A Card Game for Collecting Human-Perceived Similarity Data of Artwork Images

ZHENAO WEI1, PUIJANA PALIYAWAN2, AND RUCK THAWONMAS3, (Senior Member, IEEE)

1Graduate School of Information Science and Engineering, Ritsumeikan University, Shiga 525-8577, Japan
2Research Organization of Science and Technology, Ritsumeikan University, Shiga 525-8577, Japan
3College of Information Science and Engineering, Ritsumeikan University, Shiga 525-8577, Japan

Corresponding author: Zhenao Wei (wza583122711@gmail.com)

This work was supported in part by the Grant-in-Aid for Scientific Research under Grant 19K12291, the Japan Society for the Promotion of Science, Japan.

The authors confirm that all human/animal subject research procedures and protocols are exempt from review board approval.

ABSTRACT In this paper, we propose a card game system named “Lottery” for collecting similarity data of artwork images. It is built based on the concept of game design called “Audience Participation Game With a Purpose (APGWAP),” which is to outsource computational steps in a given task to humans via a game that allows audiences to participate in a meaningful way. The game system is streamed on Twitch. In this game system, two Artificial Intelligence (AI) players match two cards that are most similar from a set of cards in their turn and discard them. The audiences can choose their roles between Assistant and Jury. Assistant helps the AI player of a given turn choose such a pair of cards, while Jury gives a score telling how similar a chosen pair is. In the experiments, different methods for providing rewards to audiences are investigated: equal reward, random reward, and performance-based reward. Experimental results show that the performance-based reward significantly promotes personal gratification and provides a better gaming experience on several other factors. In addition, it was found in our experiment that different groups of humans provided similar similarity scores for all the pairs of images, indicating that collecting promises reliable human data that are consistent through several trials is possible using our game system.

INDEX TERMS Audience participation game, game with a purpose, similarity data, technique for data collection, card matching game.

I. INTRODUCTION

In 2020, the spread of COVID-19 has forced many businesses and events to move online. This has also resulted in online artworks exhibiting platforms receiving more attention. For instance, Google Arts & Culture1 is a cooperation between Google and museums around the world to use Google Street View to shoot real-life scenes inside the museums as well as historical and famous paintings inside those museums in high resolution for people to appreciate. WikiArt,2 an encyclopedia-style collection of visual artworks, is dedicated to collecting various genres of paintings from all around the world and making them accessible to anyone, anywhere. Ritsumeikan University Art Research Center (ARC)3 has one of the representative ukiyo-e databases that has published more than 19,000 ukiyo-e images. To enhance accessibility to a database with this large number of images, it is essential to build a search engine (e.g., [1], [2]) and mechanisms for image recommendation (e.g., [3], [4]). In order to do this, data on the similarity between images is needed.

This paper presents a card game and its system for collecting human-perceived similarity data of ukiyo-e images by extending a previous live-streaming game system proposed by Wei et al. [5]. In their work, it was found that the similarity between a pair of ukiyo-e images was perceived differently between that perceived by humans and that calculated through computation methods based on image deep features. To accurately obtain the human-perceived similarity between ukiyo-e images, Wei et al. [5] proposed a card matching game based on the concept of Game With a Purpose (GWAP) as a solution. A GWAP [6] in general lets players play the game, and while they are having fun, useful

The associate editor coordinating the review of this manuscript and approving it for publication was Maurizio Tucci.

1https://artsandculture.google.com
2https://www.wikiart.org
3http://www.db-jac.net/db/nishikie-e/search.php?enter=default
data are collected. This card matching game was inspired by a traditional two-player card game called “Old Maid” (also similar to a Japanese card game called “Baba Nuki”).

“Lottery” is a game system presented in this paper, by adapting the above game [5]. It is called “Lottery” because each of the audiences who participates in the game as an “assistant,” has a random chance to earn a lot of in-game points. Major differences to the previous game system are that Lottery is not played by human players, but two AIs, and that it is an Audience Participation Game (APG) [7]; there is a group of audiences indirectly manipulating AI actions. It is not that the previous game by Wei et al. [5] has only two players, so it would pay many sessions of gameplay to collect similarity data. On the other hand, due to integration of APG, now, it is possible for Lottery to reach a large group of people over popular game live-streaming platforms.

This paper cover the three contributions. First, it presents the first Audience Participation Game With a Purpose (APGWAP) for collecting the human-perceived similarity of artwork images. Second, three different reward methods are proposed and evaluated in terms of user experience. Third, how to measure the reliability of the system for collecting similarity data is discussed, and the proposed game system is evaluated from this perspective.

II. RELATED WORK
This section starts by giving the background of Audience Participation Game With a Purpose (APGWAP) concepts. Next, details of the existing GWAP of Wei et al. [5] are further given. Finally, the importance and guidelines to design reward methods in APGWAPs are discussed.

A. AUDIENCE PARTICIPATION GAME WITH A PURPOSE
1) APG
An APG is a multiplayer cooperative game based on a streaming media platform. The audiences of a live-streaming game can partially manipulate the characters, environment or even streamer in the game by sending commands as chat inputs. APGs blur the concept of audiences and players [7], which allows more people to participate and audiences to have more communication. This concept is not new, but recently, the rapid development of streaming media platforms has driven the development of APG games. The most famous APG game is Twitch Plays Pokémon [8], which allows viewers to enter commands as chat inputs to control the actions of the game characters in real time. Recently, audience participation media have become popular. One famous example is a reality show “Rival Peak,” which features a competition between 12 AI contestants; audiences can help contestants they like through interactions, such as telling them what they should do.

In recent years, many streaming platforms have been introduced to the market, and they are contending with each other. Some of well-known platforms are Mixer,[5] YouTube Live[6] and Twitch.[7] Among them, Twitch has been the most popular, with more than 17.5 million daily visitors, 62% of which watch game-related live broadcasts.[8] Twitch also provides APIs for obtaining chat data in real-time, which greatly facilitates the development of APGWAPs [9].

2) GWAP
The concept of GWAPs, also known as “Human Computation Games (HCGs)” [10], was originally proposed by Ahn and Dabbish [6]. The idea of GWAPs is to use manpower to solve difficult-to-calculate problems [11]. To achieve this aim, difficult tasks are broken down into smaller subtasks, and games are then designed for these subtasks. Through such games, those subtasks can be completed by game players. Finally, the results of subtasks can be merged to deal with more difficult tasks.

Compared with mainstream entertainment games, designing a good GWAP faces two challenges: 1) designing a positive gaming experience, and 2) collecting reliable data. In addition, how to balance these two challenges is also an issue. Some previous studies (e.g., [12], [13]) assert that GWAPs should put priority on collecting reliable data, while some others (e.g., Krause et al. [14]) consider that positive gaming experience is more important. There are some successful GWAPs that well balance them, such as studies by Ahn and Dabbish [6] and Cooper et al. [15].

Examples of GWAPs are given in the following. Ahn and Dabbish [16] designed the first GWAP for labeling images. Since then, GWAPs have been used to deal with the tasks of labeling or classification in various fields, including music [17], [18] and astronomy [19]. The contribution of GWAPs to machine learning is also indispensable, for example, GWAPs were used to collect training datasets for machine learning algorithms, such as labeling location information [20], [21] and collecting photos [22], [23]. Nevertheless, recruitment of game players is still a challenging issue for GWAPs.

3) APGWAP
The concept of APGWAP was first proposed by Nguyen et al. [9]. The APG part solves the problem of insufficient players in GWAPs, and at the same time, the GWAP part makes the large number of data generated by the APG part useful, thus leading to APGWAPs being beneficial for serious purposes beyond mere entertaining purpose. Nguyen et al. [9] developed an APGWAP that aims to collect user descriptions of images and thus has a different purpose from ours. It should be noted that in APGWAP, there still exists a challenging issue in balancing positive gaming experience and collecting reliable data.

B. EXISTING GWAP
To make this paper self-contained, the aforementioned game by Wei et al. [5] is described as follows:

[4]https://rivalpeak.com/
[5]https://mixer.com/
There are two players. Each of them starts with a set of cards in their hand.

In each turn, one player “draws” a card from the opponent’s hand, and then discards a pair of cards that the player thinks is the most similar pair to earn points. Points can be viewed as “scores” in general sense, but here, the term “score” will be used to represent the “similarity score” between two images of interest.” The more similar the discarded cards are, the more points the player gets.

When a pair of cards is discarded, the other player may compare cards in his/her hand with the discarded cards, and in case the player finds that he/she has a better matched card for one of the discarded cards, he/she can “bid” to obtain the discarded pair to his/her hand. Later, if the player can discard a more similar pair, he/she may earn higher points than the cost of bidding. On the other hand, instead of bidding, the player may just “pass”.

These steps are repeated until there is no card left.

As we mentioned above, this GWAP is difficult to recruit players. Therefore, we proposed an APGWAP based on this game.

C. GAME REWARD
The importance of a reward mechanism to a game has been verified [24], [25]; appropriate rewards provide players with positive gaming experiences [26]. There are 7 common methods for giving rewards: score systems, experience point systems, item granting system, resources systems, achievement systems, plot animations and pictures, and unlocking mechanisms.

A game may choose one or several reward methods to enhance the player’s gaming experience. As a card game of APGWAP where the audiences participate by voting, a “score system,” namely using digital points to reward players for their performance in the game, is chosen for our game. Because Nagle et al. [27] reported that players have their subjective preferences for game rewards, three different implementations are compared: 1) performance-based reward, 2) equal reward, and 3) random reward. Note that the terms “points” will be used for “scores” in this context.

III. PROPOSED GAME LOTTERY
The APGWAP we proposed “Lottery” is a live-streaming audience participation card matching game running on Twitch.tv, inspired by the existing GWAP [5]. This game aims to collect reliable perceptual similarity data on ukiyo-e artwork image pairs through chat commands sent by audiences on the chat area of the channel. It is called “Lottery” because the audiences participating in the game as an “assistant” have a random chance to earn a lot of in-game points as their rewards. This section provides details of the system, including audience roles and game processes.

A. ROLES IN THE GAME
The gameplay of Lottery is constructed on top of interactions among two AI players and audiences of two different roles: jury, and assistant.

1) PLAYER
Similarly to the original system [5], the game consists of 2 players playing the card game. But in this APGWAP version, both players are AIs (NPCs standing for Non-Player Characters). Their actions are as follows:

- **Draw**: The player randomly draws one card from the opponent’s cards.
- **Discard**: The player calculates the similarity of possible card pairs from cards on its hand, and then discards the pair of the most similar cards.
- **Bid**: When the opponent discards a cards pair, the player may choose to bid and obtain the two discarded cards.
- **Pass**: When the opponent discards a pair of cards, the player takes no action.

To measure similarity between the cards in a pair of interest, the AI players employ cosine similarity to both images’ deep feature vector [28]. Although Wei et al. [29] stated that the similarity obtained by using methods based on deep features of images, i.e., machine-perceived similarity, was different from human-perceived similarity, the similarity...
based on deep features can still be used during the cold-start period where the number of collected data is not sufficient. After the cold-start period, when there are sufficient human-perceived similarity data, the similarity based on the deep feature vector will be replaced with human-perceived similarity.

In the game screen (cf. Fig. 1), the avatar of player 1 (P1) is displayed on the left, while cards in its hand are at the bottom of the screen. The avatar of player 2 (P2) and its cards are on the top. Action buttons for bid and pass are shown when P2 discards.

2) JURY
The Jury Group play a role in providing “similarity scores (scores)” reflecting how similar the discarded pair of images are. While the two players are AIs, the members of Jury Group and Assistant Group are human audiences. The members of Jury Group are a portion of the audiences of the channel. A jury can get points by scoring the similarity of the pair of cards discarded by both AI players. Juries will be rewarded after the end of each game.

a: JURY GROUP’s VOTING PERIOD
When a player discards, it triggers “bid or pass” options to the other player (action buttons will be shown, as can be seen in Fig. 1). At the same time, it triggers a voting period for Jury Group (cf. Fig. 1), allowing them to score the similarity of the discarded card pair.

b: JURY GROUP’s VOTING FORMAT
Each jury can give a score by putting an integer value, from 1 to 5, right after “!” into the Twitch chat input. “1” means the two cards are not similar at all, and “5” means that the two cards are similar at the highest level in their opinion.

c: JURY GROUP’s REWARDS
Each Jury gets points based on the resulting similarity scores from all the jury’s votes, which will be rewarded at the end of each game. Several methods can be used to calculate points, which are subject to experimentation in this study.

3) ASSISTANT
The Assistant Group play a role in helping P1 to discard the most similar pair of cards. They have to spend points to recommend a pair to discard. At the end of the game they participate, each assistant will earn a lot of points at a random amount, which is in the range empirically set between 3000 and 5000. This reward is called a lottery.

a: ASSISTANT GROUP’s VOTING PERIOD
Similarly to the Jury Group, the members of Assistant Group are made up of some audiences in the current channel. In each loop (cf. Fig. 2), a certain audience can choose whether to play a role as an assistant or a jury. Namely, if they participate in selecting the most similar pair of cards during the discarding session, they cannot give a similarity score in the biding session right after that; those who want to play as juries can not do anything during card selection (the drawing session).

As shown on the right image of Fig. 1, during each P1’s discarding session, an assistant can send a command into the chat input to recommend (vote for) the card pair that he/she thinks is the most similar, in effect, giving the similarity score of the pair. After the discarding session is over, P1 will summarize the assistants’ commands and discard a pair of cards that has been most recommended by the assistants.

b: ASSISTANT GROUP’s VOTING FORMAT
The format of the command to be sent by the assistant is !x; y; z. All the three variables are integer; x and y are indexes of the chosen cards (starting from 1 as the furthest left card), while z is the similarity score given by the assistant for the pair (similar to that given by a jury).

c: ASSISTANT GROUP’s REWARDS
Each time an assistant votes, z points will be spent from his/her current points. It’s noted that if the points are insufficient, the vote will not be allowed. Each assistant gets a chance to participate in the lottery at the end of this game. The chance of winning is proportional to the assistant’s participation in the game. The more they participate, the more possibility to win.

B. GAME PROCESSES
All of the game processes are shown in Fig. 2 whose details are given as follows:

1) The game starts from the drawing session of P1, which means that P1 draws a card from P2.
2) The game enters the discarding session of P1; P1 can either choose to discard the most similar card pair, or choose the Pass button to skip this session. During this session, members of the Assistant Group can send commands to help P1 choose the most similar card pair.
3) When the discarding session of P1 ends, the game enters the bidding session of P2. During this session, P2 will decide whether to obtain the opponent’s discarded cards through a bid action. At the same time, the members of the Jury Group will evaluate the similarity of the two cards that P1 has discarded.
4) Repeat (1)-(3) until one of the players has no cards. That player and the Jury Group end the game. However, for the sake of data collection, other player will repeat the discarding session until there are no cards and the whole game ends.

C. CHEATING PREVENTION
To avoid situations where audiences assist and then vote their own selected pair, the audiences in Jury Group and Assistant Group are separated. Those who vote as an assistant cannot later vote as a jury within the same round (Fig. 2). In addition, to prevent audiences from changing their vote to
follow majority votes to gain a higher score, they are allowed to vote only once during each voting period.

D. THE WINNING SIMILARITY SCORE
At the end of Jury Group’s voting period. The winning similarity score \( w \) of a certain pair of image is obtained by the following equation:

\[
 w = \begin{cases} 
 \text{Mode}, & \text{if there is one mode} \\
 \text{Median}, & \text{otherwise}
\end{cases} 
\]  
(1)

where \( \text{Mode} \) and \( \text{Median} \) are the mode and median of similarity scores given by all the juries. In other words, that \( w \) is determined based on a majority voting, but when there are multiple modes, the median is used instead.

E. REWARD METHODS
Three different methods for providing rewards to juries are introduced: performance-based reward, equal reward, and random reward. As from their names, the first one is to give points, or rewards, to each audience based on how well they perform, the second one is to give every audience an equal number of points, and the last one is to give a random number of points to each audience. We hypothesize that the first one may encourage audience to play the game seriously, but it may induce competition and thus discourage those who do not perform well.

1) PERFORMANCE-BASED REWARD
The idea of performance-based reward is to give those who voted a similarity score closer to the winning score a higher reward. Let \( s \) be the score given by a jury, then the closer to the winning score \( w \) the value of \( s \) is, the higher points the jury will earn. This can be expressed as an equation as follows:

\[
\text{points} = 5 - |w - s|
\]  
(2)

2) EQUAL REWARD
In this reward method, all the juries get points equal to the winning score as follows:

\[
\text{points} = w
\]  
(3)

3) RANDOM REWARD
Random points from 1 to 5 are given to each jury. Different juries may thus earn different points as follows:

\[
\text{points} = \text{random}(1, 5)
\]  
(4)

IV. EXPERIMENT
We conducted the experiment to evaluate the game system in terms of game experience and reliability. We compared three different reward methods to find the best one that promotes game experience. The experiment was conducted online, using Twitch as a live streaming platform.

A. DATASET
Ukiyo-e images from the ARC databases of Ritsumeikan University Arts Research Center were used in this study. We randomly selected 12 images from each of the three largest categories. Those categories are Yakusha-e, Bijin-ga, and Meisho-e.

B. PARTICIPANTS
As we introduced in D. GAME DIFFICULTY. Thus, we asked the participants to play the role of jury.
Participants in this study were undergraduate and graduate students studying in computer-science-related fields, aged 20 to 28 with 5 females and 21 males. Their nationalities are 9 Japanese, 9 Chinese, 3 Thais, 3 Indonesians, 1 Bangladesh, and 1 Korean. However, during the 3-week experiment, four participants withdrawn. To guarantee the reliability, we manually checked the collected data from the participants. Based on long-string screening [30], data of one participant were removed as he gave the same scores on all the questions. From our observation, data from the remaining participants were applicable; data from 21 participants were left.

As shown in Table 1, the participants were divided into 3 groups. The number of participants in the three groups were 9, 9, and 8, respectively (Valid data for each group 8, 7, and 6). Each group engaged in a 3-week experiment with a different reward method each week.
TABLE 1. The weekly schedule of the experiment.

|       | Group 1                  | Group 2                  | Group 3                  |
|-------|--------------------------|--------------------------|--------------------------|
| Week 1| Performance-based reward | Random reward            | Equal reward             |
| Week 2| Equal reward             | Performance-based reward | Random reward            |
| Week 3| Random reward            | Equal reward             | Performance-based reward |

C. EXPERIMENTAL PROCEDURES

For a certain group in a week of interest, the experimental procedure in each session consists of 3 steps:

1) **Instruction:** Watch a video telling the goal of the game, providing game details, and instructing on how to play as a jury.

2) **Gameplay:** Play three games where each game uses a different category of ukiyo-e images; after each game, the total point of the current session of each player is displayed.

3) **Questionnaire:** Complete a questionnaire.

D. GAME DIFFICULTY

Wei et al. [29] introduced a concept of difficulty in terms of card matching. With respect to selection of a card pair from a set of cards, if most of the participants choose the same pair, it is considered that the difficulty of card matching for that set of cards is low. On the other hand, if opinions are divided, the difficulty is high.

To maintain the same game difficulty throughout the experiment, the initial cards, draw order, and discard order of the two AIs (NPCs) were fixed in each game of Yakusha-e, Bijin-ga, and Meisho-e. Namely, each participant sees the same game of a given ukiyo-e category each week for three weeks. However, the order of games of different categories is changed weekly based on a Latin square as shown in Table 2. In addition, the Assistant Group can interfere with the card sequence and hence affect the game difficulty. This is why in this experiment all the participants played only one role of jury. Moreover, each participant must vote during every voting period.

E. EVALUATION METHODS

The game system was evaluated in terms of game experience and reliability of collected data as described in the following.

1) **GAME EXPERIENCE**

To evaluate game experience, a questionnaire named Game User Experience Satisfaction Scale (GUESS) [31] was used. From the total of 8 factors in the original questionnaires, 6 of them were applicable to our system, and thus three factors – Narratives, Audio Aesthetic, and Visual Aesthetics—were removed. The reasons behind this removal are that 1) our game does not have narrative, 2) sound/music has not been added yet, albeit planned to, and 3) the participants are playing the same game and hence comparing Visual Aesthetics between different groups or different reward methods does not provide any information. The questions used for each factor are shown in Table 3. For each factor, 5-point Likert scale was employed, where 1 is “strongly disagree” and 5 is “strongly agree”. All the questions were shown to participants in a random order.

2) **THE RELIABILITY OF COLLECTED DATA**

The reliability of the collected image similarity data by the proposed game system was evaluated by analyzing votes from the participants. The reliability in this context indicates whether the system is capable of collecting similarity scores in a consistent manner. More specifically, when a same pair of images is presented to different groups of people, they should provide similar scores for that pair of images.

It is noted that the evaluation on gaming experience compares three different reward methods, but the reliability of data collection does not. This is because the participants’ voted before seeing their reward, so the reward method did not affect their decision.

V. RESULTS

A. GAME EXPERIENCE

As shown in Fig. 3, regardless of the reward method in use, the average values on most factors of game experience were above 3 (the neutral level in this 5-point Likert scale), which indicates that the game system was perceived positively. However, Play Engrossment and Personal Gratification were lower than 3 when random reward was used. Comparing the three different reward methods, it can be seen that
### TABLE 4. Results of Reliability analysis using p-values from statistical tests the paired sample t-test and the Wilcoxon signed-rank test (asymptotic) between mutual the performance-based reward, equal reward and random reward.

|                      | Usability | Play Engrossment | Enjoyment | Creative Freedom | Personal Gratification | Social Connectivity |
|----------------------|-----------|------------------|-----------|-----------------|-----------------------|-------------------|
| **Average (± SD)**   |           |                  |           |                 |                       |                   |
| Performance-based    | 4.12 ± 0.71 | 3.09 ± 0.69      | 3.29 ± 0.72 | 3.64 ± 0.87    | 3.26 ± 0.62          | 3.64 ± 0.78       |
| Equal                | 4.12 ± 0.72 | 3.071 ± 0.69     | 3.21 ± 0.82 | 3.38 ± 0.93    | 3.17 ± 0.80          | 3.55 ± 0.91       |
| Random               | 3.95 ± 0.79 | 2.98 ± 0.78      | 3.24 ± 0.68 | 3.43 ± 0.86    | 2.89 ± 0.71          | 3.57 ± 0.73       |
| **Shapiro-Wilk test for normality (p-value)** |          |                  |           |                 |                       |                   |
| Performance-based    | 0.018     | 0.109            | 0.026     | 0.039           | 0.009                | 0.435             |
| Equal                | 0.003     | 0.027            | 0.002     | 0.041           | 0.014                | 0.117             |
| Random               | 0.019     | 0.006            | 0.013     | 0.122           | 0.079                | 0.093             |
| **Group Analysis**   |           |                  |           |                 |                       |                   |
| Friedman Test for Repeated-Measures | 0.1525 | 0.6428 | 0.5856 | 0.1770 | 0.0336 | 0.8372 |
|                       | 0.5580 | 0.9200 | 0.9535 | 0.4724 | 0.0655 | 0.6440 |

*The p-value cannot be obtained in Wilcoxon test due to having a number of ties in this sample size.*

**FIGURE 3.** The average value of GUESS scores for performance-base reward, equal reward and random reward.

The performance-based reward method can provide the best experience in addition to Play Engrossment factors on average.

The Shapiro-Wilk test for normality was used to analyze the characteristics/distributions of collected data (similarly to what is done in [32]). Some distributions of data were found normally distributed. In order to compare the three groups, if data of all the three groups were normally distributed, the ANOVA was used, else the Friedman Test for Repeated-Measures was used. Similarly for the ad hoc analysis, if data of both groups were normally distributed, the t-test was used, else the Wilcoxon test was used.

Results of the statistical tests are shown in Table 4. For the ad hoc analysis, based on Bonferroni adjustment ($k = 3$), we use an alpha of $0.05/3 = 0.0167$ for the confidence interval of 95%. Group analysis reported no difference on all the factors. However, the ad hoc analysis reported a significant difference when comparing the Personal Gratification factor between the performance-based reward and the random reward methods ($p = 0.0096$).

### B. RELIABILITY

We examined whether the game can collect the reliable human-perceived similarity data (i.e., data are consistent through several times). We compared similarity scores of the same image pairs collected from different groups of humans. By fixing the initial cards on hands, we ensured that 7 pairs of images will be shown to all the participants to ask for their similarity scores throughout the experiment.

The results are shown in Table 5. First, Z-score normalization was applied; scores given by each participant were normalized using the mean and the standard deviation of scores that the participant had provided for the images so far. This normalization was for eliminating personal bias and differences in personal baseline (e.g., some participants usually gave high scores, while some participants usually gave low scores). After within-participant normalization, it was found that all the distributions of normalized scores by a group of participants for a pair of images were normal. Therefore, the ANOVA test was used to compare the normalized scores, for each pair of images, given by three different groups of participants. No difference was found on all pairs, indicating that scores were consistently obtained through different trials.

### C. DISCUSSION

In the results of game experience, when using random reward, the scores on Play Engrossment and Personal Gratification were lower than 3. This is reasonable because the game did not give audiences back the rewards that reflected their efforts when they seriously played the game. Comparing the three different reward methods, the performance-based reward method provided the best experience, especially on personal gratification. Therefore, this reward system will be applied to our system.

We have considered the possibility that one’s decision to provide a similarity score could be affected by the decisions of others (Keynesian Beauty Contest effect). However, we believe that when the number of participants is large, votes are expected to be diverse and the above effect should have less impact. In addition, from our pilot test, similarity scores...
or votes were found to be consistent across different groups of people, proving the reliability of the system. Moreover, we follow a practice by Nguyen et al. [9] and design a short voting time (15 seconds) to the “following others” phenomenon on voting. Nevertheless, we are aware that one’s votes may be affected by prediction of votes by other people (even if that person doesn’t see those votes), therefore we will invent some methods for better handling this issue in future work (e.g. Hide participant’s vote or quickly swipe the screen after each vote).

From the experiment, we knew that personal bias/baselines should be taken into consideration. To handle this issue, in future work, we will employ Z-score normalization in our game (on the part when AI calculates similarity scores for candidate pairs of images to be discarded or bid). Nevertheless, a limitation of within-participant Z-score normalization is that the normalization may not be applicable, or scores may not be accurate, for new participants who vote only a few times—we can consider the case when a participant has only experienced 2 pairs of image and has given a score of 5 for them. The solution we have now is that, to perform a Z-score normalization on a certain participant, the real average and the real standard deviation will be used only if the participant has a difference between his maximum and minimum vote scores greater than 1 unit (this implies that the participant has already experienced at least one similar pair and one dissimilar pair of images), otherwise, an average of 3 with a standard deviation of 1 will be used.

VI. CONCLUSION

In this paper, we introduced a card game system, based on the game design concept of APGWAP, for collecting similarity data of image pairs. An experiment in this paper offered evidence that the game system can provide preferable game experience and is capable of collecting reliable data. The performance-based reward method was found most preferable, especially in terms of promoting personal gratification of audiences in APGWAP. Data collected by the system was also found to be consistent among different player groups.

In the future, we will includes a study on game experience of assistants. This will be followed by work on applying data collected from this game system in developing a recommender system. It is expected that the game will reduce laborious work and time costs for collecting data needed for future work on development of a recommender system for an ukiyo-e artwork database.
