On Spammer Detection in Crowdsourcing Pairwise Comparison Tasks: Case Study on Two Multimedia QoE Assessment Scenarios.

Ali Ak, Mona Abid, Matthieu Perreira da Silva, Patrick Le Callet

ON SPAMMER DETECTION IN CROWDSOURCING PAIRWISE COMPARISON TASKS: CASE STUDY ON TWO MULTIMEDIA QOE ASSESSMENT SCENARIOS

Ali Ak[1], Mona Abid[1], Matthieu Perreira Da Silva, Patrick Le Callet

LS2N, CNRS, University of Nantes, France,
name.surname@univ-nantes.fr

ABSTRACT

The last decade has brought a surge in crowdsourcing platforms' popularity for the subjective quality evaluation of multimedia content. The lower need for intervention during the experiment and more expansive participant pools of crowdsourcing platforms encourage researchers to join this trend. However, the unreliability of the participant behaviors puts a barrier in the wide adoption of these platforms. Although many works exist to detect unreliable observers in rating experiments, there is still a lack of methodology for detecting unreliable observers in quality evaluation of multimedia content using pairwise comparison. In this work, we propose methods to identify irregular annotator behaviors in pairwise comparison paradigm. We compare the proposed methods' efficiency for two scenarios: quality evaluation of traditional 2D images and 3D interactive multimedia. We conducted two crowdsourcing experiments for two different Quality of Experience assessment tasks and inserted carefully designed synthetic spammer profiles to evaluate the proposed tools. Our results suggest that the detection of unreliable observers is highly task-dependent. The influence of the spammer behavior intensity and the proportion of spammers among the observers can be more severe on tasks with higher subjectivity. Based on these findings, we provide guidelines and recommendations towards developing spammer detection algorithms for subjective pairwise quality evaluation of multimedia content.

Index Terms— Spammer detection, Pairwise comparison, Crowdsourcing.

1. INTRODUCTION

Machine learning is used widely to develop quality evaluation models in the multimedia domain. These techniques rely on a large amount of data to reduce bias towards the training data. Traditionally, subjective preferences are collected via experiments in a controlled lab environment. With the surge of data-driven approaches, recent years brought increasing popularity to crowdsourcing platforms, such as Prolific[1], as an alternative method of collecting subjective preferences. However, compared to subjective experiments conducted in controlled laboratory environments, crowdsourced subjective experiments may contain a larger amount of unreliable observers due to uncontrolled experimental conditions. For this reason, the need for methodologies to analyze and improve the quality of collected subjective preferences is of rising importance. It is especially crucial for developing better models which rely heavily on training data.

Subjective Quality of Experience (QoE) assessment methodologies can be split into two categories: rating and ranking. Rating methodology has been widely adopted in literature, and well-established outlier removal and spammer detection tools exist in the standards [2]. Rating tasks ask the observers to assign a score to the displayed stimuli while ranking tasks ask the observer to compare and rank the two (or more) displayed stimuli. The most common adoption of the ranking methodology is Pairwise Comparison (PC), where observers see two stimuli at once and rank according to the research question. PC is argued as more consistent and reliable than alternative methodologies in QoE tasks due to providing a simplified task for the annotators [3]. By simplifying the annotators' evaluation task, PC methodology reduces the subject uncertainty and provides more reliable subjective preferences compared to rating methodologies. Thus, it is more suitable to crowdsourcing experiments where annotators' attention span is lower [4].

Currently, spammer detection in the QoE domain has been investigated thoroughly for rating tasks. Although spammer detection in PC paradigm has been explored thoroughly in domains such as online social networks, product reviews [5, 6], there is still a lack of well-established spammer detection methodology for PC experiments in the QoE domain. Thus, spammer detection in QoE PC experiments is an attractive topic to investigate due to the increasing need for reliable subjective preference data, increasing crowdsourcing platform usage, and lack of methodologies in the literature.

This study conducts a preliminary analysis of statistical methods that evaluate observer similarity in order to detect irregular behaviors in two different QoE assessment tasks.

[1] Ali Ak and Mona Abid contributed equally to this work as first authors.

This work was funded by the French National Research Agency as part of ANR-PISCo project (ANR-17-CE33-0005) and by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie Grant Agreement No. 765911 (RealVision).

978-1-6654-3864-3/21/$31.00 ©2021 IEEE

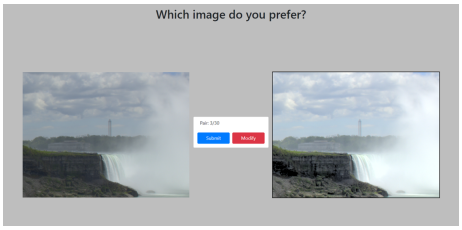

Fig. 1. Example test screen for Exp-TMO

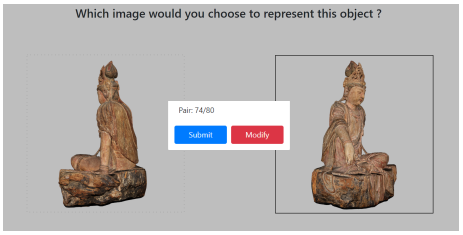

Fig. 2. Example test screen for Exp-VP

We conducted two experiments to collect subjective pairwise preferences, firstly for aesthetic quality assessment of 2D HDR images processed with different tone-mapping operators, secondly for viewpoint preferences of 3D interactive multimedia content. First, we investigate the effect of task differences on the collected subjective preferences, and we evaluate observers' agreements for each experiment using adequate statistical tools. Moreover, we evaluate the effectiveness of the considered methods by generating synthetic spammer profiles and inserting them among reliable observers while systematically increasing the spammer behavior intensity and spammer proportion.

2. SUBJECTIVE EXPERIMENT DESIGNS

We conduct two online subjective QoE assessment experiments on Prolific crowdsourcing platform to collect subjective pairwise preferences. Selected tasks have differences in subjectivity, research question, and stimuli while sharing the same experimental methodology. The first experiment, *Exp-TMO*, is conducted with traditional 2D images on a highly subjective task, *i.e., aesthetic quality assessment of tone mapped HDR images*. The second experiment, *Exp-VP*, is conducted on rendered views of 3D objects to select the most representative view of each object, *i.e., 3D viewpoint subjective preference*.

Conducted experiments have the following fundamental differences: the subjectivity of the questions directed to observers, source content being used, the purpose of the collected data. Observers were recruited through the participant pool available on Prolific crowdsourcing platform with similar requirements. In each experiment, we followed a pairwise comparison methodology. Sample screenshots from Exp-TMO and Exp-VP is presented in Fig. 1 and Fig. 2.

2.1. Dataset - Stimuli Generation

Exp-TMO [1]**:** To collect pairwise preferences of tone mapped 2D images, we used 20 HDR source contents and four different tone mapping operators (TMO). We selected 20 HDR images from Fairchild's HDR dataset and extracted crops with spatial resolution of $640 \times 480px$ [7] and converted into 8-bit representations with four TMOs from the literature: ReinhardTMO [8], KrawczykTMO [9], KimKautzTMO [10] and SemanticTMO [11]. Considering the content dependency of the aesthetic evaluation of TMO, we choose not to evaluate tonemapped images of different SRCs. Therefore, without any cross-content evaluation and with 20 SRCs and 4 HRCs, we generate 80 tonemapped images providing 120 pairs to compare.

Exp-VP [2] We selected 21 high-resolution triangle meshes with color information represented by both vertex colors and texture mapping. These models were chosen so as to ensure a variety of shapes and colors. They belong to 4 different semantic categories: human, art, animals, objects. All rendered views fit in a squared shape window with the resolution of $600 \times 600px$. The viewpoint preference corresponds to the most representative viewpoint from which the 3D object is rendered; therefore, cross-content comparisons are not meaningful in this context.

To sum up, twenty-one 3D source models are rendered under four viewpoints, resulting in a dataset of 84 rendered images and 126 unique pairs without cross-content consideration.

2.2. Experiment Setup & Protocol

Both experiments are designed to display the image content side by side on the display. The stimuli are visualized in a neutral background. No time limit is set for the experiment. Although we had no information about crowdsourcers' viewing conditions, we use restrictions on the display resolution, minimum resolution $800 \times 1300px$, to provide similar representations to observers. We also oblige the observers to use the full-screen mode to proceed with the experiment to reduce distraction during the experiment.

To decrease the noise that can occur due to crowdsourcers' lower attention span, we split the dataset into smaller parts to keep the experiment duration short. 100 unique observers have evaluated each playlist for each experiment. Although there are no demographic limitations for participants, we request a minimum 90% approval rate to increase recruited participants' reliability.

3. ANALYSIS OF TASK SUBJECTIVITY

In this section, we compare the pairwise preference results acquired from two subjective experiments to identify the influence of task differences. Firstly, we compare the number of pairs with statistically significant differences to confirm the hypothesized subjectivity difference among the two experi-

[1]Exp-TMO: ftp://ftp.polytech.univ-nantes.fr/TMOEval_Prolific

[2]Exp-VP: ftp://ftp.polytech.univ-nantes.fr/3DViewpointPref_Prolific

Table 1. Number of pairs with and without statistically significant differences.

	Sign.diff pairs	Non sign. diff pairs
Exp-TMO	$91 - 76\%$	$29 - 24\%$
Exp-VP	$121 - 96\%$	$5 - 4\%$

ments. Additionally, we calculate the inter-observer agreement of each experiment and compare the results to reveal the influence of task subjectivity on the observer preferences.

3.1. Computation of Pairwise Preferences

The outcome of the Pair Comparison experiment is a pair comparison matrix, also known as a preference matrix $\mathbf{A}$, where $\mathbf{A} = (a_{ij})_{m \times m}$. a_{ij} is the total count of preference of stimulus S_i over S_j for all observers. $a_{ii} = 0$ for $i = 1, 2, ..., m$. P_{ij} represents the probability that stimulus S_i is preferred over S_j , i.e., $Pij = a_{ij}/n_{ij}$. With n_{ij} the total number of comparisons for stimuli pair $\{S_i S_j\}$, $n_{ij} = a_{ij} + a_{ji}$. After acquiring pair comparison matrices, Barnard's exact test [12] is used to validate the statistical significance of the differences for each pair. Barnard's test results suggest that $p\,value < 0.05$ indicates a statistically significant difference with 95% confidence level for the pair in comparison.

Table 1 summarizes the result of Barnard's exact test results for both experiments. We observe a higher number of pairs with statistically significant differences in Exp-Vp compared to Exp-TMO, confirming the expected subjectivity of the QoE assessment task in Exp-TMO. Considering that 100 unique observers have evaluated each pair, we can assume that the pairs with no statistically significant differences are close in terms of preference. In aesthetic quality evaluation, as expected, we observed a higher number of close pairs.

3.2. Overall Inter-Observer Agreement

We analyzed the inter-observer agreement by calculating the correlation coefficients between the two disjoint halves of observers. When repeated for 100 iterations with randomly selected disjoint halves, the calculated correlation coefficients' median value indicates the level of general agreement among observers. The Median Pearson correlation coefficient (PCC) is used to quantify the agreement. For Exp-TMO, we acquired a PCC value of 0.9289, while PCC of Exp-VP is 0.9896. A higher PCC value of Exp-VP indicates a higher correlation, therefore a higher inter-observer agreement when compared to Exp-TMO.

Additionally, Krippendorff's alpha can be used to calculate the inter-observer agreement in a subjective experiment [13]. Alpha values range between 0 and 1, where higher values indicate a higher agreement among observers. For Exp-TMO and Exp-VP, calculated alpha values are 0.1781 and 0.6158, respectively. Krippendorff's alpha values further confirm the higher inter-observer agreement for the Exp-VP compared to Exp-TMO.

4. INDIVIDUAL OBSERVER AGREEMENT

This section introduces and suggests two different metrics to quantify the agreement between each unique pair of observers. Acquired agreement scores between pairs of observers are used as an indicator of irregular observer behavior. Firstly, we calculate Cohen's Kappa coefficient [14] between all unique pairs of observers for each experiment. Cohen's Kappa coefficient is widely adopted for measuring inter-observer agreement [15]. Secondly, we used a binary similarity metric known as Rogers-Tanimoto distance (or weighted Jaccard distance) which is also widely adopted in literature to assess the similarity between two binary vectors [16]. After analyzing the collected subjective preferences with the aforementioned agreement measures, we eliminate the observers with detected irregular behaviors.

4.1. Cohen's Kappa Coefficient

Kappa coefficient is developed to measure inter-observer agreement while considering that observers sometimes do not know the answer (or do not pay attention) and merely guesses [14]. Kappa values range between -1 and 1, where a kappa value of 1 indicates perfect agreement between a pair of observers, -1 indicates the complete disagreement, while Kappa value of 0 is considered random chance. We computed the mean Kappa coefficient of each observer obs_i compared to every other observer obs_j where $j \in \{1, N\}$ with $j \neq i$. The K_i value associated with obs_i tells us how well there is an agreement between observer i and the rest of the observers in the experiment. The plot on the right in Fig. 3 presents the distribution of observers for each experiment in terms of their Mean Cohen's Kappa coefficient.

For Exp-TMO, Kappa values are in the range $[-0.098, 0.335]$ with an overall mean value and standard deviation 0.183 ± 0.086 whereas for Exp-VP, Kappa values are in the range $[0.095, 0.760]$ with an overall mean value and standard deviation 0.619 ± 0.092. We use the Interquartile Range (IQR) to detect outlier and possible spammers. It is a measure of statistical dispersion and calculated by the difference between the 75^{th} and the 25^{th} percentiles. We identified 4 observers in the Exp-TMO and 7 observers in the Exp-VP.

4.2. Rogers-Tanimoto Dissimilarity

Pairwise preferences of each observer can be represented with a binary form, *i.e., Image A is better/worse than Image B*. It allows to use binary distance metrics for measuring similarity between observer preferences. Rogers-Tanimoto (RT) distance is widely adopted in literature, and it measures the dissimilarity between two binary vectors [17]. It is robust to varying sample sizes, and a weight vector can be used to adjust each pair's effect on the dissimilarity calculation. We used the cumulative preference of observers to generate the weights in RT dissimilarity calculation. For a given image

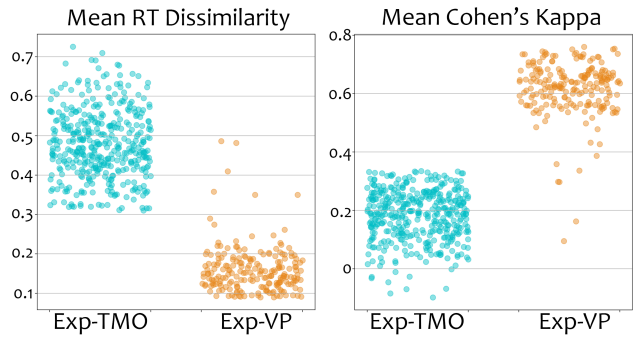

Fig. 3. Mean individual observer agreement based on RT dissimilarity and Cohen's kappa coefficient. Each dot represents an observer in corresponding experiments.

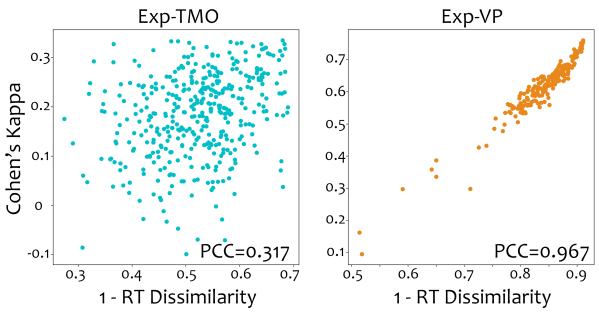

Fig. 4. Comparison of Cohen's kappa values with RT dissimilarities. For visualization purposes, 1-RT values are used.

pair $pair_i$, which compares img_A with img_B, weight of $pair_i$ is calculated as: $w_i = |P_A - P_B|/N$, where N is the number of unique observers who annotate the $pair_i$. P_A and P_B represents the number of observers who prefers img_A and img_B respectively. This allows us to generate a weight for each pair in the range of $\{0, 1\}$ where larger values indicate a stronger influence on RT dissimilarity calculation. RT dissimilarity values range between 0 and 1, and lower values indicate a greater agreement between the two observers.

The plot on the left in Fig. 3 presents the distribution of observers for each experiment in terms of their mean RT dissimilarity with the rest of the observers. For Exp-TMO, dissimilarity values are in range $[0.314, 0.733]$ with mean$\pm$ std:0.483 ± 0.086 whereas for Exp-VP, dissimilarity values are in range $[0.091, 0.491]$ with mean$\pm$ std: 0.161 ± 0.059. Based on the IQR statistical measure, we identified 8 observers in Exp-VP and 2 observers in Exp-TMO, whose associated dissimilarity is outside of IQR range(i.e., between 75^{th} and the 25^{th} percentiles).

4.3. Correlation between Cohen's Kappa Coefficient and RT Dissimilarity

The correlation of the proposed methods has been investigated in terms of pearson correlation coefficient (PCC). Fig. 4 presents the scatter plots of two agreement measures for each experiment. As reported on the figure, PCC value of -0.967 is acquired between Cohen's Kappa Coefficient and RT Dissimilarity values of all observers in Exp-VP while the PCC value for Exp-TMO is much lower in comparison with -0.313. It suggests that the individual observer measurements acquired with both methods are highly correlated when the subjectivity of the task is lower. This can be explained by the weighted calculation of RT dissimilarity measure. RT dissimilarity uses a weighted calculation as described in Sec. 4.2. It penalizes the disagreement between observer preferences for the pairs with higher statistical difference, while minimizes the influence of close pairs.

We observe a significant difference in mean observer agreement value ranges between two experiments for both of the metrics. This observation indicates that a generic threshold for mean agreement values is not sufficient since observer agreements are highly task-dependent. We also observe higher agreement in terms of mean individual observer agreements for the Exp-VP experiment compared to Exp-TMO, further confirming the difference in the two tasks' subjectivity. Finally, before moving on to synthetic spammer detection analyses, we removed the outliers suggested by both measures. In total, we identified 8 observers as outliers in Exp-VP and 5 in Exp-TMO.

5. DETECTING SYNTHETIC SPAMMERS

This analysis aims to see whether previously introduced individual observer agreement tools can identify irregular annotators *i.e., spammers*. We inserted synthetically created spammer annotators among real observers in both Exp-TMO and Exp-VP experiments. We analyze the effect of different QoE tasks on identifying spammers. Furthermore, we analyze the influence of the proportion of spammers and spammer behavior intensity on each task.

5.1. Generating Spammer Profiles

In this section, we introduce different spammer behaviors which can be observed in pairwise comparison experiments. A spammer may behave honestly at the beginning of the experiment and vote irregularly after a certain point. Therefore, we rely on randomly selected real observers to generate spammer preferences. We control the intensity of spammer behaviors in each profile with an adjustable parameter.

Random voter: Annotator randomly votes on image A or on image B. This behavior is generated by randomly sampling binary votes for each stimulus.

Repeater: Annotator has a position bias, thus providing his/her pairwise preferences based on image position *i.e. left/right, top/bottom, etc.*. We simulate this behavior by repeating a random position, *i.e., 0 or 1*

Inverted voter: Annotator provides pairwise preferences on the content that he/she does not prefer. This behavior can be caused by misunderstanding the question. We simulate this behavior by inverting real annotator preferences.

Mixed: Behaviours described above combined randomly.

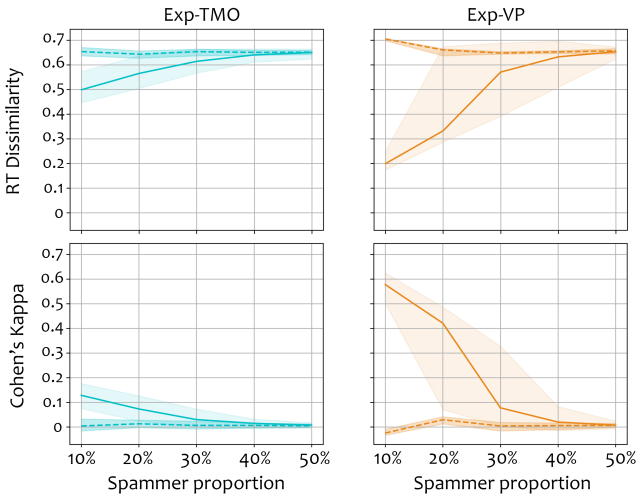

Fig. 5. Mean and 75% percentile range of RT dissimilarity and Cohen's Kappa values of real observers and spammers for varying proportion of spammers. Solid and dashed lines represent the real observers and spammers respectively for each experiment.

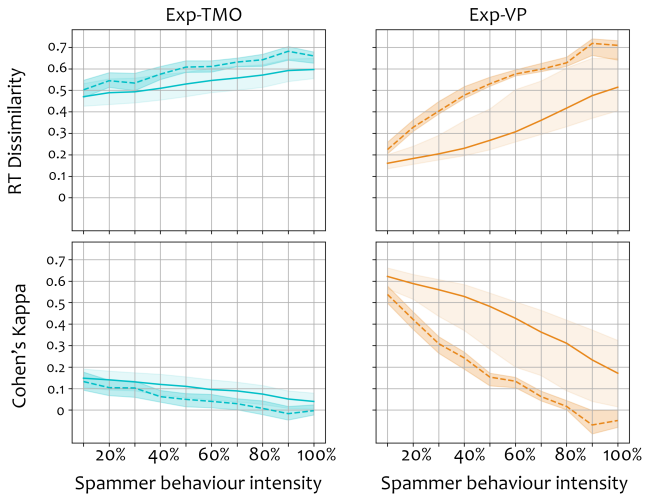

Fig. 6. Mean and 75% percentile range of RT dissimilarity and Cohen's Kappa values of real observers and spammers for varying spammer behaviour intensity. Solid and dashed lines represents the real observers and spammers respectively for each experiment.

Influence of the spammer profiles: We analyzed the influence of individual spammer profiles on the mean observer agreements. With varying spammer proportions and intensities, we calculated the observer agreements of each spammer profile. We observe no significant effect on the observer agreements with different spammer profiles under tested conditions. Nevertheless, to prevent bias towards a particular spammer profile, we included all profiles in the following analyses.

5.2. Synthetic Spammer Detection

In this section, we considered the spammer profiles described previously to generate synthetic spammer votes. Each spammer profile has an adjustable parameter to control the intensity of the spammer behavior.

As previously described, we use RT dissimilarity and Cohen's kappa coefficient to analyze observers' agreement.

Influence of spammers proportion: This analysis aims to measure the influence of the proportion of spammers on the observer agreement measures. We used an 80% fixed intensity level for each spammer's behavior intensity and systematically increased the proportion of spammers inserted into each experiment. We calculated the RT dissimilarity and Cohen's kappa coefficient of individual observers at each increment for each incremental. For each experiment, mean values and 75% percentile range of synthetic spammers and real observers are shown in Fig. 5. The first row of the figure presents the comparison in terms of RT dissimilarity measure, while the second row shows Cohen's Kappa values. Y-axes are shared along the rows to allow easy comparison. Exp-TMO and Exp-VP values are presented in the first and second columns, respectively. For each plot, dashed lines represent the mean agreement of the spammers, while solid lines represent the mean agreement of real observers in the experiment. 75% percentile range of each value is also visualized as an area around the mean values.

Fig.5 shows that the mean agreement of spammers has a narrower range than real observers in both measures. Additionally, as expected, it is shown that the higher proportion of spammers lowers the agreement of real observers for both experiments. This can be observed as an overlap in 75% percentile ranges of agreement values of real observers and spammers. We notice a significant overlap at 30% for Exp-TMO and 40% for Exp-VP. This makes the detection of the spammers difficult by using mean observer agreement values as a measure.

As previously analyzed in Section 4, observers in Exp-TMO have lower mean agreements for both measures, *i.e., higher RT dissimilarity, lower Cohen's kappa values* compared to observers in Exp-VP due to the difference in the subjectivity of the two tasks. Therefore, for a given spammer ratio, the real observers' and spammers' mean agreement are further apart in Exp-VP compared to the Exp-TMO experiment. In other words, the lower subjectivity of Exp-VP provides a higher tolerance to spammer ratio compared to Exp-TMO.

Influence of the spammer behaviour intensity: A spammer may not have malicious goals from the beginning of the experiment and provide honest opinions to a proportion of the experiment until she/he loses attention. To test the effect of intensity of such irregular behaviors, we used spammer profiles defined in Sec. 5.1. We fixed the spammer proportion to 20% and periodically increased the intensity of spammer behaviors. Fig. 6 presents the agreement of real observers and spammers with solid and dashed lines, respectively. The first row presents the mean agreement of observers in terms of RT dissimilarity, while the second column shows

the mean Cohen's kappa values. In each plot, the X-axis represents the intensity of the spammer behavior. As expected, increased spammer behavior intensity leads to a lower agreement among all observers. An important observation is that the mean agreement 75% percentile ranges of real observers and spammers are further apart even with low spammer behavior intensities in Exp-VP. Similar to the spammer proportion effect, this can be explained with the lower subjectivity of the task in Exp-VP compared to Exp-TMO. Additionally, in Exp-TMO, the overlap between the 75% percentile ranges of real observers and spammers decreases with higher spammer behavior intensity levels.

6. CONCLUSION AND DISCUSSION

In this study, we conducted comprehensive analyses of the irregular observer behaviors on two different PC QoE assessment scenarios. First, we investigate the effect of task differences on the collected subjective preferences. Krippendorff's alpha values and Barnard's test results indicate that the tested QoE assessment tasks' level of subjectivity is different. Furthermore, we evaluated mean individual observer agreements for each experiment with RT dissimilarity and Cohen's Kappa coefficient. Both metrics indicated a significant difference in terms of individual observer agreements between the experiments. Finally, we defined four different spammer profiles and analyzed the influence of individual spammer profiles, spammer behavior intensity, and spammer proportion on the observer agreement distributions. Both RT dissimilarity and Cohen's Kappa coefficient are good indicators of irregular behavior in PC QoE assessment experiments. Both metrics provide acceptable separation between agreement values of spammers and real observers up to $30 - 40\%$ spammer proportion. Our findings also indicate that the QoE assessment tasks with lower subjectivity have a higher tolerance to spammer proportion among the observers.

Our findings indicate that the inter-observer agreement is highly task-dependent. Therefore, statistical measures indicating a general "good" or "bad" agreement level can only be used relatively. In order to increase the robustness to task differences, simple thresholding of agreement measures should be avoided. Additionally, the subjectivity of the QoE assessment tasks influences the spammer tolerance of agreement measures. QoE assessment tasks with higher subjectivity should use additional precautions, such as golden units, to decrease the overlapping range of agreement values between spammers and real observers. Finally, while providing insight, using mean agreement value to detect spammers may not be sufficient. Methods should utilize approaches that can benefit from the measures between individual observers rather than relying single agreement value for each observer.

To sum up, to fully benefit from the values provided by PC methodology, better tools to detect spammers are required. The increasing popularity of the data-driven QoE assessment models and the growing adoption of crowdsourcing platforms augments this need.

7. REFERENCES

[1] "Prolific," `https://www.prolific.co/`, Accessed: March 2021. [Online].

[2] ITU-R, "Methodology for the subjective assessment of the quality of television pictures," ITU-R Recommendation BT.500-13, 2012.

[3] Emin Zerman, Vedad Hulusic, Giuseppe Valenzise, Rafał K. Mantiuk, and F. Dufaux, "The relation between mos and pairwise comparisons and the importance of cross-content comparisons," in *HVEI*, 2018.

[4] J. Ayush, S. Akash Das, P. Aditya, and W. Jennifer, "Understanding workers, developing effective tasks, and enhancing marketplace dynamics: A study of a large crowdsourcing marketplace," vol. 10, no. 7, pp. 829–840, Mar. 2017.

[5] Yingtong Dou, "A review of recent advance in online spam detection," 2019.

[6] Huajie Shao, S. Yao, Andong Jing, Shengzhong Liu, Dongxin Liu, Tianshi Wang, Jinyang Li, Chaoqi Yang, R. Wang, and T. Abdelzaher, "Misinformation detection and adversarial attack cost analysis in directional social networks," *2020 29th International Conference on Computer Communications and Networks (ICCCN)*, pp. 1–11, 2020.

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

[8] Erik Reinhard, Michael Stark, Peter Shirley, and James Ferwerda, "Photographic tone reproduction for digital images," in *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, 2002, pp. 267–276.

[9] Grzegorz Krawczyk, Karol Myszkowski, and Hans-Peter Seidel, "Lightness perception in tone reproduction for high dynamic range images," in *Computer Graphics Forum*. Citeseer, 2005, vol. 24, pp. 635–646.

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

[11] G. Abhishek, P. Mathis, H. Wolf, and D. Frédéric, "Tone mapping operators: Progressing towards semantic-awareness," in *2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 2020, pp. 1–6.

[12] George Alfred Barnard, "A new test for $2\times$ 2 tables," *Nature*, vol. 156, no. 3954, pp. 177, 1945.

[13] Klaus Krippendorff, "Estimating the reliability, systematic error and random error of interval data," *Educational and Psychological Measurement*, vol. 30, no. 1, pp. 61–70, 1970.

[14] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, 1960.

[15] Mary McHugh, "Interrater reliability: The kappa statistic," *Biochemia medica : časopis Hrvatskoga društva medicinskih biokemičara / HDMB*, vol. 22, pp. 276–82, 10 2012.

[16] Wan-Yu Lin and Daniel J Schaid, "Power comparisons between similarity-based multilocus association methods, logistic regression, and score tests for haplotypes," *Genetic epidemiology*, vol. 33, no. 3, pp. 183—197, April 2009.

[17] David J. Rogers and Taffee T. Tanimoto, "A computer program for classifying plants," *Science*, vol. 132, no. 3434, pp. 1115–1118, 1960.