Expertise screening in crowdsourcing image quality

Vlad Hosu, Hanhe Lin and Dietmar Saupe
Department of Computer and Information Science, University of Konstanz, Germany
Email: {vlad.hosu, hanhe.lin, dietmar.saupe}@uni-konstanz.de

***Abstract*—We propose a screening approach to find reliable and effectively expert crowd workers in image quality assessment (IQA). Our method measures the users' ability to identify image degradations by using test questions, together with several relaxed reliability checks. We conduct multiple experiments, obtaining reproducible results with a high agreement between the expertise-screened crowd and the freelance experts of 0.95 Spearman rank order correlation (SROCC), with one restriction on the image type. Our contributions include a reliability screening method for uninformative users, a new type of test questions that rely on our proposed database[1] of pristine and artificially distorted images, a group agreement extrapolation method and an analysis of the crowdsourcing experiments.**

***Index Terms*—crowdsourcing, image quality assessment, reliability screening, expertise screening**

I. Introduction

Experts and naive observers have very different opinions in their judgments of aesthetics [1]. Does this apply to image quality assessment as well? If it does, should we care more about expert-like opinions or those of lay-people?

Image quality relates to user satisfaction, depending on the context in which the stimulus is presented. If the goal is to satisfy everyone, including the most critical 5%, we need to precisely identify high quality images. If we cannot do this, we cannot create sufficiently good annotations to train objective IQA methods, meaning we cannot employ these methods in highly demanding applications such as selecting images for professional magazines, web-sites, etc. Therefore, expert opinions are essential, as long as they are more accurate than those of lay-people.

In this work we study the differences between expert opinions and crowd opinions. Given that expert opinions are useful, we aim at recommending a set of control methods for screening participants with the purpose to improve the reliability of the crowd for image quality assessment, thereby bridging the gap between crowdsourcing results and those of an expert group. The main purpose of the screening method is to enable crowdsourcing to yield expert-level annotations for larger image datasets at an affordable cost in comparison to engaging experts.

II. State of the art

Hekkert et al. [1] compared the judgments of experts and lay people with regard to the aesthetics of visual artworks. They found large differences which were explained on the basis of weights attributed to specific features of the artworks. Adult inexperienced subjects were guided by semantic features (i.e., content or themes) in their value judgments, whereas experienced observers judged art, irrespective of content, in terms of formal, stylistic, and relational properties.

We expect that similar mechanisms apply to value judgments like IQA. Experienced photographers are required to understand many types of image degradations and how they interact with aesthetics in order to take better photos. The expert's domain knowledge should be reflected in the difference of their quality judgments when compared to naive observers.

For a first study of the connection between experts and naive observers in IQA, we face several challenges. We need to choose the right set of images to assess, collect reliable opinions and compare them to find and explain the differences. More specifically, we study the effect of test questions on image degradations, and how these can improve the performance of the crowd.

A. Reliability testing

We collected crowdsourcing data from naive and expert participants. To ensure the quality of the results we considered several types of reliability checks that have been proposed in the literature. Crowd-workers are unreliable mainly due to a lack of attention to the task, a misunderstanding of the task, and intentional cheating. The two main approaches used to control the reliability of a crowd experiment are user rating based screening (URS), and a combination of gold standard, content, and consistency questions [2].

1) User rating-based screening (URS): Several such mechanisms have been introduced in the literature making different assumptions about worker behavior.

- In a well known URS approach, called crowdMOS, proposed by Ribeiro et al. [3], the authors detected unreliable workers based on each worker's level of disagreement with the mean opinion. Their method considered the sample correlation coefficient between the user rating of a worker and the global average rating. A fixed threshold, 0.25 in their paper, was used to remove workers that do not agree with the majority.
- The standard ITU-R BT.500 proposed a similar method for screening users [4]. The assumption is that reliable participants provide ratings close to the mean. One counts the number of each subject's high z-scored deviations from the MOS. Participants that consistently have extreme opinions, both low and high, around the mean are removed.

2) Gold standard test questions: Hossfeld et al. [2] had suggested that gold standard questions were often easy to create, even automatically. However, they cannot be directly applied to test subjective properties. They are more appropriate for objective questions related to content. These control

[1]The database consisting of 300 images with annotated degradation types is available at http://database.mmsp-kn.de.

mechanisms are not accounting for context factors, such as experimental environment or visual acuity issues.

We surmount the limitations of standard content questions by proposing a testing methodology that uses generated image quality degradations as gold standard test questions. To the best of our knowledge, this is a first time artificial degradations have been used to control IQA reliability. When applied to screening unreliable workers, during our IQA task, the method showed promising results.

B. Measuring reliability

We are interested in a consistent performance of users in IQA. The Intra-class Correlation Coefficient (ICC) is one of the most popular inter-rater reliability (IRR) measures. Among others, ICC measures the consistency or conformity among raters. Throughout the paper we use the one way random model [5]. A high ICC value means that most of the variance of the ratings is explained by the differences between individual images, and not by different rater opinions.

The range of the ICC varies widely, depending on the rating type, the experimental environment, and provenance of the participants. For instance, in Absolute Category Rating (ACR) tasks for image aesthetic appeal judgments, Siahaan et al. [6] reported an ICC of 0.40 for crowdsourcing, and 0.94 in the lab. However, in the IQA database CID2013 [7], the ICC in the lab was 0.68. In our experiment involving freelance experts, the ICC on our entire dataset of 300 images was 0.58, slightly smaller than in CID2013. However, the freelancer ICC on our subset of 100 pristine and artificially degraded images, which are more similar to the type of images used in the CID2013 dataset, the ICC was at the same value of 0.68.

III. Database creation

To validate our proposed approach, we create an IQA database of 300 images in two stage. In the first, we started by choosing 500 images from YFCC100m [8] that were all taken with high-end camera models, had an appropriate Creative Commons license, and had been captured at a focal length below 70 mm. This discouraged motion blur and noise due to low light. We then manually reduced the set down to 161 images that formed an initial un-degraded set.

We crowdsourced the un-degraded images in order to determine the types of quality degradations present in each. The top 50 images with the least amount of reported degradations were chosen as our pristine image set. This experiment will be further explained in Sec. IV-D.

The 50 pristine images were then degraded using a set of 12 types of degradations to form the set of 50 degraded images. One magnitude for each degradation was used, chosen such that degradations are clearly noticeable when compared to the pristine images. The degradations used were: over-sharpening, pixelation, jitter distortion simulating low quality super-resolution, camera grain, exposure changes, JPEG compression, color fringing, over-saturation, motion and lens blur.

In the second stage, we randomly selected 200 images from YFCC100m, different from the 50 pristine images. In both stages, the images were rescaled and center cropped to 960 × 540 to facilitate experimental use.

TABLE I
Experimental settings

	Ask for		Test on	No. of	Cost /
	degradation	quality	degradation	judgments / image	judgment (US ¢)
Freelancer (F)	✓	✓		19	7.02
Crowd 1 (C1)	✓	✓	✓	60	0.39
Crowd 2 (C2)*	✓	✓		60/150	0.19
Crowd 3 (C3)		✓		100	0.13

* Experiment setup C2 was done twice (R1 and R2), several months apart.

IV. Experiments

A. Overview

We carried out five experiments. The first experiment collected expert opinions, the remaining four were used to poll the crowd. Overall, we had three unique experimental configurations which asked about technical image quality on a 5-point ACR scale. In experiment C1 we asked about the presence of degradations and additionally provided test questions on degradations, while in C2 we only asked about degradations. The experiment involving freelancers (F) had the same setup as C2. See Table I for details.

With regard to the degradations questions, we asked users whether they noticed any degradation present in the image. If their answer was "yes" then they were asked to classify the distortion type into four major non-exclusive classes: artifacts (compression, pixelation, noise, etc), blur (incorrect focus, camera shake, motion blur, etc.), contrast (excessive sharpness, over/under-exposure, etc), colors (color shifts/fringing, over-saturation, etc). An option for a free-form answer for "Other" degradation type was also available.

Each experimental setup presented similar instructions with minimal adaption of the text description depending on the types of questions asked. For all experiments we provided examples and detailed descriptions of nine common distortion types and advised users to rate technical quality, independent of aesthetics as much as possible.

B. Freelancers experts

We posted our study on freelancer.com in four batches, over a span of several weeks, inviting workers with professional photography experience and designers to participate. We selected 19 freelancers out of 49 candidates based on their portfolios.

In addition to taking the test via the crowdflower.com interface, the freelancers were asked to fill in a questionnaire about their experience with photography or design. 12 of 19 freelancers reported to have more than three years of professional photography experience, 13 had more than 20 reviews, all with a rating above 4.8 (of 5). Each freelancer took more than one hour to finish annotating all 300 images.

C. Crowdsourcing pre-study

Before starting the main experiments, we ran a pre-study (C0) that helped us choose 50 pristine images and collect information about the perception of degradations. All crowdsourcing experiments were run on Crowdflower.com.

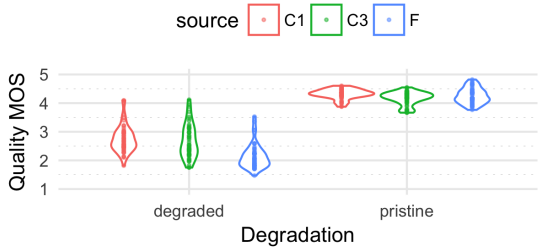


Fig. 1. Degradation ground truth vs quality MOS distributions on the 50 pristine + 50 artificially degraded image dataset, in experiments: C1 where users are tested on degradations, in C3 only presented with quality questions, and F where freelancer experts are presented with degradations and quality questions, but post-screened on 70% accuracy on test questions.

We presented 161 images (manually selected as un-degraded) and their degraded versions using our 12 types of artificial degradations. Users were asked to choose which of the 12 types of degradation is visible in each image. We used 25 gold standard test questions, directed at the presence and type of degradation. Our analysis showed that some degradations are easily confused, i.e., motion blur is confused with lens blur. Consequently, in further experiments we assigned each of the 12 types of degradation to one or more of four top level classes, in order to improve the diagonal response in the confusion matrix of the degradation types. The four classes are: blur (lens and motion blur, color fringing, over-saturation by color diffusion), contrast (exposure, over-sharpening), artifacts (JPEG compression, jitter, pixelation, grain/noise), colors (color fringing, over-saturation).

D. Main crowdsourcing experiments

The main experiments C1, C2 R1 and C2 R2 (repeated after several months) and C3 collected quality ratings for our 300 image database. In C1 and C2, users were asked about the presence and type of quality degradations as well as technical quality rating, whereas only technical quality rating were asked in C3. In C1 we used test questions, whereas in C2 and C3 none were used.

We used the 100 designed images for the test questions. In the beginning of a job, crowd workers took a quiz to identify the presence and type of degradation. Crowd workers that passed a minimum accuracy of 70% were allowed to enter the job. Apart from screening unqualified crowd workers in quiz mode, hidden test questions were also used to track qualified workers' performance on ongoing work. If at any point during work the performance of a qualified worker fell below 70% all her judgments were discarded and she was not allowed to continue. Additionally, users that finished the task too quickly, at a greater rate than one answer per second, were disqualified and their judgments were discarded.

On Crowdflower there is no option to dynamically exclude workers that have participated in a previous experiment. After the experiments finished, we made sure to keep only unique workers for each experiment.

Some initial results can be seen in Fig. 1. We show the distributions of quality MOS on the 100 designed images in three experiments. With respect to degraded images, the difference in the shape of the MOS distribution is similar between the crowd C1 and experts F, except for a global bias. However, when it comes to pristine images, we observe a much larger difference in distributions between crowd experiments and the freelance experts. Overall, experts more consistently rated degraded images as bad quality, and differentiated more quality levels for pristine images. This suggests experts have a higher sensitivity to fine degradations. The connections between the experiments remains to be analyzed later.

V. Methods

A. Removing uninformative users

We introduce a new type of reliability control, starting from a simple observation. Many crowdsourcing workers have a tendency to choose too often the same answer option. In an extreme case, when participants intentionally give inappropriate answers, this comes down to the so-called line-clicking behavior. We believe, this happens for honest workers as well. When faced with situations where the worker is not confident in her answer, she will revert to a default, safe choice. This often happens to be the middle of the ordinal scale.

We propose to screen those workers with an unusually high frequency for any single answer choice. We first compute the frequency $\phi(c)$ of each answer choice $c \in \{1..5\}$ of the ACR scores, where $\sum_{c\in\{1..5\}} \phi(c) = 1$. Let $P = \max_c(\phi(c))/(1 - \max_c(\phi(c)))$. If $P > 3$, then the user will be considered unreliable and removed. The threshold $P = 3$ was empirically chosen such that only very uninformative users were removed. For all freelancers the indicator $P < 0.8$, suggesting highly informative judgments.

B. Estimating an experiment's repeatability

An experiment's repeatability is given by the agreement between the results of two or more repetitions under similar conditions e.g. environment, type and number of participants. We estimate the agreement with respect to SROCC between the original experiment and its virtual repetition, by extrapolating the agreement between the aggregated scores of two random halves of the users when the group size is doubled.

1) Repeatability measures: An experiment's repeatability has been defined in terms of various measures, for instance with respect to Cronbach's alpha (and a standard consistency table), or with respect to a precision measure such as the absolute difference below which repeated test results should lie in, with a probability of 95%.

We propose to use a simple, but more flexible coefficient which relied on a commonly used agreement measure in IQA experiments, the SROCC. Thus, we relate repeatability to the average SROCC between two or more repetitions of an experiment. From the numbers available in the literature an IQA experiment's repeatability is very high if the SROCC between the MOS of the repetitions is above 0.95. In practice, the numbers reported are in the range 0.93-0.98. For instance the well known IQA database TID2008 [9] reports an agreement between multiple lab studies of 0.93 to 0.96 (SROCC).

An important benefit for using the Spearman correlation is its invariance to monotonic transformations of the rating scale. Hence, if one study inadvertently introduces some non-linearity this will not be counted as part of the agreement, as would happen for other measures (mean absolute error, Pearson correlation, Cronbach's alpha, etc.)

2) Estimating repeatability with more participants: When increasing the number of participants in a study, the confidence intervals of the MOS decrease, and agreement increases.

If we knew the repeatability of our freelancer study (F) in terms of SROCC, we could compare it with the observed correlation between the freelancers and the crowd experiments (C). Finding that crowd (C) and freelancers (F) have an agreement that is similar to the repeatability of the freelancers, would imply that the crowd can repeat the freelancer experiment. Therefore, we estimate the repeatability of our freelancer study, without engaging new experts in another experiment.

A first option is to perform a user-level bootstrapped sampling with replacement, and compute the MOS and SROCC between two sampled groups. However, the result is underestimating the expected agreement: the agreement of groups of 17 experts, sampled with replacement, is lower than that of 8 experts, sampled without replacement (Fig. 2).

It is worth mentioning that the Spearman-Brown prediction formula [10] connects psychometric reliability to test length and could be applied to a similar purpose as ours. However, the method is able to extrapolate intra-class correlation (ICC) to larger group sizes, and is not intended for SROCC.

We propose an alternate solution, which involves extrapolating the agreement using a function that is fitted to the bootstrapped agreements (without replacement), computed up to half the initial population size. In our case, it means we computed agreements between groups of 1 up to 8 users. We found that $f(x) = 1 - a/(x+b)$ is a good fit to our data, both for the crowd and the freelancers. We studied the extrapolation accuracy when f is fitted to part of the crowd experiments data, up to 8 average judgments per image (equivalent to a group of size 8), and evaluated on the rest of the data up to 40 average judgments per image. For crowd experiments we are required to work with average judgments per image rather than group size, as crowd workers most of the time annotate only part of the dataset. The extrapolation errors computed on the crowd data are small throughout, with a mean absolute error (MAE) of 0.0018, and thus a high accuracy.

The predicted repeatability of groups of 17 freelancers is 0.956 SROCC. The confidence interval (CI) of the estimation is small. The error is expected to be about ± 0.019 based on the CI of groups of size 8. We computed these values for 17 freelancers as two of participants did not perform sufficiently well on the degradations test questions (below 70% accuracy) and were removed from the initial group. The plot in Fig. 2 is based on the group of 17 freelancers.

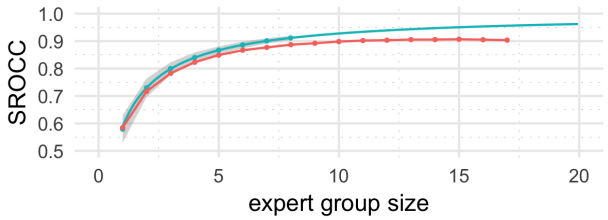


Fig. 2. Predicting the agreement of larger sized groups of experts. The sampling with replacement prediction is shown in red, and is heavily underestimating the more precise prediction of our model. Using our fitted model $f(x) = 1 - 0.79/(x + 0.92)$, groups of 17 experts are expected to have an SROCC of 0.956. Confidence intervals on the bootstrapped sampling without replacement mean SROCC are shown in gray.

VI. Analysis and Discussion

We used two types of reliability screening mechanisms. The first identifies unusual patterns of behavior among the participants (URS methods), the other filters users based on test questions. For some of our experiments we studied the effects of screening users based on their accuracy on test questions in the post-experiment analysis. We screened participants in all experiments using URS methods.

A. User rating screening

We applied three methods that perform reliability checks and screen under-performing users. We used the same threshold of 0.25 PLCC for the random clicker detection [3] method, and the standard parameters for a normal distribution for ITU-R BT.500 [4]. For our proposed method we remove uninformative users with indicator $P > 3$. The improvements in intra-class correlation (ICC) can be found in Table II.

None of the freelancers are screened by URS methods, all participants proving to be reliable, whereas we have detected unreliable users in each crowdsourcing experiment. The correlations between all experiments are shown in Table. III.

TABLE II
ICC after URS (screening), on the 200 random images set

	initial*	uninformative**	random**	outliers**	all*
F	0.551	0.551	0.551	0.551	0.551
C1	0.445	0.458	0.446	0.447	0.461
C2 R1	0.332	0.361	0.342	0.332	0.364
C2 R2	0.312	0.339	0.346	0.316	0.366
C3	0.448	0.470	0.468	0.449	0.471

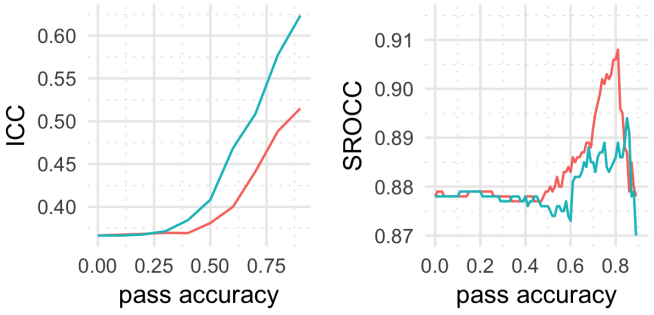
* the raw ICC before any screening is applied, and after all are applied
** independently applied screening of uninformative users, random clickers and outliers via ITU-R BT.500

TABLE III
SROCC for MOS of the 200 random images, after URS

	F	C1	C2 R1	C2 R2	C3
F (all 19)	-	0.91	0.88	0.88	0.87
C1	0.91	-	0.95	0.97	0.95
C2 R1	0.88	0.95	-	0.96	0.96
C2 R2	0.88	0.97	0.96	-	0.97
C3	0.87	0.95	0.96	0.97	-

B. Test questions screening

The screening procedure involves evaluating the accuracy of participants on test questions on the 100 designed images (50 pristine + 50 artificially degraded). The participants'



(a) (b)

Fig. 3. (a) Worker ICC of screened C2 R2 vs the pass accuracy threshold on test questions (b) SROCC between MOS of screened C2 R2 and F vs the pass accuracy used on test questions. In cyan/blue ICC or SROCC of test question on quality, and in red for degradations. In both (a) and (b) accuracy is evaluated on the 100 designed images, whereas SROCC is computed on the 200 random set.

performance (ICC and SROCC with freelancers) is evaluated on the remaining 200 random image set. This ensures a fair validation, meaning our findings generalize beyond the "training set" of 100 designed images.

We devised two testing scenarios, identification of 1. quality and 2. degradations. In order to compute a measure of accuracy for the two scenarios, we defined correct answers based on whether an image is pristine or degraded. A user's answer is correct if she chooses at least one of the valid options which are defined as follows.

1) Quality: on the 5-point scale, pristine images have valid scores of $\{4, 5\}$, whereas degraded images have $\{1, 2\}$

2) Degradations: pristine images are considered to not show any degradation, whereas degraded images can have several of the 4 main classes of perceptual degradations (Sec. IV-D).

As a result of this evaluation, two of the freelancers had scored lower than expected on the accuracy of degradations ($<$0.7) and were removed, leaving 17 freelancers with an ICC of 0.582 (previously 0.551). Both types of accuracies are shown in Table IV for each experiment, excluding C3 where we did not ask for degradations. Freelancers perform well throughout, whereas the participants in C1 which were tested on degradations have a higher average accuracy on degradations (bias caused by test questions passing accuracy of 70%). Nonetheless the crowd C1 accuracy on quality remains lower than F, but higher than C2. Participants in C2 (R1 and R2) have similar low accuracies on both degradations and quality, not being tested on any of them.

TABLE IV
AVERAGE USER ACCURACIES ON TEST QUESTIONS.

	F (all 19)	C1	C2 R1	C2 R2
degradations	0.78	0.85	0.72	0.70
quality	0.78	0.67	0.59	0.63

C. Performance of test questions screening

Until now we have shown that reliability screening via URS methods and test questions screening improve the performance of the crowd, with respect to inter-rater reliability as well as agreement with the experts. Here we analyze the consistency of the improvements due to test question screening, as well as the effects of the two testing scenarios.

TABLE V
AGREEMENT BETWEEN CROWD AND FREELANCERS ON MOS.

	MAE	RMSE	SROCC	PLCC	avg. votes**
C3	0.303	0.397	0.866	0.887	73
C2 R1	0.298	0.387	0.878	0.893	48
C2 R2	0.294	0.385	0.880	0.894	133
C1	0.267	0.340	0.906	0.918	56
C2 R2 screened*	0.263	0.340	0.909	0.918	17
C2 R1+R2 screened*	0.258	0.330	0.915	0.923	48

* users were screened based on an 80% minimum passing accuracy on degradations test questions

** average number of votes per image; more relevant than number of users

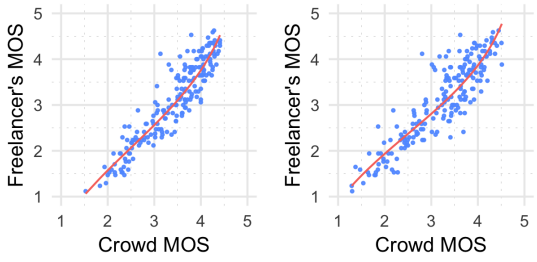
In experiment C2 R1 and R2 participants were presented with questions about degradations, however, they were not required to pass any test questions. Post-screening users based on their answers is different from actually presenting participants with test questions. When a user fails on actual test questions, she has a chance to find out the reason for her failure. However, the difference between the two cases is relatively small. As shown in Fig. 3(b), the SROCC between C2 R2 screened at 70% passing accuracy and F is 0.895, which is just slight lower than 0.906, the agreement between C1 and F (users in C1 were presented with test questions). Nonetheless, the test condition C1 is preferred over C2.

We show the changes in the crowd results when we screen participants based on different thresholds of their accuracy on quality, and as well as on degradations. In Fig. 3 we notice that as the accuracy passing threshold increases, the workers agree better. This effect is stronger when testing on quality. However, the correlation with the expert MOS does not increase as much when screening on quality accuracy. The trend is more obvious for degradations test questions, as we can see in Fig. 3 (b), meaning that those workers in the crowd that can identify degradations well, have a better chance to agree with the experts. The fast drop-off of the curves after 0.8 passing accuracy is caused by having too few judgments on some items (as low as 12 per image). Furthermore, we show that having more participants involved allows to continue increasing the agreement. We tested this by combining both C2 R1 and R2 into a single pool of users, and screening them at 80% accuracy on degradations, see results in Table V.

The analysis concludes that degradations based screening is to be preferred over quality based screening. The suggested experimental setup is one involving actual test questions, and not post-screening. However, the accuracy threshold should be increased from 70% to 80% or higher depending on the availability of highly performing workers.

D. Intrinsic differences between naives and experts

We see a consistent improvement in agreement between the crowd and freelancers as we perform tighter performance based screening. Starting with the most unconstrained experiment C3 which has the lowest agreement to the experts (F), we see a steady increase in C2 (presentation of degradations)



C1 C3

Fig. 4. Freelancer vs crowd MOS on experiments C1 and C3. The level of jitter of the data-points is larger in the case of C3, showing a lower agreement with the experts. The smoothing *logit*-type function aligns the data such that RMSE and MAE are more representative difference measures.

and C1 (test on degradations, screened at 70%). Furthermore, screening the joint users of C2 (R1+R2) at an 80% accuracy continues to increases the agreement. See Table V for more details. SROCC and PLCC were computed directly on corresponding MOS. For MAE and RMSE the scores have been first remapped to compensate for nonlinearities and range differences in the answer profile of the different groups, by fitting a *logit*-type function to the data, as seen in Fig. 4.

Screening on degradations allows us to find those groups of crowd workers that jointly (via MOS) show a better understanding of image quality, closer to that of the group of experts. However, the individual users in the screened crowd and experts perform differently. The 17 experts have a better inter-rater agreement at an ICC of 0.58 compared to the crowd in C1 standing at 0.46. It is only the MOS of the crowd that is more similar to that of experts.

The expected repeatability of the freelancer experiment, measured by the SROCC of groups of 17 freelancers was estimated to be 0.95 ± 0.02. The best screened crowd having provided an average of 48 judgments per image, stands at 0.915 agreement, meaning that it cannot perfectly match expert opinions on the entire 200 random images.

We take a closer look at individual MOS differences. A large $\delta(F, C1, I) = F_{MOS(I)} - C1_{MOS(I)}$ for image I signifies that the crowd is underestimating the quality of an image. When we rank images by decreasing δ, we notice a consistent pattern. All top 12 high δ images that have an above average $F_{MOS(I)} > 3$ represent macro-shots, having a shallow depth of field (DoF). In this kind of compositions, a large part of the image is covered by lens blur, however the content is well focused given the available DoF. See Fig. 5 for the first four images. This may mean that naive users are not aware of the limitations of macro photography and consequently discount the images' technical quality score more than the experts do.

If we remove just these 12 images from the set of 200 test images, the SROCC between C1 having on average 56 judgments per image and the 17 experts on the remaining 94% of images jumps from 0.906 to 0.945. This is a surprising finding, the screened crowd matching the expected repeatability of groups of 17 experts (0.956 ± 0.019). It suggests that the screened crowd, following the outlined procedures in this study, serves well as a substitute for limited groups of experts as long as subtle technical aspects (e.g. limitations due to lens optics) do not come into play.

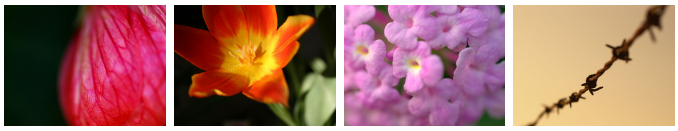


Fig. 5. Images with the largest disagreement between crowd C1 and F MOS. The crowd underestimates quality by 1.42, 1.07, 0.99 and 0.90 points respectively on a scale of $[1, 5]$.

VII. Conclusions

We reach an unexpected conclusion. Unlike aesthetics which show a large difference between experts and naive observers [1], in technical IQA the two groups of participants behave more similarly. A crowd of naive workers, having been screened for reliability and performance on IQA related test questions, is able to repeat at a lower cost (US$70) the MOS of 17 experts that are professional photographers (US$400), except for a few, specific type of images. The key to achieving this are gold standard test questions about image degradations, and URS methods. With these findings, we have created the largest IQA database [11] so far. We expect an extension of the types of degradations used to include spatially varying ones such as depth-based lens blur, could further improve the approach.

Acknowledgment

We thank the German Research Foundation (DFG) for financial support within project A05 of SFB/Transregio 161.

References

[1] P. Hekkert and P. C. W. V. Wieringen, "Beauty in the eye of expert and nonexpert beholders: A study in the appraisal of art," *The American Journal of Psychology*, vol. 109, no. 3, pp. 389–407, 1996.

[2] 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, Feb. 2014.

[3] F. Ribeiro, D. Florłncio, C. Zhang, and M. Seltzer, "Crowdmos: An approach for crowdsourcing mean opinion score studies," in *Acoustics, Speech and Signal Processing (ICASSP), International Conference on*. IEEE, 2011, pp. 2416–2419.

[4] B. T. Series, "Methodology for the subjective assessment of the quality of television pictures," *Recommendation ITU-R BT*, pp. 500–12, 2009.

[5] K. A. Hallgren, "Computing inter-rater reliability for observational data: an overview and tutorial," vol. 8, no. 1, p. 23.

[6] E. Siahaan, A. Hanjalic, and J. Redi, "A reliable methodology to collect ground truth data of image aesthetic appeal," vol. 18, no. 7, pp. 1338–1350.

[7] T. Virtanen, M. Nuutinen, M. Vaahteranoksa, P. Oittinen, and J. Hakkinen, "CID2013: A database for evaluating no-reference image quality assessment algorithms," vol. 24, no. 1, pp. 390–402.

[8] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li, "Yfcc100m: The new data in multimedia research," *Communications of the ACM*, vol. 59, no. 2, pp. 64–73, 2016.

[9] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti, "TID2008 - a database for evaluation of full-reference visual quality assessment metrics," vol. 10, pp. 30–45.

[10] H. C. de Vet, L. B. Mokkink, D. G. Mosmuller, and C. B. Terwee, "Spearman-brown prophecy formula and cronbach's alpha: different faces of reliability and opportunities for new applications," vol. 85, pp. 45–49.

[11] H. Lin, V. Hosu, and D. Saupe, "KonIQ-10K: Towards an ecologically valid and large-scale IQA database," *arXiv preprint arXiv:1803.08489*, 2018.