Quality Assessment of Crowdwork via Eye Gaze: Towards Adaptive Personalized Crowdsourcing

Md. Rabiul Islam[1(✉)], Shun Nawa[1], Andrew Vargo[1], Motoi Iwata[1], Masaki Matsubara[2], Atsuyuki Morishima[2], and Koichi Kise[1]

[1] Osaka Prefecture University, Sakai, Japan
dd104006@edu.osakafu-u.ac.jp, {nawa,awv}@m.cs.osakafu-u.ac.jp, {iwata,kise}@cs.osakafu-u.ac.jp
[2] University of Tsukuba, Tsukuba, Japan
{masaki,mori}@slis.tsukuba.ac.jp

Abstract. A significant challenge for creating efficient and fair crowdsourcing platforms is in rapid assessment of the quality of crowdwork. If a crowdworker lacks the skill, motivation, or understanding to provide adequate quality task completion, this reduces the efficacy of a platform. While this would seem like only a problem for task providers, the reality is that the burden of this problem is increasingly leveraged on crowdworkers. For example, task providers may not pay crowdworkers for their work after the evaluation of the task results has been completed. In this paper, we propose methods for quickly evaluating the quality of crowdwork using eye gaze information by estimating the correct answer rate. We find that the method with features generated by self-supervised learning (SSL) provides the most efficient result with a mean absolute error of 0.09. The results exhibit the potential of using eye gaze information to facilitate adaptive personalized crowdsourcing platforms.

Keywords: Crowdsourcing · Eye gaze · Self-supervised learning · Machine learning

1 Introduction

Crowdsourcing is widely employed as a way to achieve tasks that can be more efficiently done by human intelligence. Starting from simple labeling microtasks, researchers have broadened the scope of crowdsourcing to include tasks that require complex input [2] or creativity [5,21]. Crowdsourcing has long been discussed as a polemic topic in that research often focuses on how crowdworkers are exploited by task-providers and platforms [18,23] or focuses on how to improve task efficiency [12] and mitigate spam crowdworkers [17]. To make a fruitful society, it is necessary to prepare crowdsourcing environments that are beneficial to not only task-providers but to crowdworkers as well.

Published by Springer Nature Switzerland AG 2021
C. Ardito et al. (Eds.): INTERACT 2021, LNCS 12933, pp. 104–113, 2021.
https://doi.org/10.1007/978-3-030-85616-8_8

A key to realize such environments is to introduce more precise quality assessment methods. Currently, the assessment is primarily to evaluate "crowdworkers" to distinguish high-skill workers from low-skill and spam ones [6,15]. Once a crowdworker is classified as low-skill or spam, it is not possible to receive rewards from the work. Although spam workers deserve to receive nothing, it is not fair for low-skill workers; they should receive rewards in response to the quality of output, e.g., the number of correctly answered tasks. Generally speaking, the performance of crowdworkers depends on many factors, including the tasks themselves, personal skills and psychosomatic aspects of workers' behaviors, and their computing and living environments [4,30]. Thus, it is more reasonable and fairer to assess the quality of not crowdworkers but each piece of crowdwork. Moreover, evaluation of crowdwork allows us to adaptively change task allocation, if low performance is due to the currently assigned task. In other words, quality assessment of crowdwork is mandatory to realize adaptive personalized crowdsourcing.

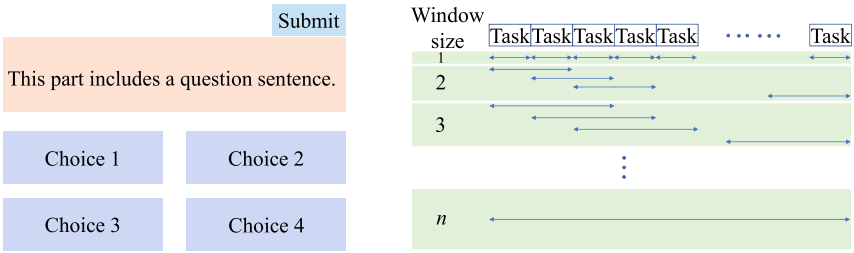
In this paper, we employ crowdworkers' eye gaze for quality assessment on tasks. It has been known that the eye gaze is influenced by confidence on an answer to a task [27], and the confidence is correlated with the correctness of the answer [10]. Thus, we can estimate the quality of crowdwork by analyzing the eye gaze. We use multiple-choice questions (MCQs) as the task and propose two different ways of feature extraction from the eye gaze: handcrafted and self-supervised learning (SSL). The findings are promising. For a large number of the tasks performed, the proposed methods, especially SSL, can estimate the performance with roughly half the error-rate as compared to a baseline estimator.

2 Related Work

Quality control has been a central issue for crowdsourcing. Quality in crowdsourcing is classified into three categories: quality model, quality assessment, and quality assurance [4]. In this work, we are looking at quality assessment. In particular, we limit our focus to computer-based methods that do not rely on evaluation by humans.

A fundamental goal of quality assessment is to identify spam crowdworkers or malicious behaviors for removal [6]. A simple way of conducting a quality assessment is using ground truth where, with known answers, we can estimate the quality of work by measuring the accuracy of the tasks [14]. However, preparing a ground truth for enough tasks is usually expensive. Another way is to evaluate the agreement in output across crowdworkers. This is also expensive because enough answers must be collected for each task. A more sophisticated way is based on crowdworkers' behavior called "fingerprinting," such as mouse usage and screen scrolling [24]. More advanced methods include ranking crowdworkers using a measure of spammers [22]. Besides, researchers have proposed time-series model [13] and cognitive abilities based model [9] to estimate quality.

Another vital point is the use of computational models. In addition to simple matching with the ground truth, game theory [20], probabilistic modeling and the EM algorithm [22], the log-normal model [28], and traditional machine learning



(a) The crowdworker is asked to select the correct choice

(b) We employed a window to cover a sequence of performed tasks

Fig. 1. (a) MCQ format and (b) window format employed in our method.

methods such as decision trees [16] have been used. To the best of our knowledge, deep learning has not yet been well employed as a tool for crowdsourcing since it generally requires a large number of task outputs with ground truth.

The technology called SSL [1,19] is a paradigm to cope with the lack of labeled data issue (details in Subsect. 3.2). The SSL has been applied in many domains [7,29], and recently to the human activity recognition task with sensor data [8,25]. In this paper, we attempt to apply the SSL technology developed to analyze eye gaze data [11] for quality assessment of crowdwork. It is important to analyze eye gaze data since it conveys vital behavioral [28], attention [26] and confidence [27] information about the user.

3 Proposed Methods

In this work, we propose methods for the quality assessment of crowdwork by estimating the correct answer rate using eye gaze information. Crowdwork involves numerous tasks; answering MCQs, labeling pictures, solving math equations, and similar. Among all, we chose the answering MCQs since MCQs present the correct and incorrect answers. Figure 1a shows the MCQs format. The eye gaze is recorded while answering MCQs on the computer screen by an eye-tracker. Finally, we propose two methods; the first one is based on handcrafted features, and the second one is based on features generated by using the SSL, where the latter eliminates the handcrafted feature engineering.

3.1 Method with Handcrafted Features

This method consists of two stages: feature extraction and estimation of the correct answer rate.

Feature Extraction. The reading behavior is characterized by a sequence of fixations and saccades [27]. Fixations appear when the gaze pauses in a point, and saccades correspond to the jumps of the gaze between fixations. We extract

Table 1. List of the selected features.

Method	No.	Feature
Handcrafted	$f1$	Number of fixations on the question
	$f2$	Number of fixations on choices
	$f3$	Number of saccades on the question
	$f4$	Number of saccades between the question and choices
	$f5$	Answering time
	$f6$	Self-confidence on the answer
Automatic generation	$f256$	Feature vector generated by SSL

features for the eye gaze data for which we want to estimate the correct answer rate by detecting fixations applying the Buscher algorithm [3] and then extract other features. Table 1 shows the six ($f1$ to $f6$) selected and extracted features.

We employ a window to cover a number of sequential tasks performed, as shown in Fig. 1b, where the number of tasks included in the window is a parameter ranging from 1 to n (all tasks). We slide the window with the step of one task. Features for describing a window is just a concatenation of features from each task. For example, let f_{ij} be a feature j from the task i. Then, for example, the feature vector representing the window of size 2 including the task (i) and $(i+1)$ is $(f_{(i)1}, ..., f_{(i)k}, \quad f_{(i+1)1}, ..., f_{(i+1)k})$.

Estimation. The feature vectors representing windows are then used to estimate the correct answer rate by employing the Support Vector Regression (SVR).

3.2 Method with Features Generated by Self-supervised Learning

This method also consists of two stages: feature extraction and estimation of the correct answer rate.

Feature Extraction. We propose an SSL method for automatic feature generation, as shown in Fig. 2, that consists of self-supervised pre-training, correctness estimation, and feature extraction stages. To handle eye gaze for this purpose, it is problematic that the size of eye gaze data varies from MCQ to MCQ. To cope with this issue, we convert the eye gaze data by plotting graphically, as shown in Fig. 3a. The red circles are eye gaze points and the x-axis belongs to the horizontal direction of Fig. 3a. The details are as follows.

The first stage is self-supervised pre-training, upper part of Fig. 2, by solving the pretext task, automatically applied to a large collection of unlabeled data. As shown in Fig. 3b to 3d, we consider three image transformations; reflection about y-axis and reflection about x-axis and 45° anti-clockwise rotation to format the pretext task. For each eye gaze image, we randomly applied one transformation or not transformed and solved a four-class classification task.

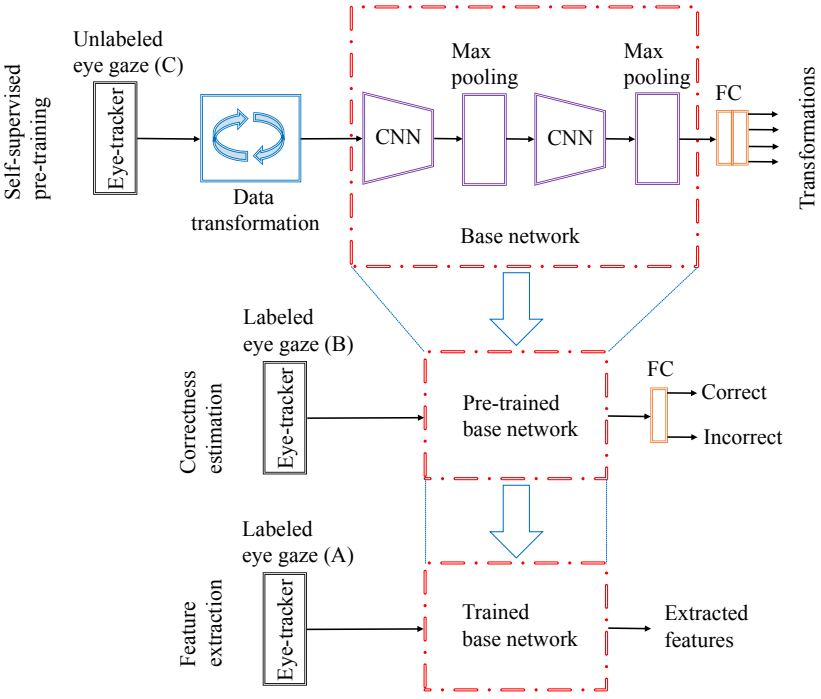


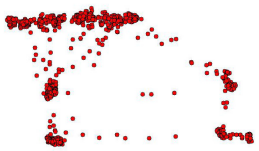
Fig. 2. The proposed method for automatic features generation using SSL. (Color figure online)

The red box in the upper part of Fig. 2 shows the base network, including two CNN blocks and a 2D max-pooling layer after each CNN block. Each CNN block consists of two 2D CNN layers. For the first and second CNN blocks, layers have 8 and 16 units, respectively. The kernel size of CNN layers is 3×3. Finally, we add a classifier consisting of two Fully Connected (FC) layers with 36 units for both. We use ReLU, softmax function, and SGD as the activation function, output layer, and optimizer, respectively. The input image size is $64 \times 64 \times 3$.

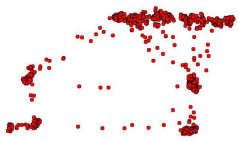
The second stage, middle part of Fig. 2, is the correctness estimation done by replacing the FC layers of the pre-trained network with an FC layer with 64 units and fine-tuning by using a labeled eye gaze dataset. The estimation of correctness is a binary classification; the answer is correct or incorrect.

In the third stage, lower part of Fig. 2, we extract features by collecting output at the end of the base network for the dataset we want to estimate the correct answer rate. The final feature vector length is 256 for each task, denoted as $f256$ in Table 1. We format windows in the way described in Subsect. 3.1.

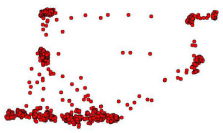
Estimation. The feature vectors representing windows are then used to estimate the correct answer rate using SVR in the same way as described in Subsect. 3.1.



(a) Eye gaze image



(b) Reflection about y-axis



(c) Reflection about x-axis

(d) Rotation

Fig. 3. Eye gaze images, (a) actual eye gaze image with no transformation applied, and (b) to (d) are transformed copies of (a).

4 Datasets

We use three datasets: labeled dataset A, labeled dataset B, and unlabeled dataset C. We did not impose restrictions in data recording sans the task directions, so that datasets are considered "in-the-wild." Data were recorded using the Tobii 4C pro upgraded eye-tracker, as shown in Fig. 4a, a sampling rate 90 Hz. We asked participants to read and answer MCQs in the format shown in Fig. 1a on a computer screen as shown in Fig. 4b. An eye-tracker fixed at the bottom of the screen records the participants' eye gaze. We used MCQs centered on four-choice English questions. Although this is not a typical crowdsourcing task since correct answers are known, it is useful for building a ground truth. All of the datasets were recorded with proper ethical clearance. The details of the datasets are as follows.

Labeled Dataset A. We recruited ten native Japanese university students and worked voluntarily. Each participant read and answered four-choice English grammatical questions on a computer screen. After answering each MCQ, the correctness of the answer is stored automatically, which constitutes the label of the dataset. In total, we collected 2,974 labeled samples.

Labeled Dataset B. We recruited 20 native Japanese university students to participate. Participants were paid 10 USD per hour for up to 4 h. We followed the same experimental procedure as above with a set of four-choice English grammatical questions. In total, 8,218 labeled samples were collected.

(a) Tobii 4C eye-tracker **(b)** Participant answering MCQs

Fig. 4. Data collection environment, (a) eye-tracker used for data recording and (b) participants' eye gaze being recorded while answering MCQs.

Unlabeled Dataset C. We recorded this dataset following the previous methods for four-choice English vocabulary questions; however, the answers remained unlabeled. We recruited 80 native Japanese high school students and worked voluntarily. In total, 57,460 unlabeled samples were collected.

5 Experiments

5.1 Experimental Conditions

The aim of our experiments is to estimate the correct answer rate using SVR, which can then be used to assess the quality of crowdwork. We used labeled dataset A for the estimation of the correct answer rate. Unlabeled dataset C and labeled dataset B are used for self-supervised pre-training and correctness estimation training, respectively, in the SSL method.

We employed the following three sets for the experiment using handcrafted features: (1) only the feature $f5$, i.e., answering time, (2) $f1$–$f4$, i.e., eye gaze features, and (3) $f5$ and $f6$, i.e., answering time and self-confidence as described in Table 1. Besides, using the feature vector generated by SSL, $f256$, we conducted one experiment. In addition to the above experiments, we employed a baseline estimator defined as, $c = \frac{1}{n}\sum_1^n c_n$ where c_n is the correct answer rate of the n^{th} window of the training dataset.

We conducted all correct answer rate estimation experiments in a participant independent way (leave one participant out cross-validation). As an evaluation metric, we used an absolute error that is calculated as $|c_t - c_p|$ where c_t and c_p are the true and predicted correct answer rate, respectively, for a window. We changed the window size from one to the maximum possible size of 102.

5.2 Results

Figure 5 shows the experimental results. It describes the change of mean absolute error in estimating the correct answer rate with the window size. For smaller windows, the mean absolute error decreases sharply for all methods, although it is relatively high. This indicates that the quality assessment by estimating the

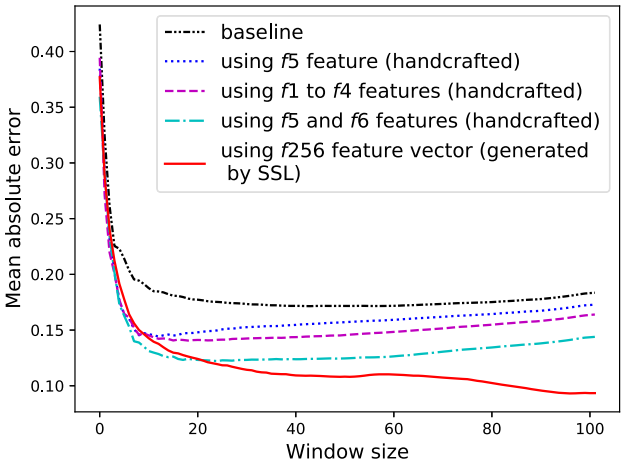


Fig. 5. Result of the correct answer rate estimation experiments.

correct answer rate is not an easy task by just taking into account the behavior for a short period of time. However, for larger windows, the tendency is different. As compared with the baseline, all proposed methods worked better. Among all handcrafted features, the use of $f5$ and $f6$ produced the best result. This is because self-confidence includes rich information about the correctness [10, 27], though it requires additional efforts by crowdworkers to declare the self-confidence for each task. The best performance was obtained by using the feature vector generated by SSL. At the largest window, the mean absolute error was 0.09. Note that in the feature vector generated by SSL, we do not include self-confidence manually so that they are easier to employ.

The best proposed method offers an absolute error around 0.1, which is 50% less than the baseline. This shows the advantage of using eye gaze information for quality assessment. We consider that the results show a new possibility of quality assessment using eye gaze—a richer fingerprint of crowdsourcing tasks.

6 Conclusion and Future Work

In this paper, we presented machine learning methods for the quality assessment of crowdwork by using eye gaze data, answering time, and self-confidence. The results are promising, especially with the SSL, and show the possibility that biometric data can be used to evaluate work quickly. With this, personalized adaptive crowdwork that is based on individual tasks is feasible. In the future, further experimentation on different types of tasks need to be conducted in order to gauge the suitability of the method and decouple it from burdensome tasks such as confidence labeling. Another important area that needs special focus is on leverage this technology for good, benefiting both crowdworkers and task-providers. This means developing platforms with clear ethical guidelines and regulations to ensure crowdworkers' rights.

Acknowledgments. This work was supported in part by the JST CREST (Grant No. JPMJCR16E1), JSPS Grant-in-Aid for Scientific Research (20H04213, 20KK0235), Grand challenge of the iLDi, and OPU Keyproject.

References

1. Amis, G.P., Carpenter, G.A.: Self-supervised ARTMAP. Neural Netw. **23**(2) (2010)
2. Baba, Y., Kashima, H.: Statistical quality estimation for general crowdsourcing tasks. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2013, Chicago, USA, pp. 554–562. ACM (2013)
3. Buscher, G., Dengel, A., Elst, L. V.: Eye movements as implicit relevance feedback. In: CHI 2008 Extended Abstracts on Human Factors in Computing Systems, CHI EA 2008, Florence, Italy, pp. 2991–2996. ACM (2008)
4. Daniel, F., Kucherbaev, P., Cappiello, C., Benatallah, B., Allahbakhsh, M.: Quality control in crowdsourcing: a survey of quality attributes, assessment techniques, and assurance actions. ACM Comput. Surv. **51**(1), 1–40 (2018)
5. Dontcheva, M., Morris, R.R., Brandt, J.R., Gerber, E.M.: Combining crowdsourcing and learning to improve engagement and performance. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2014, Toronto, Ontario, Canada, pp. 3379–3388. ACM (2014)
6. Gadiraju, U., Kawase, R., Dietze, S., Demartini, G.: Understanding malicious behavior in crowdsourcing platforms: the case of online surveys. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI 2015, Seoul, Republic of Korea, pp. 1631–1640. ACM (2015)
7. Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. CoRR, arXiv.abs/1803.07728 (2018)
8. Haresamudram, H., et al.: Masked reconstruction based self-supervision for human activity recognition. In: Proceedings of the 2020 International Symposium on Wearable Computers, ISWC 2020, Virtual Event, Mexico, pp. 45–49. ACM (2020)
9. Hettiachchi, D., van Berkel, N., Hosio, S., Kostakos, V., Goncalves, J.: Effect of cognitive abilities on crowdsourcing task performance. In: Lamas, D., Loizides, F., Nacke, L., Petrie, H., Winckler, M., Zaphiris, P. (eds.) INTERACT 2019. LNCS, vol. 11746, pp. 442–464. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-29381-9_28
10. Ishimaru, S., Maruichi, T., Dengel, A., Kise, K.: Confidence-aware learning assistant. arXiv:2102.07312 (2021)
11. Islam, M.R., et al.: Self-supervised deep learning for reading activity classification. arXiv preprint arXiv:2012.03598 (2020)
12. Jiang, H., Matsubara, S.: Efficient task decomposition in crowdsourcing. In: Dam, H.K., Pitt, J., Xu, Y., Governatori, G., Ito, T. (eds.) PRIMA 2014. LNCS (LNAI), vol. 8861, pp. 65–73. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-13191-7_6
13. Jung, H., Park, Y., Lease, M.: Predicting next label quality: a time-series model of crowdwork. In: AAAI Conference on Human Computation and Crowdsourcing. Association for the Advancement of Artificial Intelligence, Pittsburg, USA (2014)
14. Kazai, G., Kamps, J., Koolen, M., Milic-Frayling, N.: Crowdsourcing for book search evaluation: impact of hit design on comparative system ranking. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, pp. 205–214. ACM (2011)
15. Kazai, G., Kamps, J., Milic-Frayling, N.: Worker types and personality traits in crowdsourcing relevance labels. In: Proceedings of the 20th ACM International

Conference on Information and Knowledge Management, CIKM 2011, Glasgow, Scotland, UK, pp. 1941–1944. ACM (2011)
16. Kazai, G., Zitouni, I.: Quality management in crowdsourcing using gold judges behavior. In: Proceedings of the Ninth ACM International Conference on Search and Data Mining, WSDM 2016, San Francisco, USA, pp. 267–276. ACM (2016)
17. Kuang, L., Zhang, H., Shi, R., Liao, Z., Yang, X.: A spam worker detection approach based on heterogeneous network embedding in crowdsourcing platforms. Comput. Netw. **183**, 107587 (2020)
18. Kwek, A.: Crowdsourced research: vulnerability, autonomy, and exploitation. Ethics Hum. Res. **42**(1), 22–35 (2020)
19. Liu, X., Weijer, J.V.D., Bagdanov, A.D.: Exploiting unlabeled data in CNNs by self-supervised learning to rank. IEEE Trans. Pattern Anal. Mach. Intell. **41**(8), 1862–1878 (2019)
20. Moshfeghi, Y., Huertas-Rosero, A.F., Jose, J.M.: Identifying careless workers in crowdsourcing platforms: a game theory approach. In: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, pp. 857–860. ACM (2016)
21. Oppenlaender, J., Milland, K., Visuri, A., Ipeirotis, P., Hosio, S.: Creativity on paid crowdsourcing platforms. In: Proceedings of 2020 CHI Conference on Human Factors in Computing Systems, CHI 2020, Honolulu, USA, pp. 1–14. ACM (2020)
22. Raykar, V.C., Yu, S.: Eliminating spammers and ranking annotators for crowdsourced labeling tasks. JMLR **13**(16), 491–518 (2012)
23. Ross, J., Irani, L., Silberman, M. S., Zaldivar, A., Tomlinson, B.: Who are the crowdworkers? Shifting demographics in mechanical Turk. In: CHI 2010 Extended Abstracts on Human Factors in Computing Systems, CHI EA 2010, Atlanta, Georgia, USA, pp. 2863–2872. ACM (2010)
24. Rzeszotarski, J.M., Kittur, A.: Instrumenting the crowd: using implicit behavioral measures to predict task performance. In: Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology, UIST 2011, Santa Barbara, California, USA, pp. 13–22. ACM (2011)
25. Saeed, A., Ozcelebi, T., Lukkien, J.: Multi-task self-supervised learning for human activity detection. In: Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, vol. 3, no. 2, p. 30 (2019)
26. Tsai, M., Hou, H., Lai, M., Liu, W., Yang, F.: Visual attention for solving multiple-choice science problem: an eye-tracking analysis. Comput. Educ. **58**(1), 375–385 (2012)
27. Yamada, K., Kise, K., Augereau, O.: Estimation of confidence based on eye gaze: an application to multiple-choice questions. In: Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers, UbiComp 2017, Maui, Hawaii, pp. 217–220. ACM (2017)
28. Yuasa, S., et al.: Towards quality assessment of crowdworker output based on behavioral data. In: 2019 IEEE International Conference on Big Data, Los Angeles, USA, pp. 4659–4661. IEEE (2019)
29. Zeng, A., Yu, K., Song, S., Suo, D., Walker, E., Rodriguez, A., Xiao, J.: Multi-view self-supervised deep learning for 6D pose estimation in the Amazon Picking Challenge. In: 2017 IEEE International Conference on Robotics and Automation (ICRA), Singapore, Singapore, pp. 1383–1386. IEEE (2017)
30. Zhuang, M., Gadiraju, U.: In what mood are you today? An analysis of crowd workers' mood, performance and engagement. In: Proceedings of the 10th ACM Conference on Web Science, WebSci 2019, Boston, Massachusetts, USA, pp. 373–382. ACM (2019)