penalizes the dissimilarities on pairs with higher statistical difference. As shown, observers in the *Exp-Lab* experiment have a higher agreement among themselves as compared to the online experiments. Additionally, we observe a higher agreement among observers in *Exp-Online* experiment compared to *Prolific*.

TABLE III
KRIPPENDORFF ALPHA INTER-OBSERVER AGREEMENT

	All pairs				Signf. Diff. Pairs			
	Plist-1	Plist-2	Plist-3	Plist-4	Plist-1	Plist-2	Plist-3	Plist-4
Exp-Lab	0.2244	0.2512	0.2020	0.3229	0.2856	0.3274	0.3214	0.3579
Exp-Online	0.1653	0.2420	0.1571	0.2496	0.2328	0.3157	0.2187	0.4420
Prolific	0.1576	0.1904	0.1424	0.2224	0.1871	0.2602	0.1958	0.3048

Table III presents the results of Krippendorff alpha evaluation on each experiment. Krippendorff's alpha value measures the inter-observer agreement in an experiment [27]. Higher alpha values indicate higher agreement among the observers. Since our online experiments are split into 4 playlists, we sampled each observer's data accordingly for the *Exp-Lab* experiment. The first 4 columns represent the alpha values on all pairs in corresponding playlists. Additionally, last 4 columns represent the alpha values for the pairs with statistically significant differences. Similar to the RT dissimilarity analysis results, we observe higher agreement among observers in *Exp-Lab* experiment compared to online experiments. For all the experiments, observers show higher agreement on pairs with statistically significant differences.

C. Effect of Number of Observers

A standard way to observe how the number of observers affects the experiment results is by bootstrapping. We create subsets of observers with incremental size from a randomly shuffled list of all observers and evaluate the experiment results over all created subsets. When repeated for a significant number of iterations, we can determine the required number of observers for a desired level of certainty. We start by shuffling our total list and select 10 observers and their preference data. We use Barnard's test to check whether a statistically significant difference exists between the pairs with these 10 observers. The next observer's preferences are then combined with the existing preferences and Barnard's test is conducted again. This step of increasing the subset size is repeated until the maximum number of observers is reached. This procedure is repeated 1000 times separately for each experiment. As a result of this bootstrapping, we can observe for each stimuli pair for 1000 instances, when the observer preference converges to a conclusion with significant difference. Difficult pair of images may result in delayed convergence beyond the maximum number of observers available resulting in incorrect inference. To increase the reliability of the results within the maximum number of observers, we check the consistency of the convergence for each pair, i.e, non-fluctuating Barnard's test results for the observers between $N-5$ and N, where N is the maximum number of observers. In order to compute an expected baseline, independent of the observer order, we also calculate the Barnard's test results for maximum number of observers over 100 permutations. Finally, the distribution of results acquired through bootstrapping over 1000 iterations are compared with the expected baseline results calculated over 100 permutations to find a measure of certainty. Fig. 6 presents the outcome of the permutation test. Y axis represents the certainty in percentage where 100% indicates a perfect match of the convergence distribution between the two permutation test of 1000 and 100 iterations for all pairs in the evaluation. Each color represents an experiment. Solid lines represent certainty percentages over all pairs whereas dashed lines represent certainty percentages only for the common pairs with statistically significant difference. As visualised, to reach the same certainty of the *Exp-Lab* experiment with 35 observers, *Exp-Online* requires 40 observers, and *Prolific* requires 50 observers. Similarly. for the common pairs with statistically significant difference among all experiments, to reach the same certainty of the *Exp-Lab* experiment with 35 observers, *Exp-Online* requires 25 observers, *Prolific* requires 60 observers.

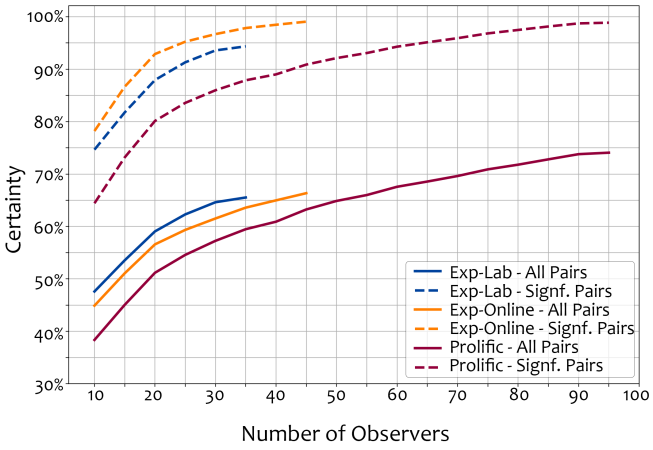

Fig. 6. Effect of number of observers on the certainty of the acquired results over 100×1000 permutations. Y axis is the percentage of pairs which reach to the final conclusion with corresponding number of observers at X axis.

V. CONCLUSION AND DISCUSSION

In this study, we conducted three different experiments with systematic changes to investigate the possibility of using crowdsourcing platforms for aesthetic evaluation of TMOs. First, we collected subjective data in a controlled laboratory environment to acquire expected desired pairwise preferences. The second experiment was conducted online via a private call to the same recruitment channel to isolate the effect of uncontrolled experiment conditions. Finally, we conducted the same experiment on Prolific with the participants pool available on the website.

Comparing the three experiments revealed that the online experiments provide desirable similarity in terms of subjective preferences. Furthermore, effect of Prolific participants pool on the cumulative pairwise preferences is favorable and brings a degree of certainty after reaching certain number of observations per stimuli. We see a higher variation among observers' subjective preferences in *Exp-Online* and *Prolific*. This is not a surprising outcome considering the uncontrolled environmental conditions of the experiments. Finally, we compared the certainty of the collected subjective preferences with varying number of observers. To reach the desired level of certainty, *Prolific* requires higher number of observers overall

when compared to other experiments. Considering the lower cost of recruitment through Prolific and the availability of a wider audience, we find Prolific advantageous in terms of certainty acquired per resource spent.

Hence, through extensive analysis we confirm that Prolific can be safely used to collect subjective preferences on aesthetic evaluation of TMOs. We believe that this conclusion can be generalized to other aesthetic image quality evaluation tasks which do not depend highly on viewing conditions. Finally, we also observe that, depending on the expected certainty compared to the in-lab experiment, the required number of observers to evaluate each pair of stimuli lies between 50 to 60 for a full pair comparison design.

REFERENCES

[1] L. Krasula, K. Fliegel, P. Le Callet, and M. Klima, "Objective evaluation of naturalness, contrast, and colorfulness of tone-mapped images," *Proceedings of SPIE - The International Society for Optical Engineering*, vol. 9217, 08 2014.

[2] H. Yeganeh and Z. Wang, "Objective Quality Assessment of Tone-Mapped Images," *IEEE Transactions on Image Processing*, vol. 22, no. 2, pp. 657–667, 2013.

[3] K. Wolski, D. Giunchi, N. Ye, P. Didyk, K. Myszkowski, R. Mantiuk, H.-P. Seidel, A. Steed, and R. K. Mantiuk, "Dataset and Metrics for Predicting Local Visible Differences," *ACM Trans. Graph.*, vol. 37, Nov. 2018.

[4] X. Cerdá-Company, C. A. Párraga, and X. Otazu, "Which tone-mapping operator is the best? A comparative study of perceptual quality," *CoRR*, vol. abs/1601.04450, 2016.

[5] M. H. Kim, J. Kautz, *et al.*, "Consistent tone reproduction," in *Proceedings of the Tenth IASTED International Conference on Computer Graphics and Imaging*, pp. 152–159, ACTA Press Anaheim, 2008.

[6] G. Krawczyk, K. Myszkowski, and H.-P. Seidel, "Lightness perception in tone reproduction for high dynamic range images," in *Computer Graphics Forum*, vol. 24, pp. 635–646, Citeseer, 2005.

[7] E. Reinhard, M. Stark, P. Shirley, and J. Ferwerda, "Photographic tone reproduction for digital images," in *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, pp. 267–276, 2002.

[8] A. Goswami, M. Petrovich, W. Hauser, and F. Dufaux, "Tone Mapping Operators: Progressing Towards Semantic-awareness," in *2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pp. 1–6, IEEE, 2020.

[9] L. Krasula, M. Narwaria, K. Fliegel, and P. Le Callet, "Influence of HDR reference on observers preference in tone-mapped images evaluation," in *2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX)*, pp. 1–6, 2015.

[10] "Prolific." https://www.prolific.co/. Accessed: Oct 2020. [Online].

[11] "Amazon Mechanical Turk." https://www.mturk.com. Accessed: Oct 2020. [Online].

[12] "Microworkers." https://microworkers.com/. Accessed: Oct 2020. [Online].

[13] D. Kundu, D. Ghadiyaram, A. C. Bovik, and B. L. Evans, "Large-Scale Crowdsourced Study for Tone-Mapped HDR Pictures," *IEEE Transactions on Image Processing*, vol. 26, no. 10, pp. 4725–4740, 2017.

[14] A. Ak, A. Goswami, W. Hauser, P. Le Callet, and F. Dufaux, "A Comprehensive Analysis of Crowdsourcing for Subjective Evaluation of Tone Mapping Operators," in *Image Quality and System Performance, IS&T International Symposium on Electronic Imaging (EI 2021)*, 2021.

[15] Z. Ying, H. Niu, P. Gupta, D. Mahajan, D. Ghadiyaram, and A. Bovik, "From Patches to Pictures (PaQ-2-PiQ): Mapping the Perceptual Space of Picture Quality," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3572–3582, 2020.

[16] T. Hoßfeld, M. Hirth, J. Redi, F. Mazza, P. Korshunov, B. Naderi, M. Seufert, B. Gardlo, S. Egger, and C. Keimel, "Best Practices and Recommendations for Crowdsourced QoE—Lessons learned from the Qualinet Task Force "Crowdsourcing","

[17] P.-Y. Hsueh, P. Melville, and V. Sindhwani, "Data quality from crowdsourcing: a study of annotation selection criteria," in *Proceedings of the NAACL HLT 2009 workshop on active learning for natural language processing*, pp. 27–35, 2009.

[18] U. Gadiraju, S. Möller, M. Nöllenburg, D. Saupe, S. Egger-Lampl, D. Archambault, and B. Fisher, "Crowdsourcing versus the laboratory: towards human-centered experiments using the crowd," in *Evaluation in the Crowd. Crowdsourcing and human-centered experiments*, pp. 6–26, Springer, 2017.

[19] L. Krasula, M. Narwaria, K. Fliegel, and P. Le Callet, "Influence of HDR reference on observers preference in tone-mapped images evaluation," *2015 7th International Workshop on Quality of Multimedia Experience, QoMEX 2015*, 07 2015.

[20] J. Li, R. K. Mantiuk, J. Wang, S. Ling, and P. L. Callet, "Hybrid-MST: A Hybrid Active Sampling Strategy for Pairwise Preference Aggregation," 2018.

[21] M. D. Fairchild, "The hdr photographic survey," in *Color and imaging conference*, vol. 2007, pp. 233–238, Society for Imaging Science and Technology, 2007.

[22] F. Banterle, A. Artusi, K. Debattista, and A. Chalmers, *Advanced High Dynamic Range Imaging (2nd Edition)*. Natick, MA, USA: AK Peters (CRC Press), July 2017.

[23] ITU-T, "Methodologies for the subjective assessment of the quality of television images." ITU-R Recommendation BT.500-14, ITU-R Std., Oct 2019.

[24] B. Gardlo, S. Egger, M. Seufert, and R. Schatz, "Crowdsourcing 2.0: Enhancing execution speed and reliability of web-based QoE testing," in *2014 IEEE International Conference on Communications (ICC)*, pp. 1070–1075, 2014.

[25] G. A. Barnard, "A new test for $2\times$ 2 tables," *Natur*, vol. 156, no. 3954, p. 177, 1945.

[26] D. J. Rogers and T. T. Tanimoto, "A Computer Program for Classifying Plants," *Science*, vol. 132, no. 3434, pp. 1115–1118, 1960.

[27] K. Krippendorff, "Estimating the Reliability, Systematic Error and Random Error of Interval Data," *Educational and Psychological Measurement*, vol. 30, no. 1, pp. 61–70, 1970.