Faster Image Rating using a CAPTCHA Inspired Task

Siyi Fan

1. Introduction

With the development of high capacity supervised learning algorithms, there has been notable progress in the performance of image detection and classification. Lagging behind the growth in computational modeling, the available image datasets are becoming outdated in terms of size and density (Yu et al., 2016). The primary limitation is the lack of data. The input data with ratings provided by human raters would be especially useful for training and testing purposes since they can provide insights into whether the local processing is correctly modeled or whether further adjustments are warranted. Moreover, training images are spread across many categories, which results in the density of image examples in current datasets being quite low (Yu et al., 2016).

The bottleneck to building a supervised image dataset with high density is the high cost of manual labeling. Traditionally, manual image ratings have only been carried out on Likert scales rating a single image or ranking scales comparing two images at the same time, which involves an enormous time commitment for both raters and experimenters to create the image database.

In the present study, we have begun to circumvent this bottleneck and address this issue by proposing CAPTCHA-style image rating method. The objective is to validate CAPTCHA-style image rating method against the conventional image rating methods. Participants were presented with twelve images simultaneously each trial and asked to click on "top 4" images in terms of preference. **Our research question is: Will CAPTCHA-style image rating would be faster than conventional image rating methods?** Specifically, we have two hypotheses:

Hypothesis1: CAPTCHA-style image rating will be faster than conventional image rating. That is, participants will spend less time rating the same amount of images using CAPTCHA-style than using Likert scale.

Hypothesis2: CAPTCHA-style image ratings will be positively correlated with single-image style ratings previously collected by Likert scale ratings of naturalness, preference, and order-disorder dimensions for each image.

2. Literature Review

2.1 Image Rating Tasks

While judging preferred images is a subjective task, it has been an area of active study for a long time and substantial progress has been made in evaluating those images.

2.1.1 Single-image Rating With Likert Scale.

The first line of image rating task is to display a single image on the screen at a time and have participants rate that image using a Likert scale. Kaplan et al. (1972) studied whether images of natural or urban environments will influence preference by presenting slides depicting non-spectacular, relatively local places to participants. For each slide, the participants were asked to rate the image on complexity and preference using a 5-point rating scale ranging from "not at all" to "a great deal". Berman et al. (2008) compared the restorative effects on cognitive functioning of interactions with natural versus urban environments. Picture viewing lasted approximately 10 min, during which participants rated on a scale of 1 to 3 how much they liked each picture. The Likert scale is a commonly used assessment tool. It evaluates how participants feel about stimuli and within a specific classification. One con of using a Likert scale is that people may respond too quickly, simply assigning an average score to each of the items without fully considering the question.

2.1.2 Pairwise Comparison Rating With Ranking Scales.

An alternative method is a ranking scale. It allows participants to compare individual elements to each other and rank the choices in order of preference. A common ranking scale approach in

image evaluation is pairwise comparison, where an individual compares two images side by side and chooses which image they prefer. A downside of pairwise comparison is that people are forced to make a choice between two items, when they may actually perceive them equally. Gur et al. (1997) compared five-category ordinal ratings of relative quality to a two-alternative forced-choice approach by presenting pairs of laser-printed images (labelled as "A" and "B") to participants. For the five-category ordinal scale, participants were asked to rank order the relative differences between the two images ranging from "Image 'A' is much better than image 'B'" to "Image 'A' is much worse than image 'B'". For the two-alternative forced-choice scale, participants were forced to choose one image in each pair ("A" or "B") as the better image. The researchers found a significantly lower ability to detect small differences in perceived image quality when using a five-category ordinal scale and participants are highly sensitive to image differences during a side-by-side review. Therefore, participant behaviors can be affected by the image-rating scales used. Phelps et al. (2015) compared the accuracy of pairwise comparisons with that of a Likert scale in biomedical image assessment. Using a pairwise comparison test, two images were presented side-by-side on one slide and subjects were required to determine which one is sharper. Using a Likert scale, a single image was displayed on each slide and subjects assigned a sharpness score between 1 (least sharp) and 10 (sharpest) for that image. The results showed that pairwise comparison method was easier because it did not require the subjects to remember the upper and lower limits of the Likert scale.

In summary, these classic image rating tasks have been applied successfully in order to answer psychological questions so that fifty years later, researchers are still using single-image and pairwise comparison formats to gather image ratings. However, these tasks have a major limitation. Participants are required to spend a lot of time and energy to rate the wide array of

images. To this end, the current study attempts to fill this void by designing a new image rating method to improve the efficiency of human-in-the-loop annotation and make the rating process more convenient for participants.

2.3 Image Rating CAPTCHA

In this paper, we introduce a new image rating format using CAPTCHA with humans who view 12 images at the same time and click on 'top 4' preferred choices. CAPTCHA (Completely Automated Public Turing Test to Tell Computers and Humans Apart) is a program that generates and grades tests that most humans can pass, but current computer programs fail (Von et al., 2003). Such a program has many applications for website security, including online polls, email services, search engine bots, worms and spam, and preventing dictionary attacks. Image-based CAPTCHA requires users to perform an image recognition task, such as: click all of the images that meet a requirement. By using the similar way for our study, participants were asked to view 12 images simultaneously displayed during each trial and click on 4 images they preferred most. For each image, one click assigned one point to it.

3. Experiment

3.1 Methods

3.1.1 Participants.

The target sample size was 120 participants, the optimal number of participants in the initial sample based on the results of a simulation. Volunteers participants were recruited using TurkPrime to complete the online study via the online labor market Amazon Mechanical Turk (AMT), in which workers perform human intelligence tasks (HITs) for requesters.

3.1.2 Stimuli and Tasks.

Materials. The 1030 images utilized in the current study were real-world environmental scenes selected from the Scene Understanding (SUN) image database (http://vision.princeton.edu/projects/2010/SUN/; Xiao, Hays, Ehinger, Oliva, & Torralba, 2010), which contains 899 categories and 130,519 images, diverse in semantic content (e.g., indoor, urban, and nature). In the online experiment, the 1030 full-color 120*160 pixel photographs of real-world, indoor and outdoor scenes were used.

Procedure. Participants viewed 1030 scene images in total during 87 trials, including 14 attention check trials. On each trial, 12 different scene images arranged in 3 discrete rows with random item placement were presented on a plain white background. Twelve was the optimal set of images that could be displayed simultaneously to the participants without overcrowding the display. After scanning 12 images on the screen, participants were instructed to click on the "top 4" images in terms of preference. Once an image got clicked on, it was assigned 1 point. Otherwise, it was assigned 0 points. The prompt "Select 4 images you like and click submit" was written above 12 images on each trial. For the attention check trials, subjects were told to drag the blurry image to the trash can before selecting 4 images they preferred. Finally, participants were given a random code that they had to submit an AMT to verify their completion of the experiment.

3.1.3 Data and Code Availability.

The data analysis was performed using custom MATLAB scripts, which were available at https://github.com/siyiii/fp-siyiii. All statistical tests were performed using MATLAB R2016b. The *fitglme* function was used to perform GLMM analyses, the *corrcoef* function was used to calculate the effect size (95% CI) of Pearson's correlations, and the *polyfit* function was used to calculate the 95% confidence band of a regression.

3.2 Results (still working on it…)

Simulation. To determine the required number of subjects (P*) and ideal number of images (N*) subjects should click on of the 12 images displayed on the screen at a time, the following computer simulation was conducted. First, we calculated the inter-rater correlation for the actual dataset, which was created by Environmental Neuroscience Lab at the University of Chicago (https://enl.uchicago.edu/). Next, we generated a simulated dataset with P (P=30, 60, 120) participants rating 1030 images and computed the inter-rater correlation for this simulated dataset. To mimic the idiosyncrasies of actual preference rating, "sbjweight" was added to the simulated preference rating for each image. We modified the "sbjweight" so that the inter-rater correlation for this simulated dataset was close to the inter-rater correlation in the actual dataset (simulated preference = actual preference + sbjweight * randn(size(actual preference))). Then, we shuffled the order of all images order and sorted N (N=2, 3, 4) images out of 12 images with the highest simulated preference ratings (simulated preference) displayed on each trial. Based on the simulated preference ratings, those selected N images would be clicked on in CAPTCHA-style. Therefore, we assigned point 1 to each sorted image and point 0 to all unsorted images, generating a new dataset with 1030 preference rating. Finally, we calculated (Table 1) and plotted (Figure 1) the correlations between the preference rating in the new dataset and actual preference rating. The correlation increased as the number of subjects increased, and the number of images clicked on each trial increased, reaching the highest value with P = 120 and N = 4. In this way, we could approximate the best parameters (P* = 120, N* = 4) and decrease sampling error.

Table 1.

Correlations of Simulated Preference Rating and Actual Preference Rating

	N = 2	N = 3	N = 4
P = 30	0.88	0.92	0.95

P = 60	0.90	0.94	0.96
P = 120	0.92	0.95	0.97

Figure 1. The correlation between actual preference rating and simulated preference rating

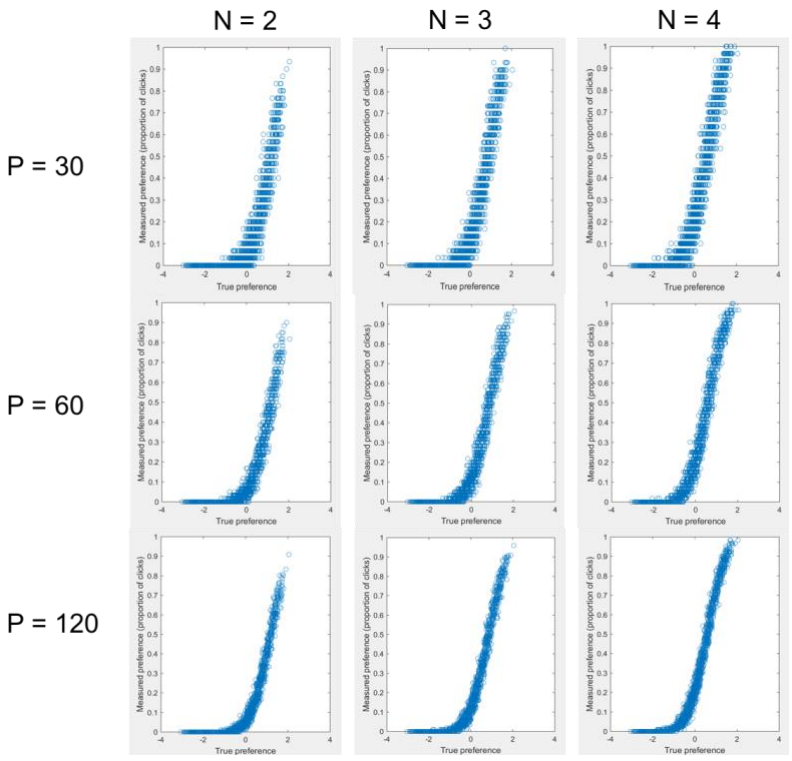

Figure 1. The correlation between actual preference rating and simulated preference rating. The true preference indicates the actual preference rating. The measured preference indicates the average simulated ratings across all participants P.

Descriptives. Will finish after data collection…

Correlation. Will finish after data collection…

Position bias. (Will finish after data collection…)The influence of position and trial bias was investigated using a generalized linear mixed-effects model (GLMM) in MATLAB, with the binary dependent variable "Click" or "Not Click" predicted from (1) Image and (2) image position during each trial. In this analysis, the fixed effects were image (1, 2, …, 1030) and image position

(1, 2, …, 12). Also, random intercepts for participants were included in the model. … GLMMs incorporate individual trial-level data and are therefore preferable over traditional ANOVA models that aggregate trial-level data and cannot account for stimuli differences.

References

Berlyne, D. E. (1963). Complexity and incongruity variables as determinants of exploratory choice and evaluative ratings. *Canadian Journal of Psychology/Revue canadienne de psychologie*, *17*(3), 274.

Berman, M. G., Jonides, J., & Kaplan, S. (2008). The cognitive benefits of interacting with nature. *Psychological science*, *19*(12), 1207-1212.

Berman, M. G., Hout, M. C., Kardan, O., Hunter, M. R., Yourganov, G., Henderson, J. M., ... & Jonides, J. (2014). The perception of naturalness correlates with low-level visual features of environmental scenes. *PloS one*, *9*(12), e114572.

Day, H. (1968). The importance of symmetry and complexity in the evaluation of complexity, interest and pleasingness. *Psychonomic Science, 10*(10), 339-340.

Gur, D., Rubin, D. A., Kart, B. H., Peterson, A. M., Fuhrman, C. R., Rockette, H. E., & King, J. L. (1997). Forced choice and ordinal discrete rating assessment of image quality: a comparison. *Journal of digital imaging*, *10*(3), 103-107.

Kaplan, S., Kaplan, R., & Wendt, J. S. (1972). Rated preference and complexity for natural and urban visual material. *Perception & Psychophysics, 12*(4), 354-356.

Kardan, O., Demiralp, E., Hout, M. C., Hunter, M. R., Karimi, H., Hanayik, T., ... & Berman, M. G. (2015). Is the preference of natural versus man-made scenes driven by bottom–up processing of the visual features of nature?. *Frontiers in psychology*, *6*, 471.

Kardan, O., Henderson, J. M., Yourganov, G., & Berman, M. G. (2016). Observers' cognitive states modulate how visual inputs relate to gaze control. *Journal of Experimental Psychology: Human Perception and Performance, 42*(9), 1429.

Khosla, A., Raju, A. S., Torralba, A., & Oliva, A. (2015). Understanding and predicting

image memorability at a large scale. In Proceedings of the IEEE International Conference on Computer Vision (pp. 2390-2398).

Kotabe, H., Kardan, O., & Berman, M. G. (2016). Can the High-Level Semantics of a Scene be Preserved in the Low-Level Visual Features of that Scene? A Study of Disorder and Naturalness. In CogSci.

Kotabe, H. P., Kardan, O., & Berman, M. G. (2016). The order of disorder: Deconstructing visual disorder and its effect on rule-breaking. *Journal of Experimental Psychology: General*, *145*(12), 1713.

Kotabe, H. P., Kardan, O., & Berman, M. G. (2017). The nature-disorder paradox: A perceptual study on how nature is disorderly yet aesthetically preferred. *Journal of Experimental Psychology: General*, *146*(8), 1126.

Martindale, C., Moore, K., & Borkum, J. (1990). Aesthetic preference: Anomalous findings for Berlyne's psychobiological theory. *The American Journal of Psychology*, 53-80.

Phelps, A. S., Naeger, D. M., Courtier, J. L., Lambert, J. W., Marcovici, P. A., Villanueva Meyer, J. E., & MacKenzie, J. D. (2015). Pairwise comparison versus Likert scale for biomedical image assessment. *American Journal of Roentgenology*, *204*(1), 8-14.

Schroeder, H. W., & Anderson, L. M. (1984). Perception of personal safety in urban recreation sites. *Journal of leisure research*, *16*(2), 178-194.

Seresinhe, C. I., Preis, T., & Moat, H. S. (2017). Using deep learning to quantify the beauty of outdoor places. Royal Society open science, 4(7), 170170.

Seresinhe, C. I., Moat, H. S., & Preis, T. (2018). Quantifying scenic areas using crowdsourced data. Environment and Planning B: Urban Analytics and City Science,

45(3), 567-582.

Thayer, R. L., & Atwood, B. G. (1978). Plants, complexity, and pleasure in urban and suburban environments. *Environmental Psychology and Nonverbal Behavior, 3*(2), 67-76.

Torralba, A. (2009). How many pixels make an image?. *Visual neuroscience*, *26*(1), 123-131.

Von Ahn, L., Blum, M., Hopper, N. J., & Langford, J. (2003, May). CAPTCHA: Using hard AI problems for security. In *International Conference on the Theory and Applications of Cryptographic Techniques* (pp. 294-311). Springer, Berlin, Heidelberg.

Wohlwill, J. F. (1968). Amount of stimulus exploration and preference as differential functions of stimulus complexity. *Perception & Psychophysics*, *4*(5), 307-312.

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., & Torralba, A. (2010, June). Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 3485-3492). IEEE.

Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., & Xiao, J. (2015). Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*.