Crowdsourcing-based Multimedia Subjective Evaluations: A Case Study on Image Recognizability and Aesthetic Appeal

Judith Redi
INSY, TU Delft
j.a.redi@tudelft.nl

Tobias Hoßfeld
University of Würzburg
tobias.hossfeld@uni-wuerzburg.de

Pavel Korshunov
MMSPG, EPFL
pavel.korshunov@epfl.ch

Filippo Mazza
IRCCyN, ECN
filippo.mazza@irccyn.ec-nantes.fr

Isabel Povoa
INSY, TU Delft
misabelpovoa@gmail.com

Christian Keimel
Technische Universität München
christian.keimel@tum.de

ABSTRACT

Research on Quality of Experience (QoE) heavily relies on subjective evaluations of media. An important aspect of QoE concerns modeling and quantifying the subjective notions of 'beauty' (aesthetic appeal) and 'something well-known' (content recognizability), which are both subject to cultural and social effects. Crowdsourcing, which allows employing people worldwide to perform short and simple tasks via online platforms, can be a great tool for performing subjective studies in a time and cost-effective way. On the other hand, the crowdsourcing environment does not allow for the degree of experimental control which is necessary to guarantee reliable subjective data. To validate the use of crowdsourcing for QoE assessments, in this paper, we evaluate aesthetic appeal and recognizability of images using the Microworkers crowdsourcing platform and compare the outcomes with more conventional evaluations conducted in a controlled lab environment. We find high correlation between crowdsourcing and lab scores for recognizability but not for aesthetic appeal, indicating that crowdsourcing can be used for QoE subjective assessments as long as the workers' tasks are designed with extreme care to avoid misinterpretations.

Categories and Subject Descriptors

I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*perceptual reasoning, representations, data structures, and transforms*; H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems—*evaluation/methodology, video*

Keywords

Crowdsourcing, Aesthetics, QoE, Subjective evaluations.

CrowdMM'13, October , 2013, Barcelona, Spain.

DOI: 10.1145/2506364.2506368

1. INTRODUCTION

Crowdsourcing (CS) is a powerful tool for gathering subjective ground truth for large multimedia collections. Big amounts of users (microworkers) can be reached to accomplish a set of small tasks in exchange for a symbolic payment, which is particularly convenient when large user studies have to be conducted. By designing appropriate micro-tasks, sufficiently reliable data can be gathered in an inexpensive and time-effective way. As a result, CS has become a popular tool for media tagging [4], investigation of cognitive responses to media fruition [8], evaluation of privacy filters [13], etc.

Research on Quality of Experience (QoE) [15] relies on understanding user preferences in terms of perceptual quality and overall enjoyment of multimedia. To this end, studies are conducted in a controlled Laboratory (Lab) environment, with fixed lighting and experimental conditions [19], since the goal is to collect information on the user sensitivity to impairments in the media signal [7] and the related quantification of their annoyance. In this context, CS has often been considered not appealing for QoE research, as it would not guarantee the necessary level of environmental control to provide reliable data. Lately, however, the sensitivity-centric definition of QoE has been challenged, and it was shown that QoE depends also on user preferences and personality, context of media usage, and quality of the interaction with the system [15, 20]. With the acceptance of this more encompassing definition of QoE, the interest in using CS for QoE research has grown significantly [2, 7, 11]. Nevertheless, some doubts remain regarding the extent to which CS can provide reliable QoE data.

To understand the benefits and limits of using crowdsourcing in QoE evaluations, we look into how well QoE ratings collected in a controlled lab environment can be replicated by a crowdsourcing experiment. Being QoE a multifaceted quantity [15], we focus specifically on aesthetic appeal, which has been recently shown to play an important role in QoE judgments [20]. Understanding aesthetic appeal of media is of major interest for the multimedia community, which has indeed devoted a lot of effort to it lately [10], although often based on ground truth collected through social media platforms. Interestingly, very few efforts have been made towards quantifying the aesthetic appeal in a more controlled way. In this study, we make a first attempt at collecting more rigorous ground truth on aesthetic appeal of consumer images in a lab environment, and we check to what extent CS can be used to collect the same type of information.

We conducted an experiment in a controlled lab environment, in which the aesthetic appeal of 200 consumer images was rated by 14 paid participants in a single stimulus setup [19]. Along with this quantity, participants also rated the level of recognizability of the content of the image. This second quantity is related to perceptual fluency [17], which is known to have an effect on the aesthetic appeal of works of art. In this study, we wanted to check whether this effect was preserved also when judging the aesthetic appeal of consumer images.

We then replicated the same experiment in a crowdsourcing setting, by using the Microworkers[1] platform. About 390 workers from 16 countries evaluated (subsets of) our images, ensuring a variety in cultural and social backgrounds, which are known to impact aesthetic preferences. Adaptations to the protocol were needed to allow controlling the reliability of the workers, and checks were made using both control questions and timestamp information prior to analyzing the results and comparing them with the Lab data.

In the remainder of this paper, after a brief review of existing work on user studies on aesthetic appeal and crowdsourcing (Section 2), we describe the experimental protocol followed in the Lab and its adaptation to the crowdsourcing evaluation (Section 3). In Section 4, we analyze the reliability of the CS workers, and based on reliable workers only, in Section 5, we compare the outcomes of crowdsourcing experiment with lab experiment. We draw conclusions and possible future extensions of this study in Section 6.

2. BACKGROUND

Being able to model aesthetic preferences of users is a major concern for modern multimedia research. Information on image aesthetic appeal can help in retrieval and recommendation tasks, as well as in optimizing visual Quality of Experience [10, 20]. Before computational models can be created that reliably predict the aesthetic appeal of an image [9], in-depth knowledge is needed on actual user aesthetic preferences. This is not a trivial task, as aesthetic preferences are typically considered to be highly subjective and related to personal implicit experiences [17], cognitive biases, and personal opinions and memories [18]. Nevertheless, some research on the matter has been conducted by means of self assessments, eye tracking experiments and physiological measurements [22]. Color and saliency have been shown to play a major role in the aesthetic and emotional impact of an image [1, 23]. Furthermore, correlation between aesthetic ratings and familiarity has been reported in [3]. Content recognizability has been shown to have an influence on aesthetic appeal in [14, 17], and abstract paintings were found to be less likely appreciated by people with respect to immediate works of art [16].

Based on these studies and on classical geometrical canons (e.g., rule of thirds and golden ratio), researchers in computational aesthetics have proven to be able to capture useful information for the aesthetic assessment of images [9, 3, 16]. Nevertheless, reliable prediction of the aesthetic appeal of images is still to be achieved. To work towards that goal, computational aesthetic researchers need to rely on ground truth of how users judge the aesthetic appeal of large image collections. Since obtaining this sort of data from controlled experiments is expensive in time and cost [9], more and more researchers turn to community-contributed resources (i.e., from popular online image databases, such as *Photo.net* used in [3]) for data collection. These platforms, however, lack a strict protocol for image assessment and some users can create fraudulent accounts to increase their ratings, leading to unreliable evaluations. In this scenario, crowdsourcing seems to be an in-between solution, offering both the opportunity to reach out to large communities of users and controlling the aesthetic evaluation procedures.

Crowdsourcing is a further development of the outsourcing principle, where the granularity of work is reduced to small tasks that can be accomplished within a few minutes to a few hours and do not require a long-term employment. Tasks are often highly repetitive (e.g., image annotation) and are usually grouped in larger units, referred to as *campaigns*. Most *employers* submitting tasks to an anonymous crowd use a mediator in the form *crowdsourcing platforms* that maintains the crowd, manages the employers campaigns and handles the reimbursement of the workers on behalf of the employer after successful completion of the the tasks.

Amazon's Mechanical Turk (MTurk)[2] and Microworkers are typically used commercial Crowdsourcing platforms. MTurk is the largest crowdsourcing platform and is often used in research, as well as in commercial third-party applications; however, it allows only US residents or companies to submit tasks to the platform. The platform used in this contribution, Microworkers, allows not only international employers, but also worker diversity [5], whose geographic location can be chosen directly by the employer.

When it comes to subjective QoE evaluation tasks, Crowdsourcing tests require the presentation and assessment of different media in a suitable web-interface. Instead of implementing an appropriate interface separately for each QoE test, existing and publicly available frameworks as the *Qudrant of Euphoria* [2] and *QualityCrowd* [11] can be used. Chen's *Quadrant of Euphoria* provides an online service for the QoE evaluation of audio, visual, and audio-visual stimuli using pairwise comparison of two different stimuli in an interactive web-interface, where the worker can judge which of the two stimuli has a higher QoE. In contrast, the *QualityCrowd* framework is not an online service, but a complete open-source platform designed especially for QoE evaluation with crowdsourcing. It can be modified with relatively low effort for different assessment tasks (e.g., single or double stimulus) and provides a simple scripting language for creating campaigns including multi-modal stimuli, training sessions and control questions.

3. EVALUATION METHODOLOGY

We investigated aesthetic appeal and its relationship with some of the features analyzed in Section 2 by means of both a Lab-based and a Crowdsourcing-based experiment. To do so, we designed a within-subjects experiment, in which every participant had to evaluate several aspects of a set of images in a single stimulus setup [19]. Four quantities, namely aesthetic appeal, color likeability, familiarity, and recognizability were inspected in the lab environment. In the crowdsourcing setup only two quantities were inspected to simplify the task: recognizability ('how well can you understand what is represented in the image?') and aesthetic appeal ('how beautiful do you think is the image?').

3.1 Image material

We used a database of 200 images, out of which 56 corresponded to the ones used in [20], 26 were crawled from the web, and 118 were selected from the private collection of an amateur photographer. Images were chosen to encompass a wide range of image contents as generally available online, based on their classification into the categories used by *500.com*, an online database for both expert and amateur photography. As a result, images were chosen that could be classified into categories typically used in computer vision research (e.g., Landscapes and People), frequently occurring in social networks (e.g., Food and Fashion) and covering different

[1] http://microworkers.com/

[2] https://www.mturk.com/mturk/

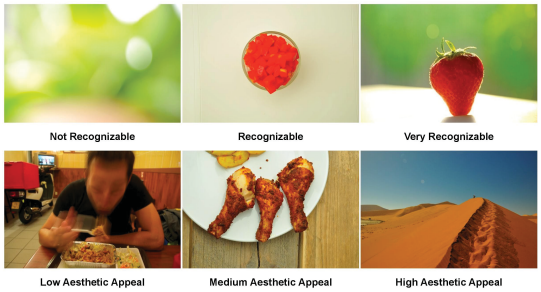

Figure 1: Example of images with different levels of recognizability and aesthetic appeal.

levels of familiarity and recognizability (e.g., Celebrities and Abstract). Images were also selected to roughly span a wide range of aesthetic appeal, based on the ratings already awarded to some of them on the website *500px.com*.

3.2 Lab-based Experiment

Fourteen paid participants took part in the Lab experiment, mostly originating from Europe. They were initially briefed about the general setup and their task. Then, they went through four short training sessions (each used 3 images, reflecting the evaluation scale) to ensure (1) the participant's acquaintance with the task and (2) the anchoring of the scoring scale for each quantity.

Each participant was then asked to assess color likeability, familiarity, recognizability and aesthetic appeal of each of the 200 images. They used four (one per quantity) 5-point discrete numerical scales, ranging from 1 being the lower score and 5 being the higher score. Semantic labels were added at the ends of each scale ("Bad Color" and "Excellent Color", "Not Familiar" and "Very Familiar", "Not Recognizable" and "Very Recognizable" or "Bad aesthetic appeal" and "Excellent aesthetic appeal", respectively). To avoid distraction during the image observation, these scales were kept in a follow-up separated screen.

To avoid fatigue effects that could harm the data collection procedure, due to the elevated number of images, the dataset was randomly split in two sets of 100 images each, to be evaluated by the same participant in two sessions, to be performed in different days. Each session lasted on average 40 minutes per participant, including a short break after scoring the first 50 images to minimize fatigue.

The experimental set-up followed the ITU-R BT.500 recommendation [19] and throughout the whole experiment, no time constraint was given for image observation and scoring.

3.3 Crowdsourcing Experiment

To repeat the experiment in a crowdsourcing environment, it was necessary to deal with two issues: (1) the fact that crowdsourcing tasks should not last longer than 5 to 10 minutes and (2) the risk of unreliable behavior of some of the workers, because of the distributed and remote nature of the test environment. Some adaptation in the experimental protocol was therefore needed to address these issues.

First of all, instead of two sessions with 100 images each, the crowd-based test consisted of 13 campaigns with 20 images each, where 5 of the images were the same for all campaigns to allow re-alignment and scale anchoring purposes. These 5 images corresponded to the 0th, 25th, 50th, 75th, and 100th percentiles of the distribution of all aesthetic quality scores as determined in the lab-based evaluation of the 200 images. The remaining 15 images per campaign were unique to each campaign. Due to this split, each worker in the crowd-based test only evaluated a subset of the original image set. Each worker could also participate in multiple campaigns.

To address the second issue, we implemented reliability control mechanisms to identify and filter out ratings from unreliable users or wrong test conditions [7]. Details on various reliability mechanisms for crowdsourcing experiments can be found in [6] and references therein. Unreliable user rating may be caused by language problems or wrong test conditions due to software errors or hardware incompatibilities, and need to be filtered out in order to avoid a falsification of QoE results. Additionally, there may also be *cheating* users who try to submit invalid or low quality work in order to reduce their effort while to maximizing their received payment, especially when this is very small [21]. We included therefore content questions [12, 7] in each campaign of 20 images after the 5th and 15th images. Furthermore, we targeted countries with an adequate proficiency in the English language, with an English speaking population larger than 10 million people or than 50% of the total population, as all test instructions were provided in English only. In order to limit the workers' participation to specific geographic regions, we used the Microworkers platform. We identified three regions in which workers could correspond to the above characteristics. Region 1 (CS-R1) corresponded to North America and major English speaking countries, such as USA, UK, Canada, and Australia, region 2 (CS-R2) corresponded to Western Europe, including workers from France, Germany, Italy, Ireland, the Netherlands, and Sweden, and region 3 (CS-R3) corresponded to Asia, including workers from Bangladesh, India, Pakistan, Philippines, Singapore, and Thailand. Each campaign was therefore replicated three times for each of the three geographic regions considered in this evaluation, resulting in a total of 39 campaigns.

We used the QualityCrowd [11] framework due to its flexibility and therefore easy adaptation to the task of aesthetics and recognizability evaluation. Similarly to the lab test, we also included a mandatory training to introduce the worker task and the same images used for the recognizability training in the lab experiment were used to allow workers practicing with the experimental interface. Each worker was presented with the image to be evaluated in a web interface that also provided two discrete five point scales to rate the content recognizability and aesthetic appeal of the shown image, similar to the computer-based interface used in the lab test. It is important to note that both questions were displayed on the same page as the corresponding evaluated images with recognizability question being on the left and aesthetic on the right, both below the image.

For each of the 39 campaigns, 30 different users participated and rated 20 images for 0.30 USD. In total, 28,080 images were rated consuming about 85 working hours at a total cost of 351 USD.

4. CROWDSOURCING RELIABILITY

Before comparing Lab and crowdsourcing results, the reliability of the crowdsourcing users has to be analyzed in order to identify and filter out unreliable user ratings. In the following, the results from the 13 different crowdsourcing campaigns are investigated. As mentioned in Section 3.3, each worker could participate in multiple campaigns. Figure 2 shows the histogram of the number of campaigns conducted by a single worker. It can be seen that regions CS-R1 and CS-R2 lead to similar results, while CS-R3 was significantly different. For CS-R1 and CS-R2, 6.61 and 7.36 campaigns were completed on average per user, respectively. Asian users (CS-R3) on average participated only in 2.47 campaigns. While at most $13 \cdot 30 = 390$ different workers could have participated per region, there were only 59 (CS-R1), 53 (CS-R2), and 158 (CS-R3) differ-

ent workers, respectively. The higher user diversity in R3 may be caused by higher competition, as the workers are mainly located in Asia for Microworkers.com [5]. As a consequence, 14 and 15 workers from CS-R1 and CS-R2 are able to participate in all 13 campaigns, while no one from CS-R3 completes all campaigns.

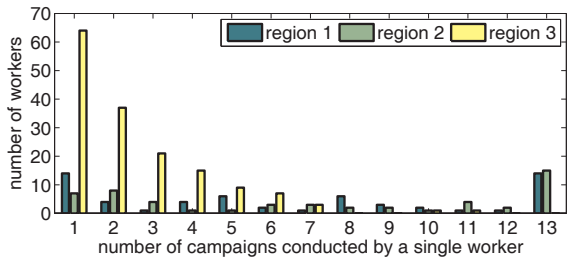

Figure 2: Number of campaigns conducted by a single worker

As a prerequisite to define a worker as 'reliable', all content questions about the images had to be answered correctly by an individual user. Figure 3 shows that the ratio of 'reliable' workers is similar for CS-R1 and CS-R2 with about 90 % over all campaigns. In contrast, only 70 % of workers from CS-R3 correctly answered all content questions. This discrepancy could be due to both language problems or cheating; either way, evaluations from these users could not be considered reliable, and were filtered out from the analysis presented in Section 5.

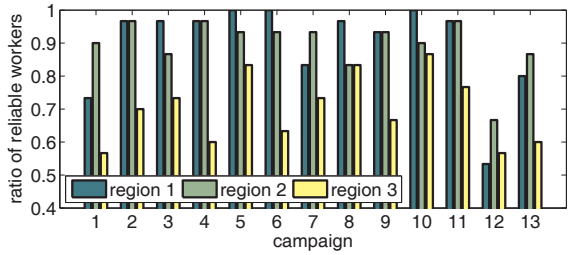

Figure 3: Ratio of workers who answered all questions correctly

The completion time per task was also considered for each user. A median task completion time of 3 min is observed for CS-R1 and CS-R2, while CS-R3 leads to 4 min. Taking a closer look at the mean task completion time reveals that for CS-R1 and CS-R2 the median task completion time is close to the mean completion time. However, for users in CS-R3, the average task completion time is significantly larger than the mean values. Thus, there are users with very large observation times for some images. The observation time per image is measured as the time from when the image is displayed until the time the user rating is given. Figure 4 shows the cumulative distribution function (CDF) of the standard deviation of the image observation duration per user in the different regions. Again, the curves from CS-R1 and CS-R2 overlap. However, the results for R3 are significantly different. In order to filter out users not rating seriously and being distracted during the subjective test, all users with a standard deviation of the image observation time larger than 20 s were rejected. This value was chosen to accommodate possible variations in download speeds of different users but reject users with significantly high variations in completion times.

Finally, unreliable participants were also identified as those rating images in a way that is significantly different with respect to the rest of the population. These outliers were also detected according to [19] and excluded from the subsequent analysis.

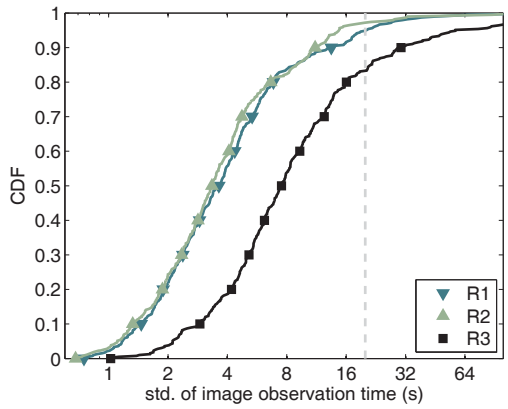

Figure 4: Cumulative distribution function (CDF) of the standard deviation (STD) of the image observation duration per user

Eventually, 14 % (CS-R1), 13 % (CS-R2), and 43 % (CS-R3) of the workers were filtered out, respectively.

5. LAB VS. CROWDSOURCING RESULTS

We computed normalized MOS (Mean Opinion Scores) from the lab and crowdsourcing experiments according to [19]. For lab-based scores, after rejection of one outlier participant, we normalized the ratings of each participant by subtracting from each individual score the mean score value for that participant and dividing it by the standard deviation of all the ratings of that same participant. Similarly, the scores were normalized for each crowdsourcing campaign separately. Although the original scores can also be compared, our experiments showed that the normalization allows for a better comparison, demonstrating all the disparities clearer. Normalized MOS for recognizability and aesthetics were computed as the mean values of normalized ratings given by all workers/participants who evaluated an image,but separately for each of the three regions (CS-R1, CS-R2, and CS-R3).

The primary goal of this study was to check whether subjective image judgments collected in a crowdsourcing and lab environments were consistent. As a starting point, we checked the degree of inter-participant consistency. We expected a similar level of inter-participant consistency across Lab and CS experiments to indicate a comparable level of understanding of the task and of the underlying image construct to be rated (either recognizability or aesthetics). Such similarity would in turn allow for a more fair comparison of the MOS. We computed thus the standard deviation of the scores assigned to the same image by all the participants evaluating it. High values of standard deviation for an image indicate high disagreement across participants on the judgment of that image. Table 1 shows the mean values of the standard deviation across all images in the database.

Participants were quite consistent in rating both recognizability and aesthetics. Furthermore, the degree of consistency is rather stable across Lab and CS conditions, with an exception for the crowd-

Table 1: Average standard deviation of individual scores across all images and participants

	Lab	CS-R1	CS-R2	CS-R3
recognizability	0.6590	0.6213	0.6430	0.7716
aesthetics	0.8164	0.7061	0.7198	0.7902

sourcing data obtained from CS-R3 in the recognizability scoring task. The results suggest that across all experiments participants were able to score images with an acceptable and similar degree of consistency, which allows for further comparison of the Mean Opinion Scores gathered in the experiment.

As a second step, we checked whether the MOS obtained from the Lab and CS experiment were similarly distributed. One way to test this is to check whether the MOS values for lab and CS originate from two distributions with the same median. We tested this by means of the Kruskal-Wallis test (MOS for Lab, CS-R1, CS-R2 and CS-R3 were found not to be normally distributed, hence the need for a non-parametric test). The test revealed that neither the recognizability MOS (df = 3, chi = 7.49, p = 0.0578) nor the aesthetics MOS (df = 3, chi = 1.37, p = 0.7126) had significantly different medians across Lab, CS-R1, CS-R2 and CS-R3. This can be visually inspected in Figure 5, where Lab and CS MOS distributions are shown to be spread around a similar range, without systematic scoring differences (e.g., aesthetics always scored lower in the lab experiment). Systematic differences were also excluded by running a Mann-Whitney U-test among all possible distribution pairs, which in all cases gave negative response.

From Figure 5 it is also noticeable that the distributions of Lab and CS scores do not always nicely overlap. To quantify this, we checked to what extent Lab and CS MOS were linearly correlated. The results are reported in Table 2. Interestingly, Lab and Crowdsourcing MOS are quite well correlated for the recognizability construct (above 0.8 except for R3, for which the correlation drops significantly); MOS obtained from CS participants originating from different geographical areas are also acceptably consistent. Consistency across geographic areas is maintained for the aesthetic scoring; however, this is not the case for the correlation between lab and crowdsourcing scores, for which a visible drop occurs. Especially CS-R3 MOS have little predictive power for the Lab scores of the same images (correlation coefficient of 0.23).

Table 2: Linear correlation between LAB MOS and CS MOS

	recognizability			aesthetics		
	CS-R1	CS-R2	CS-R3	CS-R1	CS-R2	CS-R3
Lab	0.869	0.856	0.652	0.398	0.418	0.228
CS-R1	-	0.956	0.752	-	0.932	0.750
CS-R2	-	-	0.791	-	-	0.794

To further investigate this mismatch between CS and Lab results, we checked whether the CS data would preserve the insights on the measured construct emerged from the obtained Lab data. To test this, we looked into the relationship between recognizability and aesthetic Lab scores. These two quantities were found to be not correlated (correlation coefficient of 0.19, Table 3). When computing the same quantity for the three CS experiments, we found instead recognizability scores to be highly correlated to aesthetic scores (above 0.85 for all regions). Again, we found a discrepancy between the Lab results and the CS results, probably due to the difference in scoring aesthetics.

The main surprise of the crowdsourcing experiments is that while recognizability shows high correlation with lab-based scores, aesthetics doesn't. There are a few ways to explain this phenomenon. In principle, the discrepancy between Lab and CS results could be due to a different interpretation of the aesthetic quality scoring task in the CS settings. However, participants were found to be equally consistent when scoring in Lab or CS (see Table 1), which suggests an equal clarity of the tasks. Another possible explanation is that in the lab test, participants had to evaluate four quantities, whereas in the CS experiment they focused only on recognizability and aesthetics. This may have primed participants, favoring an unconscious association of the two quantities. A third explanation could be that some microworkers are careless in the way they complete their task. If they may try to answer the first question (on recognizability) honestly, for the second question (on aesthetics) they could just replicate the judgment expressed for recognizability, to minimize their effort. This reasoning is supported by the fact that recognizability and aesthetics MOS in crowdsourcing tests are highly correlated, whereas this is not the case in the Lab.

Table 3: Correlation between recognizability and aesthetic MOS in lab and crowdsourcing experiments

	Lab	CS-R1	CS-R2	CS-R3
correlation	0.196	0.869	0.896	0.888

6. CONCLUSION

In this paper, we compared lab and crowdsourcing-based evaluations of image aesthetic appeal and content recognizability. We found that crowdsourcing workers can be quite consistent with lab participants in scoring recognizability, whereas this is not the case for aesthetic appeal. Further analysis of the results suggests that crowdsourcing can be used for this type of subjective assessments, but the evaluation methodology needs to be designed carefully to avoid misinterpretations or cheating by the online workers. In particular, priming, confusion or cheating effects may arise from the evaluation of two different quantities in the same task.

As current results do not indicate a clear cause for the discrepancy between lab and crowdsourcing scores, we intend to conduct another round of crowdsourcing experiments to clarify the matter further. To investigate confusion and cheating effects, a possibility would be to have workers repeating the same campaign with a reversed order of the questions (first aesthetics and then recognizability) or just one of the two questions at a time.

ACKNOWLEDGEMENTS

This work was conducted in the framework of COST Action IC1003 – Qualinet with partial support from the EC funded Network of Excellence VideoSense. Special thanks to photographer Mark Dekker (http://gplus.to/markdekker) for allowing to use his photos in the experiments and to Christopher Pangerl who helped organizing and conducting crowdsourcing evaluations.

7. REFERENCES

[1] O. Axelsson. Towards a psychology of photography: Dimensions underlying aesthetic appeal of photographs. *Perceptual and Motor Skills*, 105(August 2002):411–434, 2007.

[2] K.-T. Chen, C.-J. Chang, C.-C. Wu, Y.-C. Chang, and C.-L. Lei. Quadrant of euphoria: a crowdsourcing platform for QoE assessment. *Network, IEEE*, 24(2):28–35, Mar. 2010.

[3] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Studying aesthetics in photographic images using a computational approach. In *Proceedings of the 9th European conference on Computer Vision - Volume Part III*, ECCV'06, pages 288–301, Berlin, Heidelberg, 2006. Springer-Verlag.

[4] A. Doan, R. Ramakrishnan, and A. Y. Halevy. Crowdsourcing systems on the world-wide web. *Commun. ACM*, 54(4):86–96, Apr. 2011.

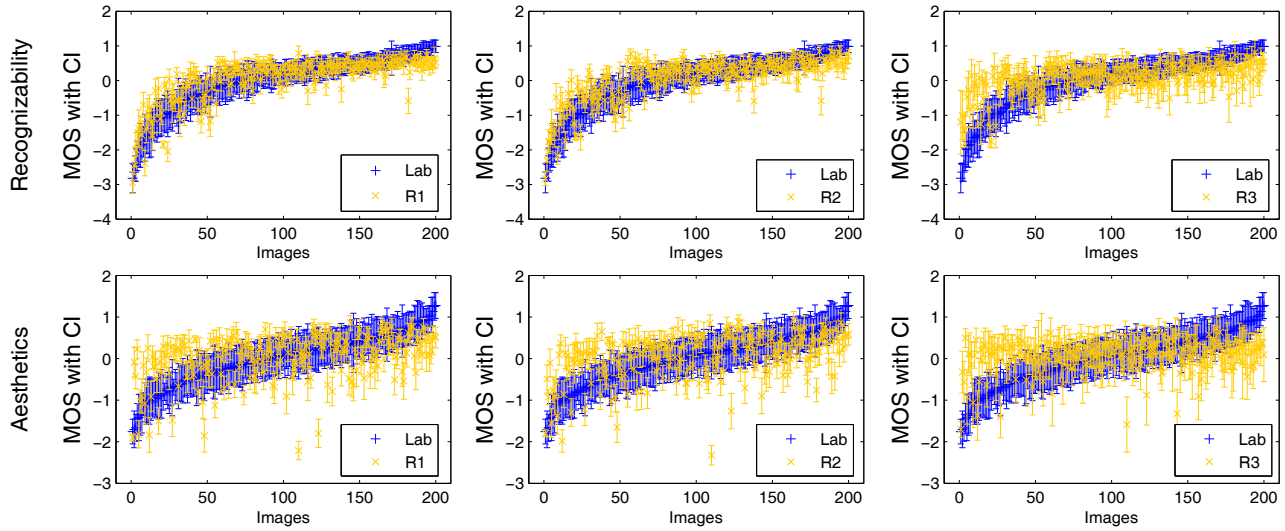

Figure 5: Comparison of Lab MOS with the CS MOS for the three scoring regions, for both recognizability (above) and aesthetics (below). Lab MOS (blue '+' markers) are sorted according to their magnitude.

[5] M. Hirth, T. Hoßfeld, and P. Tran-Gia. Anatomy of a Crowdsourcing Platform - Using the Example of Microworkers.com. In *Workshop on Future Internet and Next Generation Networks (FINGNet)*, pages 322–329, Seoul, Korea, June 2011.

[6] T. Hoßfeld, C. Keimel, M. Hirth, B. Gardlo, J. Habigt, K. Diepold, and P. Tran-Gia. CrowdTesting: A Novel Methodology for Subjective User Studies and QoE Evaluation. Technical Report 486, University of Würzburg, Feb. 2013.

[7] T. Hoßfeld, M. Seufert, M. Hirth, T. Zinner, P. Tran-Gia, and R. Schatz. Quantification of youtube qoe via crowdsourcing. In *Multimedia (ISM), 2011 IEEE International Symposium on*, pages 494–499, Dec. 2011.

[8] P. Isola, J. Xiao, A. Torralba, and A. Oliva. What makes an image memorable? In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 145–152. IEEE, 2011.

[9] D. Joshi, R. Datta, E. Fedorovskaya, Q.-T. Luong, J. Wang, J. Li, and J. Luo. Aesthetics and emotions in images. *Signal Processing Magazine, IEEE*, 28(5):94–115, 2011.

[10] D. Joshi, R. Datta, E. Fedorovskaya, Q.-T. Luong, J. Z. Wang, J. Li, and J. Luo. Aesthetics and emotions in images. *Signal Processing Magazine, IEEE*, 28(5):94–115, 2011.

[11] C. Keimel, J. Habigt, C. Horch, and K. Diepold. Qualitycrowd - a framework for crowd-based quality evaluation. In *Picture Coding Symposium (PCS), 2012*, pages 245–248, May 2012.

[12] A. Kittur, E. H. Chi, and B. Suh. Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 453–456, New York, NY, USA, 2008. ACM.

[13] P. Korshunov, S. Cai, and T. Ebrahimi. Crowdsourcing approach for evaluation of privacy filters in video surveillance. In *Proceedings of the ACM Multimedia 2012 Workshop on Crowdsourcing for Multimedia*, CrowdMM'12, pages 35–40, Nara, Japan, Oct. 2012.

[14] J. Lassalle, S. Member, L. Gros, T. Morineau, and G. Coppin. Impact of the content on subjective evaluation of audiovisual quality: What dimensions influence our perception? In *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, pages 1–6, 2012.

[15] P. Le Callet, S. Möller, and A. Perkis. Qualinet white paper on definitions of quality of experience (2012). European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003), Mar. 2013.

[16] C. Li and T. Chen. Aesthetic visual quality assessment of paintings. *Selected Topics in Signal Processing, IEEE journal*, 3(2):236–252, 2009.

[17] W. A. Mansilla, A. Perkis, and T. Ebrahimi. Implicit experiences as a determinant of perceptual quality and aesthetic appreciation. In *Proceedings of the 19th ACM international conference on Multimedia*, pages 153–162. ACM, 2011.

[18] S. Plous. *The Psychology of Judgment and Decision Making*. McGraw-Hill Series in Social Psychology. McGraw-Hill Education, 1993.

[19] Recommendation. ITU-R BT.500-13. Methodology for the subjective assessment of the quality of television pictures. International Telecommunication Union, Geneva, Switzerland, 2012.

[20] J. A. Redi. Visual quality beyond artifact visibility. *Proc. SPIE*, 8651:86510N–86510N–11, Mar. 2013.

[21] S. Suri, D. Goldstein, and W. Mason. Honesty in an online labor market. In *Human Computation: Papers from the 2011 AAAI Workshop*, pages 61–66, 2011.

[22] W. Tschacher, S. Greenwood, V. Kirchberg, S. Wintzerith, K. van den Berg, and M. Tröndle. Physiological correlates of aesthetic perception of artworks in a museum. *Psychology of Aesthetics, Creativity, and the Arts*, 6(1):96–103, 2012.

[23] L.-k. Wong and K.-l. Low. Saliency-enhanced image aesthetics class prediction. In *2009 16th IEEE International Conference on Image Processing (ICIP)*, pages 997–1000, Cairo, Nov. 2009. Ieee.