Worker Qualifications for Image-Aesthetic-Assessment Tasks in Crowdsourcing

Yudai Kato,[1] **Marie Katsurai,**[1] **Keishi Tajima**[2]

[1] Doshisha University
[2] Kyoto University
{yudai.kato, katsurai}@mm.dohisha.ac.jp, tajima@i.kyoto-u.ac.jp

Abstract

Image aesthetic assessment has been a trending topic in the research field of multimedia information retrieval. Crowdsourcing can be an efficient approach for collecting manual assessment results to construct an image dataset associated with aesthetic scores. This study explores a strategy for setting worker qualifications to participate in an image-aesthetic-assessment task. Our current experiments based on the AVA dataset indicated that the target subjective task requires highly experienced workers to produce ratings similar to photographers' ones.

Introduction

Automatic assessment of images' aesthetic quality has been actively studied to facilitate several applications, such as image retrieval and editing. The traditional approach used handcrafted features (e.g., brightness and composition) to train an aesthetics classifier. Recent studies demonstrated that deep learning based on images and their subjective ratings could automatically extract visual features related to beauty; for example, convolutional neural networks have yielded significant performance improvements over conventional visual features (Deng, Loy, and Tang 2017; Zhang, Miao, and Yu 2021). These supervised learning approaches usually require a large amount of data associated with manually assigned aesthetic information. Since training using unreliable, biased, or small data will affect prediction performance, we should consider how to efficiently collect aesthetic scores that are carefully assessed.

Crowdsourcing is an efficient approach to collecting a large number of answers in a short time and at a low cost for constructing an annotated dataset. Several studies have illustrated the effectiveness of crowdsourcing even for subjective evaluation (Redi et al. 2013). Related studies, such as those involving image sentiment or emotion analysis, have also employed crowdsourcing to construct a labeled dataset (Katsurai and Satoh 2016; Korovina et al. 2018). Crowdworkers include spammers who complete many tasks with little effort to maximize their rewards and inexperienced workers who do not understand the task properly. Prior studies have eliminated the data of such workers using postprocessing methods, such as outlier detection and answer aggregation (e.g., majority voting). However, only a few studies have been conducted on preventing the participation of such workers before ordering tasks. Thus, in this study, we explore a strategy for setting worker eligibility requirements to stabilize the quality of the results for image aesthetic assessments, which is a subjective task. Specifically, using an existing dataset for visual aesthetic analysis as the ground truth, we quantitatively evaluated the performance of each qualification setting.

Dataset and Platform

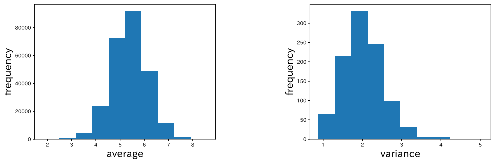
As a crowdsourcing platform, we used Amazon Mechanical Turk (MTurk)[1]. MTurk provides functions to improve the quality of tasks, such as the ability to select eligible workers and reject submitted tasks based on the quality of answers. Some advanced filtering conditions require an additional fee. Our current study focuses on the qualification conditions that workers must meet to submit reliable results for the image-aesthetic-assessment task.

Our experiments used the AVA dataset (Murray, Marchesotti, and Perronnin 2012), which comprises 255,530 images collected from the DPChallenge website[2]. DPChallenge is a contest in which professional and amateur photographers submit photos under a theme called "challenge" and receive aesthetic scores ranging from 1 to 10 from other participants. The contest evaluators are good judges of visual aesthetics, so these scores are considered relatively reliable. Each image in the AVA dataset has 78 to 549 ratings, with an average of 210. Several previous studies on the AVA dataset have calculated the average of the ratings to produce a single aesthetic score per image (Deng, Loy, and Tang 2017).

Since it is cost-prohibitive to collect the ratings for all of the images in the AVA dataset, we first randomly chose 20 challenges associated with more than 100 images before arbitrarily extracting 50 images from each challenge. The total number of test images used in our experiment was 1,000. Figure 1 shows the histograms of the averages and variances of the ratings over 1,000 test images. Most of the images have average ratings between 5.4 and 6.0, and opinions for some images were polarized among photographers.

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

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



(a) Averages' distribution. (b) Variances' distribution.

Figure 1: Histograms of the averages and variances of the ratings over 1,000 test images from the AVA dataset.

Experimental Settings

Design of evaluation screens. In our crowdsourcing experiment, a single task comprised the aesthetic evaluation of 50 test images, and this was regarded as a human intelligent task (HIT) in MTurk terminology. To create an evaluation situation that is similar to the one used for the AVA dataset, the title and a short description of the challenge corresponding to the test images were presented to the worker. The workers were then instructed to perform the aesthetics evaluation while taking the challenge's concept into account. The evaluation was made on a scale of 1 (non-aesthetic) to 10 (aesthetic).

Qualifications. Table 1 lists the five conditions for worker qualification, which are combinations of the number of tasks approved among the previous submissions of a worker and their percentage, which was represented by the approval rate. The higher the approvals a worker has, the more familiar they are with MTurk. Workers with low approval rates are likely to be spammers who submitted unacceptable results to other users' tasks. Condition 1 is often used in crowdsourcing-based experiments (Robinson et al. 2019). Further, Condition 2 allows the inclusion of workers who are more likely to be spammers, whle Condition 3 will allow the inclusion of more new workers. Condition 4 severely restricts eligibility to recruit many high-quality workers who are familiar with MTurk. Finally, Condition 5 corresponds to no restriction with regard to worker qualifications.

Language and location. Our task instructions were written in English. It was reported in Goodman, Cryder, and Cheema (2013) that the quality of the results for tasks that are in English deteriorated if the workers were not limited to those living in the U.S. Therefore, we selected the U.S. as the workers' location.

Orders and rewards. We ordered multiple HITs by changing the time and day so that each worker's activity time does not affect the task results of each qualification condition. The workers were paid $0.25 for each completed HIT. In the experiment, we allowed individual workers to participate in only a single qualification condition's HITs. Each test image received 25 ratings for each qualification condition.

Result

We computed the average and variance of the ratings obtained from multiple workers for each test image. Using all the test images, we calculated the correlation coefficient and mean absolute error (MAE) between the MTurk and the AVA ratings in terms of averages and variances, respectively.

Table 1: The five worker qualification conditions compared in this study. None means no limit.

Condition	# of approved tasks	Approval rate
1	over 100	over 95%
2	None	under 95%
3	under 100	None
4	over 5,000	over 98%
5	None	None

Table 2: Comparison of relationships with AVA ratings in terms of averages and variances for different qualification conditions.

	Average		Variance	
Condition	correlation	MAE	correlation	MAE
1	0.32	1.06	-0.06	2.38
2	0.29	1.39	0.12	2.46
3	0.36	1.18	0.03	3.29
4	**0.43**	**0.81**	-0.02	3.17
5	0.29	1.55	0.12	2.39

Table 2 compares the results for different worker qualification conditions. As expected, the average scores produced by workers that satisfied Condition 4 were the closest to the AVA dataset. Condition 1 was a relatively severe restriction; but interestingly, it did not lead to any significant differences from the other conditions, except for Condition 4. This result implies that the commonly used qualifications are insufficient for subjective tasks that require expertise. We also found that Condition 3 yielded better results than Condition 2 even though it did not consider the approval rate of workers. Further, we observed almost no linear relationship in the variances between the MTurk ratings and the AVA ratings for the test images; this indicates that the rating consistency differed among these two sets of ratings. This might be because some of the test images were edgy or subject to interpretation; photographers' opinions on such images often varied (Murray, Marchesotti, and Perronnin 2012), and a layman might not be able to identify any aesthetics. We will further investigate this issue in future studies.

Conclusion and Future Work

Using the AVA dataset's ratings, we investigated how different MTurk worker qualifications produced different results for image-aesthetic-assessment tasks. The results of our current experiments demonstrated the effectiveness of limiting eligibility to only those workers who had been approved for thousand tasks and had a high approval rate of over 98%. We also found that the standard criterion, which was often used in related studies, was insufficient for the target subjective task; detailed instructions or training to guide the workers might be necessary. These results will be informative for other subjective tasks that require understanding contexts, semantics, and sentiments.

We will conduct similar experiments for other subjective tasks. Our future work will also include an analysis of the relationship between crowdsourcing time and worker qualification conditions.

References

Deng, Y.; Loy, C. C.; and Tang, X. 2017. Image Aesthetic Assessment: An experimental survey. *IEEE Signal Processing Magazine*, 34(4): 80–106.

Goodman, J. K.; Cryder, C. E.; and Cheema, A. 2013. Data Collection in a Flat World: The Strengths and Weaknesses of Mechanical Turk Samples. *Journal of Behavioral Decision Making*, 26(3): 213–224.

Katsurai, M.; and Satoh, S. 2016. Image sentiment analysis using latent correlations among visual, textual, and sentiment views. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2837–2841.

Korovina, O.; Casati, F.; Nielek, R.; Baez, M.; and Berestneva, O. 2018. Investigating Crowdsourcing as a Method to Collect Emotion Labels for Images. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–6.

Murray, N.; Marchesotti, L.; and Perronnin, F. 2012. AVA: A large-scale database for aesthetic visual analysis. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2408–2415.

Redi, J. A.; Hoßfeld, T.; Korshunov, P.; Mazza, F.; Povoa, I.; and Keimel, C. 2013. Crowdsourcing-Based Multimedia Subjective Evaluations: A Case Study on Image Recognizability and Aesthetic Appeal. In *Proceedings of the 2nd ACM International Workshop on Crowdsourcing for Multimedia*, 29–34.

Robinson, J.; Rosenzweig, C.; Moss, A. J.; and Litman, L. 2019. Tapped out or barely tapped? Recommendations for how to harness the vast and largely unused potential of the Mechanical Turk participant pool. *PLoS ONE*, 14(12).

Zhang, J.; Miao, Y.; and Yu, J. 2021. A Comprehensive Survey on Computational Aesthetic Evaluation of Visual Art Images: Metrics and Challenges. *IEEE Access*, 9: 77164–77187.