A Comprehensive Analysis of Crowdsourcing for Subjective Evaluation of Tone Mapping Operators

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

A Comprehensive Analysis of Crowdsourcing for Subjective Evaluation of Tone Mapping Operators

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

Abstract

Tone mapping operators (TMO) are pivotal in rendering High Dynamic Range (HDR) content on limited dynamic range media. Analysing the quality of tone mapped images depends on several objective factors and a combination of several subjective factors like aesthetics, fidelity etc. Objective Image quality assessment (IQA) metrics are often used to evaluate TMO quality but they do not always reflect the ground truth. A robust alternative to objective IQA metrics is subjective quality assessment. Although, subjective experiments provide accurate results, they can be time-consuming and expensive to conduct. Over the last decade, crowdsourcing experiments have become more popular for collecting large amount of data within a shorter period of time for a lesser cost. Although they provide more data requiring less resources, lack of controlled environment for the experiment results in noisy data. In this work[1], we propose a comprehensive analysis of crowdsourcing experiments with two different groups of participants. Our contributions include a comparative study and a collection of methods to detect unreliable participants in crowdsourcing experiments in a TMO quality evaluation scenario. These methods can be utilized by the scientific community to increase the reliability of the gathered data.

Figure 1. *Cropping example with all the SRC.*

Introduction

HDR imaging is a pivotal step towards hyper-realism and immersive media consumption. It allows capturing and rendering more details in a scene compared to traditional capture and display devices. Since normal displays cannot render HDR content to its capacity and HDR displays are not yet mainstream, TMOs are used to compress the dynamic range and render captured HDR content on to traditional displays. TMOs map the tonal values to preserve the general perception and visual cues. In the process they can introduces artifacts which reduce the aesthetic quality of the content. Hence, in order to optimize TMOs according to human preferences, quality evaluation is crucial.

Assessing the quality of tone mapped HDR images is a non-trivial task considering the subjective nature of the problem. Tone mapping quality can be assessed by using objective quality metrics such as TMQI [1] or through a subjective experiment involving human participants. Although both are beneficial for certain tasks, subjective experiments with enough number of participants provide more accurate evaluation.

Subjective experiments for TMO quality evaluation can be conducted with or without a reference. Reference HDR stimulus can be presented on an HDR screen side by side with the tone-mapped stimulus. This provides participants a reference point when assessing the quality of tone-mapped images. Alternatively, evaluation of tone-mapped image can be conducted without providing a reference point to human subjects. Although, assessing the quality of the same stimuli, each experimental design provides answer to different questions in TMO quality evaluation scenario [2]. While, having a reference image can help to assess fidelity of the TMO, no-reference scenario can answer the preference of an observer among different TMOs [2].

In the last decade, crowdsourcing platforms, such as Amazon Mechanical Turk (AMT) [3] and Microworkers [4] have gained popularity among the scientific community to conduct subjective experiments. While these platforms help to collect data from a larger population, researchers have less control on the environment in which the experiment is conducted. Furthermore, compared to an in-lab experiment, crowdsourcing experiments contain higher number of unreliable participants.

Identifying unreliable observers is not straightforward for aesthetic focused subjective experiments. Primary reason is aesthetic comparisons don't have a ground truth unlike subjective experiments measuring image fidelity or similarity. In other words, each observer can respond differently to same stimuli. Thus, reliability of observers cannot be evaluated by their aesthetic choices. In this work, we identify different constraints for subjective evaluation via crowdsourcing. We focus on methods which can be used

[1]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)

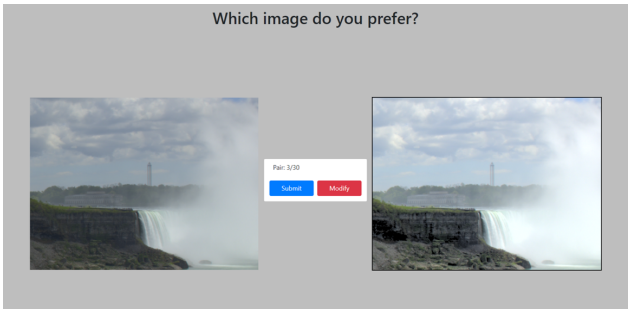

Figure 2. *Example test screen*

in aesthetic evaluation of images to process the retrieved data in order to extend the reliability of evaluation.

Related Works

We start by summarizing existing research on TMOs and their subjective evaluation. Additionally, we delve into datasets targeted towards IQA and existing efforts towards detecting spurious data and unreliable observers in regards to such dataset compilation.

Over the past decades, TMOs have been widely researched. Functionally, TMOs are of two types: *global* and *local*. Global TMOs apply the same luminance compensation throughout the image, whereas local TMOs take into account the spatial neighborhood of each pixel. Reinhard et al. [5] introduce a TMO inspired by the photographic *Zone system* to globally scale exposure and *dodge and burn* to compress the dynamic range of the image.

We follow a recent comparative subjective study of several classical TMOs provided by Cerda-Company et al. [6] to further select two highly rated TMOs. Kim et al. [7] propose a global TMO based on the log-luminance adaptation of human visual cortex. As a local approach, Krawczyk et al. [8] proposed a TMO based on a probabilistic model of lightness perception. They decompose an HDR image into areas of consistent luminance (lightness framework) and map each framework by adjusting the perceived 'white' point.

Although, objective image quality metrics can be utilized to assess the quality of an image, subjective experiments are the most robust and reliable method of analysing aesthetic preference of images. For the evaluation of tone mapped images, several subjective studies exist in the literature[2][9]. Evaluation of the tone mapped content can be done with or without a reference, ie. the presence of the original stimuli. This design choice depends on the use-case of the study. A fidelity evaluation would benefit from the presence of a reference whereas for an aesthetic preference evaluation a no-reference study is prefered [2].

Previous literature studies in subjective evaluation of tone mapped content have mostly been conducted in a controlled lab environment with physically attending participants. Although, platforms such as AMT[3] and Microworkers[4] gained popularity in the last decade, they have not been utilized much for subjective evaluation of tone mapping operators. To best to our knowledge, we are only aware of one study by Kundu et al.[10]. Although the dataset contains 1811 images, it is not a tone mapping operator evaluation study. Additionally, authors used several methods to deal with unreliable observers such as gold standard

	KK	Kraw	Rein	STMO
KK	0	33	16	15
Kraw	6	0	5	7
Rein	23	34	0	17
STMO	24	32	22	0

Figure 3. *Tone mapped stimuli and pair comparison matrix*

images where results gathered from an in-lab experiment have been compared to participants' answers. Subsequently, participants who make more mistakes than previously set threshold are considered to be unreliable.

Experimental Design

As discussed earlier, both full-reference and no-reference methodologies can be utilized for tone mapping evaluation. In a crowd-sourcing platform, conducting a full reference experiment for tone-mapping operator evaluation is practically difficult as it requires an HDR screen for each participant. Hence we follow a no-reference methodology for our experiment.

Pair comparison (PC), rating and ranking are three major methods which are used for subjective quality evaluation. Previous studies [11] have shown that forced-choice pairwise comparison method provides the most accurate results. It follows a simpler task and hence is less demanding for the participants. Therefore we select the PC method for our experiment.

Participants are not allowed to use smartphones or tablets. In addition, display resolution is set to 1080p as it is widely used for commercial purposes. This ensures same presentation of stimuli for every participant. Although display resolution is controlled during the experiment, participants are free to adjust viewing distance. Two tone mapped images of 480p resolution are shown side by side and participants are requested to indicate their preference. Participants are given permission to observe each stimuli as long as needed. At the end of their evaluation, participants provide their choice by clicking on their preference and confirming it.

We use the publicly available Fairchild's [12] HDR dataset as source content (SRC) to generate tone mapped images. The image resolution in Fairchild database is fairly large, therefore, we have scaled down and created systematic crops of 480p resolution. Selected crops filtered by dynamic range and standard deviation in order to promote challenging scenes for tone mapping. Filtered crops are then clustered based on TMQI [1] scores. Finally, we select 20 SRC crops among the clusters. 4 tone mapping algorithms, ReinhardTMO [5], KrawczykTMO [8], KimKautzTMO [7] and SemTMO [13] are selected from the literature. Each SRC is tone mapped with aforementioned TMOs. Apart from SemTMO, we use the matlab HDR Toolbox by Banterle et al. [14] for the other TMOs. The parameters of each TMO are optimised to maximise their respective TMQI [1] scores.

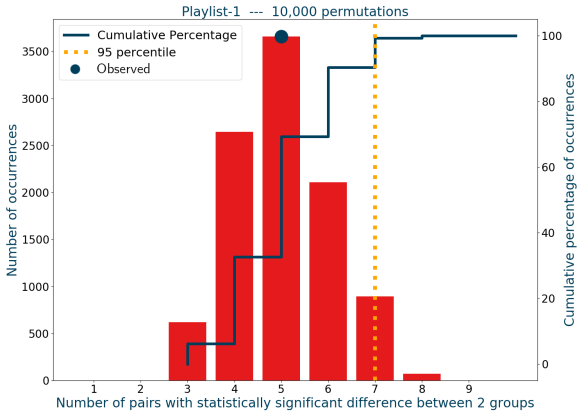

Figure 4. *Permutation test results between the two experiment for Playlist-1*

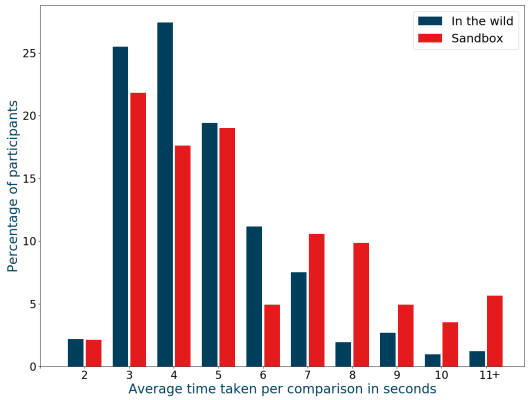

Figure 5. *Distribution of participants based on average time taken per comparison for each experiments. A bar at n^{th} second on the plot corresponds to $(n-1,n]$ seconds range.*

Finally, we compiled a dataset containing 80 tone mapped images and 120 unique pair-wise comparisons. We split the dataset into 4 playlists of 30 pairwise comparisons each to keep the experiment duration short and traceable.

Two separate experiments are conducted using the aforementioned dataset. Both experiments are conducted online via AMT platform. Although all design choices remain exactly identical, the method of recruitment of participants differs between the two experiments. For the first experiment, for each playlist, 100 participants are recruited from AMT workers, 400 in total. We call this experiment 'In the wild'. Minimum requirement of previous Human Intelligence Tasks (HIT) conducted by AMT workers is set to 500 with a minimum approval rate of 0.98 (range of 0 to 1). For the second experiment, AMT Sandbox is used to conduct the experiments and around 40 participants are recruited for each playlist through connections from previously conducted experiments. We will call this experiment 'Sandbox'. We hypothesise that the 'In the Wild' experiment may have unreliable observers. As a result, we may expect behavioral differences between the two group of participants. We will first investigate the differences between the two experiment, in terms of preferences. We will then provide different analysis in order to identify the participants with unexpected behavior which leads to suspicion.

Analysis of Pairwise Comparison Results

Since content and design of both experiments are exactly the same, with a large enough number of participants, one would expect similar preferences from both groups. In this section, we compare and analyse the results of both experiments. Since PC is used for the experiments, preferences are stored in the form of pair comparison contingency matrices (PCM). An example PCM can be found in the Fig 3. Each cell of the table represents the number of times observers have chosen the TMO at the row in comparison to TMO at the column.

Barnard's Exact Test

One common way to determine whether there is a statistically significant difference between a pair of images over pair-comparison data is to check Barnard's test [15] results. Barnard's test is an alternative to Fisher's test[16] to determine statistically significant differences for $n \times n$ contingency tables. For large n, effect of discreteness of Fisher's statistical test reduces and renders Fisher's test more powerful compared to Barnard's test. However, for 2×2 contingency tables, Barnard's test has been observed to be more powerful than Fisher's test [17].

Comparing the results from a significantly different pairs perspective, we observe a difference between both experiments. Among 120 image pairs in total, both experiments are in agreement for 76 of the pairs.

Permutation Test

Barnard's test can also be used to determine whether both distributions are statistically different for a given image pair. While directly comparing the two experiments for each of the image pairs is useful, a permutation test is necessary to ensure the validity of this information. For this purpose, we conducted a 10k fold permutation test. With each iteration, two random observers, one from each experiment, are swapped. The new count of statistically different pairs is calculated. This procedure is applied to each playlist separately.

Result of the permutation test indicates that the Barnard's test results from comparing both experiments are within the 95 percentile range. In other words, we can claim with a confidence of 95% that the observed difference between the two experiment is not random. Fig. 4 illustrates the permutation test results for playlist-1. Histogram shows the distribution of statistically significant pairs between both experiments for 10000 permutations. The dark solid line represents the cumulative percentage value on the right vertical axis, while the dashed yellow line shows the 95 percentile threshold. Observed difference (5 pairs out of 30 possible) is represented by the dark colored dot on the histogram. It is clear that the observed difference is below the 95 percentile threshold. For each playlist, permutation test has been conducted and the results indicate the same.

Detection of Unreliable Observers

As discussed and documented in the previous section, we observe statistically significant differences between the result of

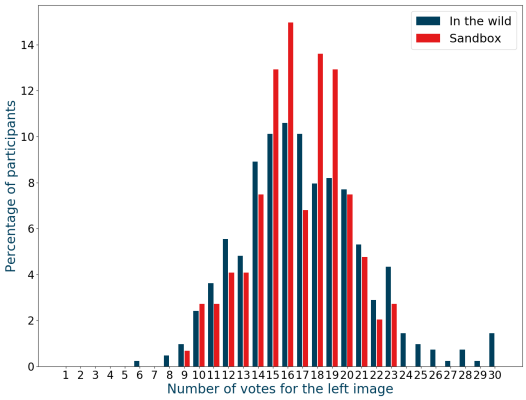

Figure 6. *Distribution of participants based on the number of times the left image is chosen over the right one.*

both experiments. Although these differences can be partly attributed to some uncontrolled factors of the experimental environment, we believe that unreliability of observers within the AMT worker population is one of the most important reasons. To test our hypothesis, we propose a set of methods to identify unreliable observers. By comparing participants from 'In the wild' experiment and 'Sandbox' experiment, we can distinguish such behaviours.

Timing Analysis

We have not set time limits for either experiments. Participants have been allowed to respond as quickly or as slowly as they wish. On comparison, we observed that participants from 'In the wild' experiment have been faster on average. This might be considered as an indication of the difference of attention span between both groups during the experiment. Histogram of the average time spent per comparison by participants is plotted in Fig. 5. The timing alone is not a robust condition to detect an unreliable observer. However, it can be used as an indicator to flag and further analyse observers who spend significantly less time on a comparison since it may be a sign of carelessness. With this in mind, we set an empirical threshold of 2 seconds per comparison and identify 10 participants who spend less than that on average and marked those data as suspicious.

Voting Pattern Analysis

In this section, we evaluate participants' behaviour in terms of their voting patterns. We aim at identifying whether if a participant follows a certain discernible pattern while voting, such as voting left/right image consecutively, alternating left-right votes perfectly, etc. In addition to voting patterns, we also analyse the distribution of left/right votes for each participants. In Fig. 6, distribution of the participants by the number of times they have voted for the left image is displayed, for a playlist of 30 images. We observe normal distribution for both experiments. For 'In the wild' experiment, some observers show suspicious behaviours such as voting for the image on the left side for more than 24 times. With a random sampling of paired test images, it is unlikely to happen. In addition, the comparison with the 'sandbox' experiment confirms it. As a result, we have identified 15 participants from 'In the wild' experiment as suspicious.

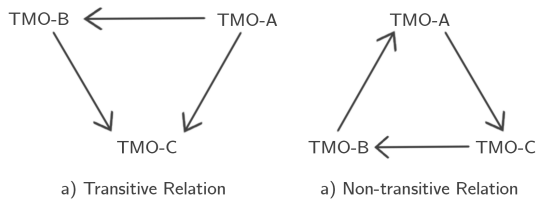

Figure 7. *Example cases of a transitive relation and a non-transitive one. Arrow directions indicate observer preferences for a tone mapped image over another.*

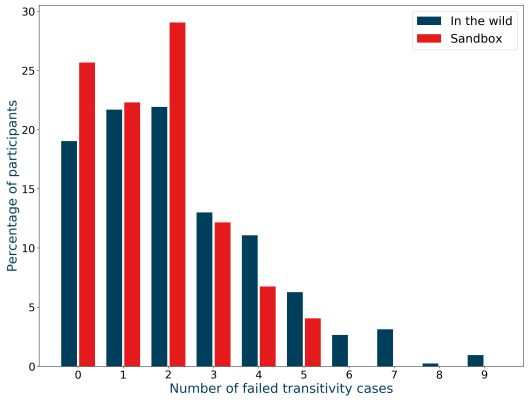

Figure 8. *Distribution of the participants for each experiment based on the number of failed transitivity relation out of possible 20.*

Transitivity Analysis

In mathematics, a relation ">" is transitive over a set M if and only if $\forall x,y,z \in M$, $x > y$ and $y > z$ implies the condition $x > z$. In the context of our experiments, we check whether a transitivity exists between comparison of each tone mapped image for each participant. Two example cases are shown in Fig. 7. Arrow directions indicate an observer's preference of a tone mapped image over another tone mapped image. A behaviour is marked as suspicious if the relation is non-transitive.

It is important to note that preferences that do not satisfy a transitive relation do not imply a dishonest participant, due to the subjective nature of the image comparison task. However, since the conditions for both experiments are exactly identical, we expect to observe similar behaviours from both participant groups. With this in mind, we have counted the number of occurrences of non-transitive relations for each participant in both experiments and compared them. As it can be observed in Fig. 8, 'In the wild' experiment has more participants with higher number of failed transitivity relations compared to 'Sandbox' experiment. Nearly 10 percent of the participants in the 'In the wild' experiment have at least 6 failed transitivity relations out of 2o possibilities. Conversely there is no participant in the 'Sandbox' experiment with more than 5 failed transitivity relations. Thus, 29 participants from 'In the wild' experiment are marked as suspicious.

Discussion and Conclusion

After conducting the aforementioned analysis, we have identified 54 participants in total as suspicious. Among 400 participants, 10 are identified by timing analysis, 15 are filtered further by voting pattern analysis and finally 29 are identified by transitivity analysis. After pruning the PCMs from unreliable records, agreement between the two experiment increased according to Barnard's Exact Test results. Among 120 pairs in total, number of pairs where both experiments are in agreement increased from 76 to 82.

In addition to the analyzed methods, one common method to identify unreliable participants is the gold standard image analysis[10]. It requires a set of annotated comparisons with known outcomes. Stimuli from this set of comparisons is then inserted into the crowd sourcing experiment playlists and participants' preference are checked for consistency. It is expected to be in line with the known outcomes. However, it is particularly difficult to find gold standard images for aesthetic evaluation. Participants do not always share the same opinions. Evaluating participants by aesthetic preferences may lead to incorrect unreliability detection. Another method that can be adapted is to repeat image pairs[10]. Randomly selected stimuli are presented more than once to each observer and participants are expected to provide consistent answers. The disadvantage of both of these methods is the requirement of considerable amount of resource for implementation. Crowd sourcing experiments are shorter compared to an in-lab experiment due to lower attention span of the participants. A playlist with 30 stimuli where 5 stimuli are repeated requires to spend 20% more resources.

We have investigated several methods to detect unreliable observers for TMO evaluation in a pair comparison experiment. Identifying unreliable observers is particularly difficult for aesthetic evaluation experiments due to lack of ground truth for aesthetic preferences. Thus, we have focused on methods that does not require an expected answer such as gold standard units. Methods proposed in this work can be adapted to other aesthetic preference subjective experiments without additional resources. As future work, we are planning to conduct the same experiment in a controlled lab environment to further investigate crowd-sourced data and the effect of environmental factors on the quality evaluation of TMOs.

References

[1] 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.

[2] 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.

[3] "Amazon mechanical turk." https://www.mturk.com. Accessed: Oct 2020. [Online].

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

[5] 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.

[6] X. Cerda-Company, C. Parraga, and X. Otazu, "Which tone-mapping operator is the best? a comparative study of perceptual quality.," *Journal of the Optical Society of America. A, Optics, image science, and vision*, vol. 35, no. 4, p. 626, 2018.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] R. K. Mantiuk, A. Tomaszewska, and R. Mantiuk, "Comparison of four subjective methods for image quality assessment," in *Computer graphics forum*, vol. 31, pp. 2478–2491, Wiley Online Library, 2012.

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

[13] 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.

[14] 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.

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

[16] R. A. Fisher, "On the interpretation of χ 2 from contingency tables, and the calculation of p," *Journal of the Royal Statistical Society*, vol. 85, no. 1, pp. 87–94, 1922.

[17] C. Mehta and P. Senchaudhuri, "Conditional versus unconditional exact tests for comparing two binomials," 01 2003.

Author Biography

Ali Ak received his B.Sc. in Mathematics from Bilkent University(2014) and he earned a Master of Science degree at Visual Computing in University of Nantes(2018). As from September 2018, he has joined LS2N in University of Nantes within the framework of RealVision ITN. His work is focused on perceptual evaluation of image quality.