Improving Deep-Feature Image Similarity Calculation: A Case Study on an Ukiyo-e Card Matching Game Lottery

ZHENAO WEI[ID][1]**, PUJANA PALIYAWAN**[ID][2]**, AND RUCK THAWONMAS**[ID][3]**, (Senior Member, IEEE)**

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

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

ABSTRACT The purpose of this study is to improve AI players in Lottery. Lottery is a card matching game designed based on the concept of Audience Participation Game With a Purpose. Namely, it lets live streaming game audiences take part in gameplay and collects image similarity data perceived by these audiences. The game employs two AI players that initially calculate image similarities based on deep features (deep-feature similarities). In our previous study, it was found that similarities, between a certain pair of images, perceived by machines or computers – calculated based on deep features – were different from similarities perceived by humans. This would make gameplay by the AI players unbelievable, in other words, non-human-like. This study, therefore, proposes to use a linear model, built based on pre-collected human data, for improving the deep-feature similarities. The amount of human data required to make the model stable is also discussed. Experimental results show that the linear model only requires small quantities of human data to greatly improve the deep-feature similarities. At the same time, our results also show that the game Lottery is indispensable. This is because the linear model can only make the calculated similarity closer to that of humans, but there is still discernible difference; in order to obtain accurate similarity between images, it is necessary to collect a certain amount of human-perceived similarity data.

INDEX TERMS Audience participation game, game with a purpose, audience participation game with a purpose, technique for similarity data collection, card matching game, human-perceived similarity, deep feature similarity.

I. INTRODUCTION

Streaming media platforms are increasingly receiving attention, for example Mixer,[1] YouTube Live[2] and Twitch.[3] With live streaming games growing in popularity, a game design concept trend called Audience Participation Game (APG) [1], [2] has emerged and blurred the line between audiences and players by allowing audiences to partially control the game. Furthermore, some researchers integrated this APG concept with another game design concept called Game With a Purpose (GWAP) [3], [4], resulting in a hybrid concept called Audience Participation Game With a Purpose (APGWAP) [5], [6]. The key idea of APGWAP is to design an interesting game that can be broadcast on streaming platforms where audiences can participate, while at the same time the game publisher/streamer can collect a large amount of useful data from gameplay. This is thanks to the number of daily visitors to a popular site like Twitch exceeding 17.5 million, 62% of which watch game-related live broadcasts.[4]

Lottery [6] is an APGWAP designed by our group for collecting similarity data on ukiyo-e images. It is a card game between two AI players, where in each turn, an AI in charge discards the most similar pair of cards from the set of cards in its hand. Human audiences can help Player 1, one of the AIs, choose the most similar card pair, by providing similarity scores for image pairs in its hand. For pairs with

The associate editor coordinating the review of this manuscript and approving it for publication was G. R. Sinha[ID].

[1]https://mixer.com/
[2]https://www.youtube.com/live
[3]https://www.twitch.tv/

[4]https://twitchadvertising.tv/

no human-provided similarity input, i.e., when Player 1 has no human helps, it would calculate their similarity scores by using a method based on deep features. Previous work [7] pointed out that such scores are different from those likely to be provided by humans, which makes the AI players play in a non-human like fashion. Therefore, this work aims to solve this problem by using a regression model; it is expected that deep-feature scores adjusted by a regression model will be closer to scores given by humans.

The contributions of this work are as follows:

- We demonstrate that regression models trained with human-perceived similarity data can effectively improve the quality of deep feature similarity.
- We examine the minimum number of votes required to build a stable model.
- We find the minimum number of human data for each image pair to switch from the improved deep-feature similarity to the similarity based on the human votes received so far, which in turn proves the importance of Lottery in its data-collection role.

II. RELATED WORK

A. UKIYO-E IMAGE DATABASE

Ukiyo-e is a popular print art in Japan during the Edo period. Because of its low price and often depicting everyday scenes [8], it has been loved by the public since the 17th century [9], and is also an important part of art history [10]. A number of ukiyo-e public databases exist. For example, the database of the Art Research Center (ARC)[5] of Ritsumeikan University, used in our research, has over 19,000 ukiyo-e images.

B. DEEP FEATURE SIMILARITY

"Deep features" is a term used to refer to a set of features derived from the image itself. It has become widely used; for example, it was first used for style transfer by Gatys et al. [11]. Then, Chu and Wu [12] and Matsuo and Yanai [13] discussed a variety of expressions for deep features. Among such expressions, Wei et al. [14] found that *Cosine* is the most suitable expression for ukiyo-e images. *Cosine* is a feature matrix obtained by the cosine-similarity calculation among pairs on the feature maps of the conv5_1 layer in the deep learning model VGG-19. Recently, such features extracted from VGG-19 are still widely used, e.g., Zhang and Yamasaki [15] reported their effectiveness in image recommendation.

C. MULTIPLE FEATURE FUSION

Kayhan and Fekri-Ershad [16] proposed to fuse four kinds of non-deep-learning image features as the representation of images: modified local binary patterns, local neighborhood difference pattern, gray level co-occurrence matrix, and color histogram. However, their method only fuses non-deep-learning features, so there is room for deep learning

features to be added for fusion. In addition, their method's performance is sensitive to weight selection.

Pathak and Raju [17] fused multiple image features, both non-deep-learning features and deep learning features, and showed that more features in fusion lead to improved performance. However, as with [16], their performance is also sensitive to weight selection.

Lu et al. [18] proposed a feature-fusion method based on information entropy theory and relevant feedback to adapt the weights of image features. However, their relevant feedback of a feature of interest is based on how good the feature is in selecting images that have the same category as the query image. This indicates that their work does not directly focus on similarity that is close to human perception.

D. LINEAR REGRESSION

The linear regression method presents the relationship between independent and dependent variables by fitting a linear equation to the data [19]. Its advantages are ease of use and interpret-ability [20], and it is often used for prediction tasks (e.g., [21], [22]). For some special cases, linear regression even outperforms deep learning networks when the number of training samples is small (e.g., [23]).

E. AUDIENCE PARTICIPATION GAME WITH A PURPOSE

APG is a type of game in which audiences can manipulate the characters or game environment through various methods, such as sending commands as chat messages. It reduces the boundary between audiences and players as well as promotes social communication [24]. The most famous APG on the Twitch platform is arguably Twitch Plays Pokémon [25]. In addition, the concept of APG has recently been integrated with other fields, such as the APG reality show "Rival Peak.[6]"

The concept of GWAP, invented by Ahn and Dabbish [26], breaks down difficult tasks into smaller subtasks, then allows players to solve those subtasks – through the game – and finally merges the results of the subtasks to deal with the original task. It has been applied to tasks such as labeling, classification, and collection in many fields, including music [27], [28], astronomy [29], and machine learning [30], [31]. Because it uses manpower to solve difficult-to-calculate problems [32], GWAPs are also known as "Human Computation Games (HCGs)" [33].

APGWAP is a hybrid concept recently proposed by Nguyen et al. [5]. It combines advantages from the above two concepts. In this combination, APG helps reaching large numbers of people (audiences), and GWAP enables getting useful data from them.

F. LOTTERY: APGWAP FOR COLLECTING UKIYO-E SIMILARITY DATA

Lottery [6] is an APGWAP designed to collect human-provided similarity data for pairs of ukiyo-e images. Lottery

---

[5] http://www.dh-jac.net/db/nishikie-e/search.php?enter=default

[6] https://rivalpeak.com/

is inspired by a traditional card game called Old Maid [34]. It is designed to stream on, but not limited to, Twitch.tv. There are three types of parties in Lottery: Player, Jury, and Assistant. Players are two AIs, while Juries and Assistants are human audiences who participate in gameplay through chat-message commands.

The gameplay flow of Lottery is shown in **Fig. 1**. This figure shows AI 1's turn. On AI 2's turn, which is after the Bid & Jury and before the next Draw, the positions of AI 1 and AI 2 are swapped in the figure. The audiences can only see the cards of AI 1, which is the side they can help. At the beginning of each game, both AIs will be allocated a certain number of cards (6 cards or images per player by default). On each turn, the player of that turn draws one of the opponent's cards and then discards a pair of the two most similar cards from its hand. If the opponent thinks it can more similarly match a card in its hand with one of the cards in the discarded pair, it can choose to bid and then obtain both discarded cards. This process continues until no card is left – when cards run out in the hand of one player, that player ends their part of the game, leaving the other to finish their cards.

The roles of Jury and Assistant are in the following. Juries judge the cards that each player discards by giving a score that evaluates how much they are similar. This is done by inputting !*x*, where *x* is an integer score from 1 (least similar) to 5 (most similar). Assistants help Player 1 choose the most similar pair of cards by inputting a command in the form of !*x*; *y*; *z*, where *x*, *y* are the indices of the cards in Player 1's hand, and *z* is the similarity score of the pair.

Card pair selection by Player 1 or Player 2 is directly based on the deep-feature similarity (cosine similarity between the aforementioned *Cosine* features of images) without using audiences' similarity scores (votes). However, it was found that there is difference in perceptual similarity between humans and machines, i.e., even normalized to have the same scale, similarity scores calculated using deep-features are significantly different from those that humans provide [7]. This makes gameplay by both AI players non-believable.

III. IMPROVING DEEP-FEATURE SIMILARITY USING A REGRESSION MODEL

In this paper, we propose to use a linear-regression model based on collected similarity scores by humans to reduce the difference described in the end of II-E. The resulting similarity from the model is called "improved deep-feature similarity." In addition, in this work, both AI players make a card-selection decision based on not only the improved deep-feature similarity but also on the average of the votes, received up to the current time from both juries and assistants, for each image pair of interest.

Let X, x_i, and y_i be a set of image pairs, the deep-feature similarity of image pair i, and the human-perceived similarity (the average of votes) of i, respectively, where $i \in X$. Our goal is to improve x_i using a regression model. The resulting value (the improved deep-feature similarity) is denoted as $\widehat{y_i}$. As mentioned above, a linear model is used for the estimation
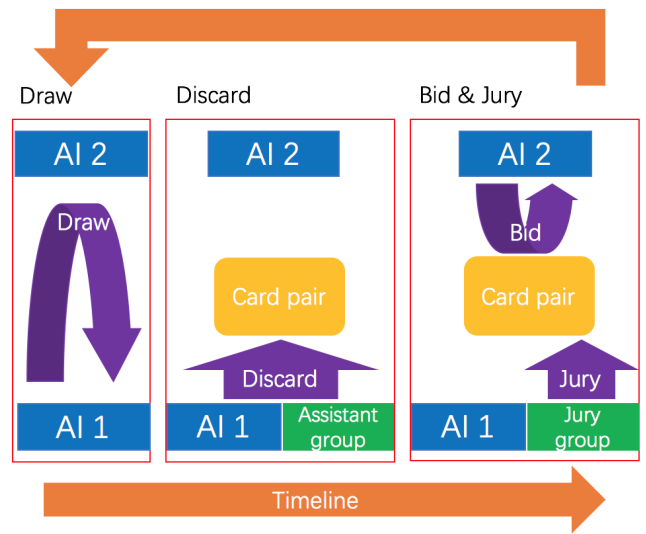

FIGURE 1. Lottery's gameplay flow.

(cf. Equation (1)); such a linear model is built using the data of pairs that have votes. It is noted that linear regression was chosen out of multiple candidate models (cf. V-B).

$$\widehat{y_i} = ax_i + b \qquad (1)$$

where a and b are the slope and the constant of this linear function.

Let s_i be the similarity score that an AI player uses for image pair i. It is considered that the more votes, the less important the model becomes. Thereby, when an image pair has less than, say, t votes, they will be substituted with the improved deep-feature similarity. When an image pair has t or more votes, its similarity score will solely be the average of its votes. This is summarized in (2). The value of t is investigated through an experiment in this study.

$$s_i = \begin{cases} \frac{1}{n} \sum_{j=1}^{n} y_{ij} & \text{if } n \geq t \\ \widehat{y_i} & \text{otherwise} \end{cases} \qquad (2)$$

where n and y_{ij} are the number of votes and the jth vote, respectively, for image pair i, and $\widehat{y_i}$ is the improved deep-feature similarity in (1).

To demonstrate the role of (2), let us assume an AI player of interest is choosing the most similar image pair among five image pairs: A, B, C, D and E. Deep-feature similarity, scaled to have a range from 1 to 5, gives scores of 4.24, 4.16, 4.16, 4.04, and 4.22, respectively. With the model in (1), the scores become 2.75, 2.56, 2.54, 2.24, and 2.68. Regardless of using (1), the AI player will select pair A. However, let us assume further that human scores of 3.36 and 3.80 exist for pairs C and D and that their number of votes is above t, the AI player will now choose pair D according to (2), rather than choosing pair A. This example is based on real data, which are pairs #50, #51, #52, #53, and #54 in **Fig. 2**.
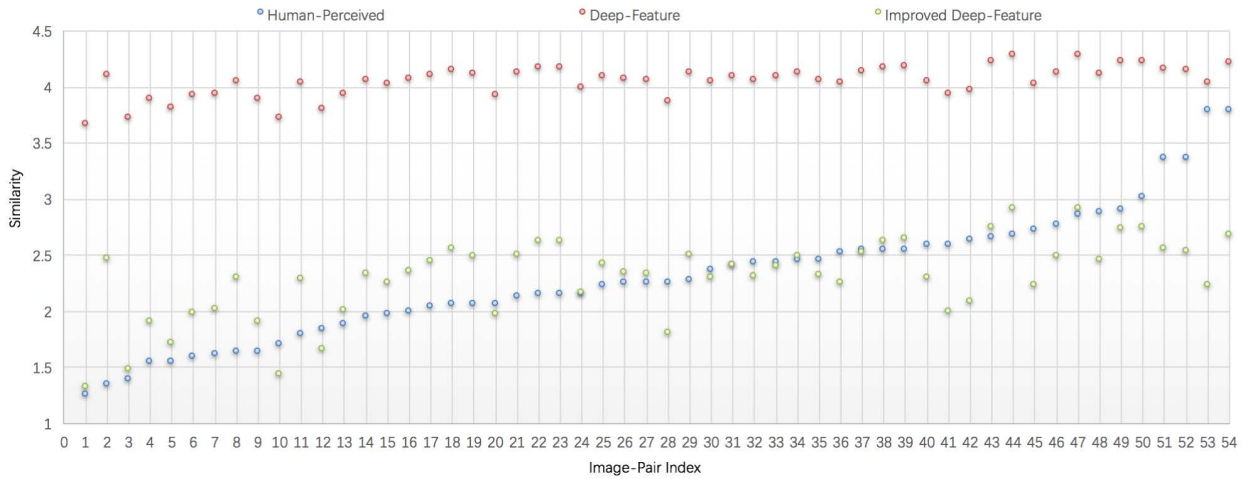
FIGURE 2. Comparison of human-perceived similarity, deep-feature similarity, and improved deep-feature similarity.
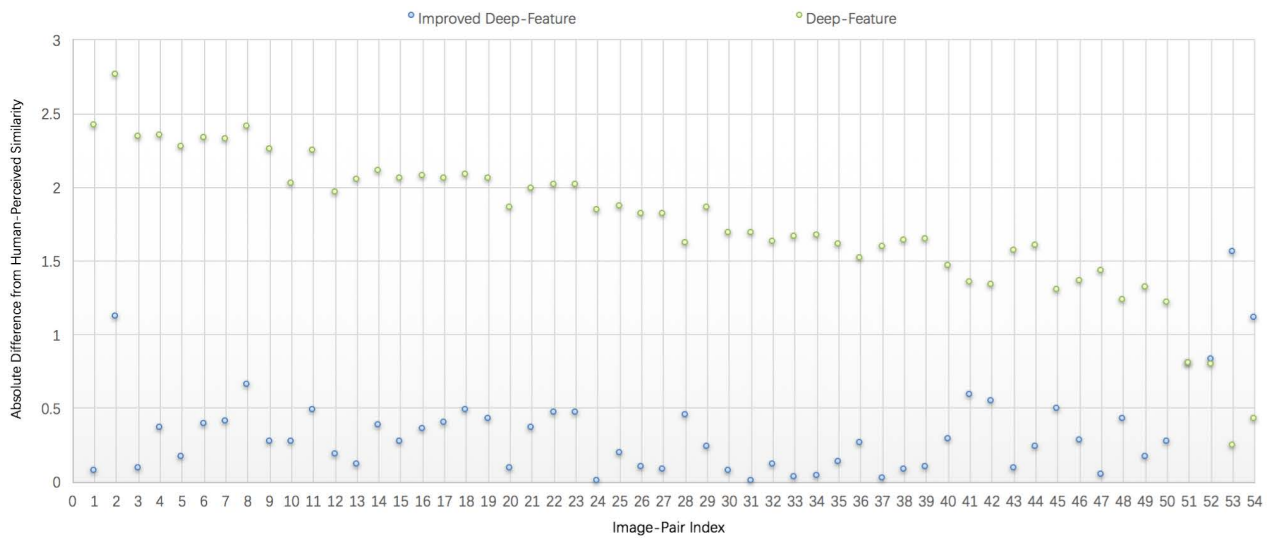

FIGURE 3. Absolute difference of either the improved deep-feature similarity or the deep-feature similarity from the human-perceived similarity.

IV. EXPERIMENT

An experiment was conducted for three major purposes as follows:

1) To prove that a regression model (the linear model in (1)) can improve the deep-feature similarity, by verifying if it outputs values closer to those perceived by humans or votes (cf. V-A).

2) To find the minimum number of votes needed to make the regression model in (1) stable, by locating the number of votes used in training of the model where the performance starts to saturate (cf. V-B).

3) To examine whether vote collection and hence the Lottery game itself are still needed for new image pairs when such a model exists, by deriving a point (t in (2)) where the quality of the improved deep-feature similarity starts to fall behind that of the average vote (cf. V-C).

Supplementary material such as all the experimental data can be found on our Open Science Framework page.[7]

A. DATASET (IMAGE PAIRS)

Ukiyo-e images used in this research are from the aforementioned ARC database. We randomly selected 12 images from three image categories that have the largest number of images: Yakusha-e, Bijin-ga, Meisho-e. Images in each category were equally divided into three sets, thereby having in total nine image sets. **Table 1** shows all the image sets used in this experiment.

B. COLLECTION OF HUMAN DATA (QUESTIONNAIRE)

The four images in each set were used to create pairs of images, resulting a combination of six pairs or questions.

[7]https://osf.io/v4fpc/

TABLE 1. All the image sets used in this experiment.

Index	Image 1	Image 2	Image 3	Image 4
Set 1 (Yakusha-e)				
Set 2 (Yakusha-e)				
Set 3 (Yakusha-e)				
Set 4 (Bijin-ga)				
Set 5 (Bijin-ga)				
Set 6 (Bijin-ga)				
Set 7 (Meisho-e)				
Set 8 (Meisho-e)				
Set 9 (Meisho-e)				

An online questionnaire was used to present 54 questions (6 pairs x 9 sets) in random order to participants and ask for their similarity scores. As in Lottery, scores are integer values ranging from 1 to 5.

C. PARTICIPANTS

Forty-four participants engaged in this study. They were undergraduate and graduate students studying in computer-science-related fields, aged 19 to 29, 38 males and
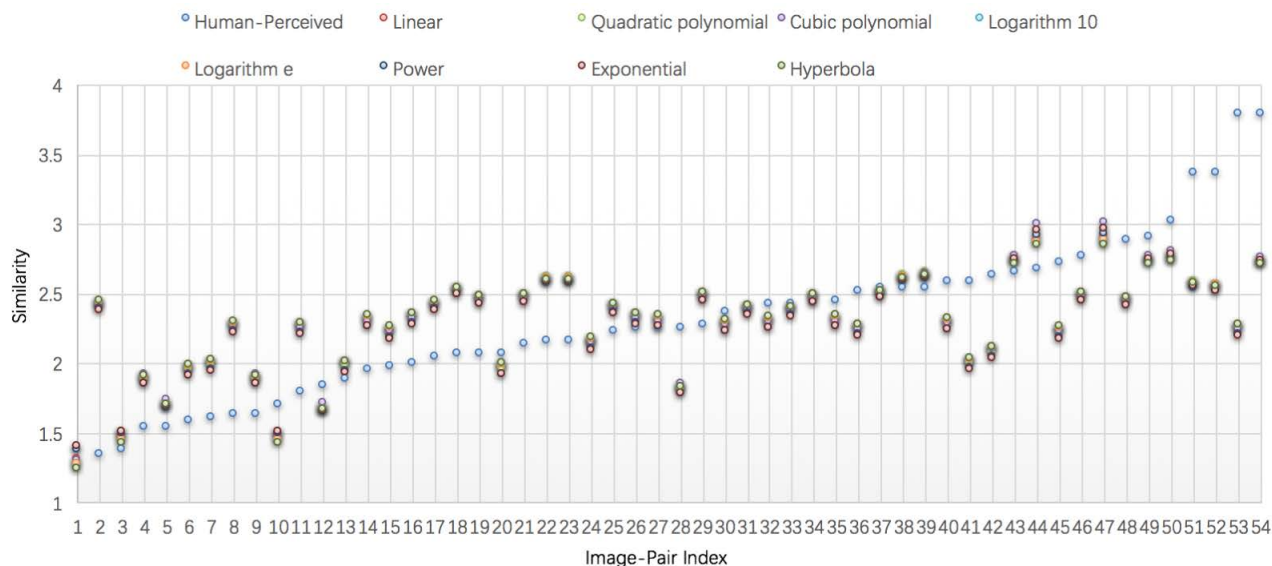
FIGURE 4. Comparison of the human-perceived similarity and eight regression models.
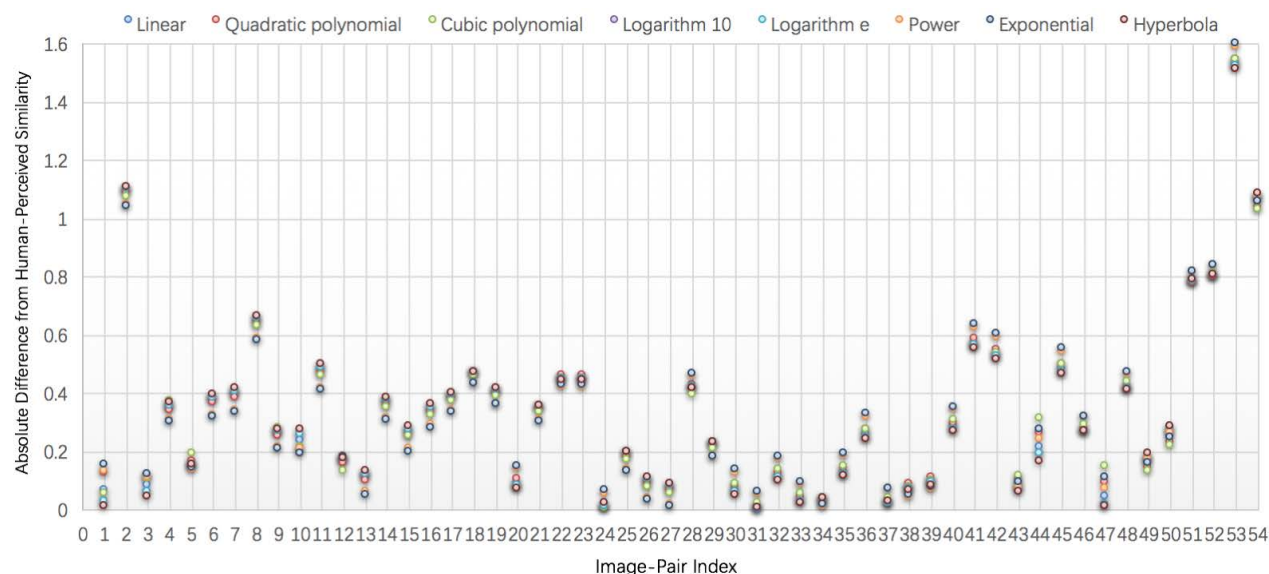

FIGURE 5. Absolute difference of each of the eight regression models from the human-perceived similarity.

6 females. A link to the questionnaire was sent to them via email.

V. RESULTS

A. IMPROVEMENT OF IMAGE SIMILARITY WITH REGRESSION

Here we compare three types of similarities: human-perceived similarity, (original) deep-feature similarity, and improved deep-feature similarity. Since the second one is a cosine similarity, with an output range of from 0 to 1, it is scaled to have a range from 1 to 5, which is the voting range by audiences. All the votes are used in taking the vote average

to obtain human-perceived similarity values and in training a regression model to obtain improved deep-feature similarity values. All figures in this sub-section, **Fig. 2 and Fig. 3**, have the x-axis sorted in ascending order of human-perceived similarity.

Fig. 2 shows the results of the three similarities, where the linear regression model in (1) is used for the improved deep-feature similarity. Differences can be clearly seen when comparing the deep-feature similarity (the red points) with the human-perceived similarity (the blue points). The improved deep-feature similarity (the green points) is found to be much closer to the human-perceived similarity.

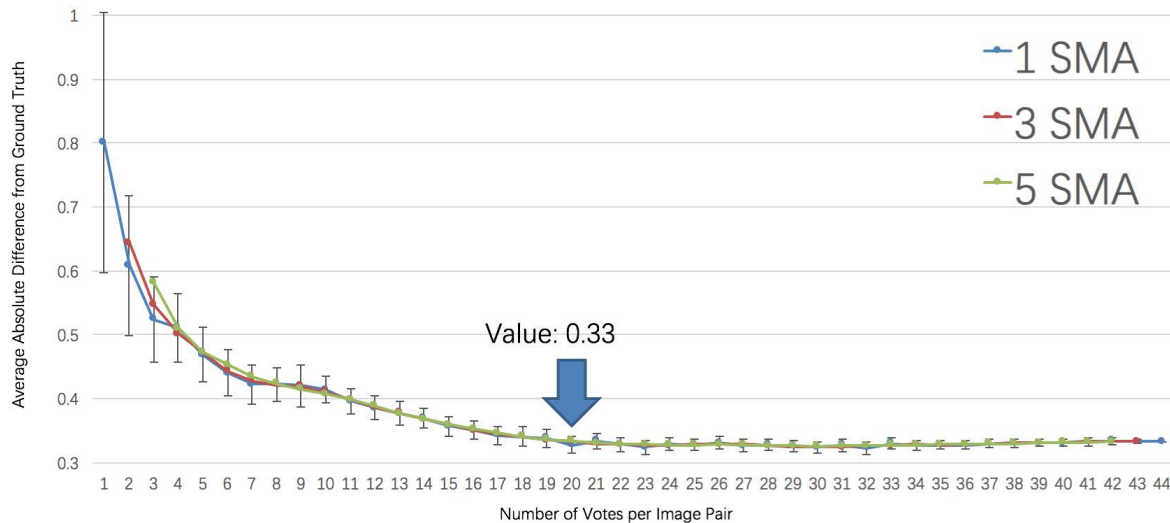|	Linear	Quadratic polynomial	Cubic polynomial	Logarithm 10	Logarithm e	Power	Exponential	Hyperbola
R-squared	0.3934	0.3945	0.3958	0.3919	0.3919	0.3943	0.3935	0.3896
Average absolute difference	0.3225	0.3248	0.3245	0.3215	0.3215	0.3220	0.3236	**0.3213**

FIGURE 6. Linear model's average absolute differences from the ground truth and their error bars; results of the simple moving average (SMA) with three window sizes are also shown.

Fig. 3 shows the absolute difference of either the improved deep-feature similarity or the deep-feature similarity from the human-perceived similarity. The results confirm that the former is much lower than the latter on most image pairs. The improved deep-feature similarity has higher values of difference only when the human-perceived similarity is near or above 3.5, i.e., from the 52th image pair.

B. COMPARISON OF REGRESSION MODELS

As described earlier, linear regression is used in our work. This is substantiated in this sub-section. Here, we compare the predicted accuracy of multiple commonly used models: linear, quadratic polynomial, cubic polynomial, logarithm 10, logarithm e, power, exponential, and hyperbola. As done in the previous sub-section, all the votes are used in obtaining human-perceived similarity values and training each regression model, and relevant figures, **Fig. 4** and **Fig. 5**, have their x-axis sorted in ascending order of human-perceived similarity.

Fig. 4 and **Fig. 5** show prediction results from each model and their differences from the human-perceived similarity, respectively. **Table 2** shows for each model both the R-squared and the average absolute difference, among all the image pairs, from the human-perceived similarity. Since there is no much difference among the models, linear regression was chosen due to its low complexity.

C. DISCUSSION ON THE REQUIRED AMOUNT OF DATA

In order to find the minimum number of votes required to effectively build a linear model, we examine how the training size – defined as the number of votes used for training per image pair – affects the prediction accuracy. For each training size r, all combinations $44Cr$ are evaluated, except for each r from 3 to 41 with their $44Cr$ larger than 1000 where 1000 combinations are randomly chosen for evaluation.

In addition, three-fold cross validation is applied. In particular, 18 images in each of the three categories are grouped into three sets. Thereby, in each fold of cross valuation, 36 image pairs (12 per category) are used for training a regression model with training size r; and the remaining 18 image pairs (six per category) are used for testing, from which the average absolute difference (AAD), among all the aforementioned combinations and the three folds, from ground truth is obtained. Note that henceforth the ground truth for each image pair is defined as the human-perceived similarity when all of its 44 votes are used.

Fig. 6 shows the results, where blue dots represent AADs, each associated with a grey error bar (± standard deviation). As the number of votes increases, the AAD gradually decreases. Data smoothing is also applied to the results by using the simple moving average with a sliding window size of 1 (the blue solid line), 3 (the red solid line), and 5 (the green solid line). It can be observed that all the three lines converge to 0.33 when there are 20 votes. As a result, we argue that
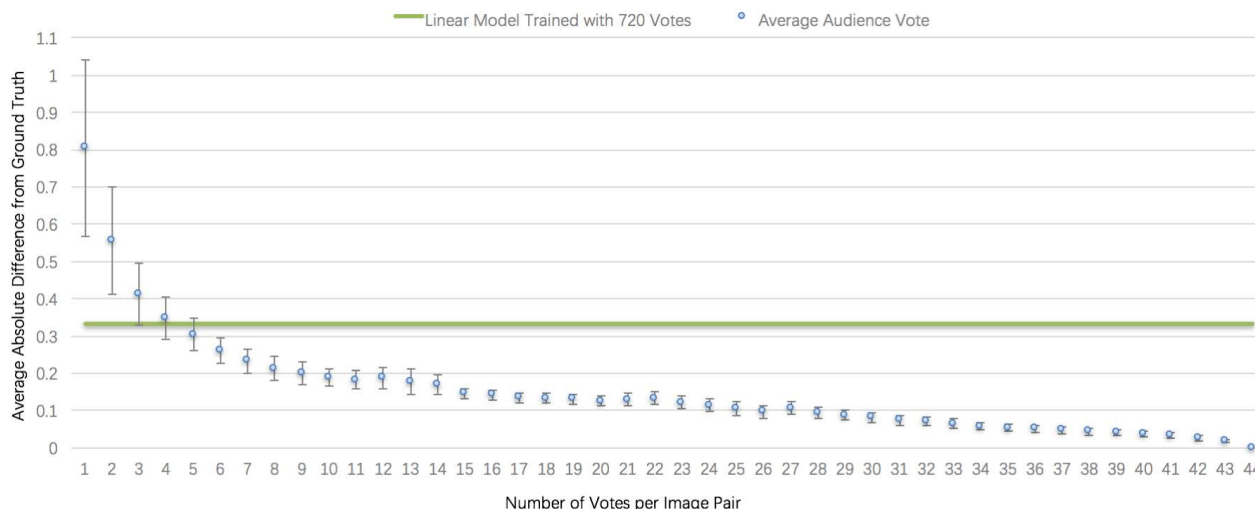
FIGURE 7. Average vote's average absolute differences from the ground truth and their error bars.

there should be at least 720 votes (20 votes × 36 image pairs) for achieving a stable linear-regression model in this task.

D. DISCUSSION ON THE NECESSITY OF LOTTERY

In the previous sub-section, we have illustrated the minimum number of votes to train a stable regression model. A question then arises: once a model is built, do we still need to use Lottery to collect similarity data for new image pairs or should we just simply use the trained model to obtain their similarity values? To answer this, we conduct an analysis similar to V-C, but replacing the linear regression model with the average vote at each point on the x-axis.

Fig. 7 shows AADs of average votes and their associated error bars. In this figure, the AAD of the linear model trained with 720 votes, 0.33 discussed in the previous sub-section, is also overlaid as the green horizontal line. It can be seen that the linear model outperforms the average vote up to and including the point where each image pair has four votes. This indicates that t in (2) should be set to 5. In addition, the results here confirm that Lottery is still needed for collecting votes, at least five votes per image pair, to obtain reliable similarity data, used for example in building an image recommender system.

E. SUMMARIZED FLOW FOR OTHER APPLICATIONS

Below is a general procedure to apply our approach to other systems.

1) Pre-collect a certain amount of human similarity data of the target application.
2) Train and evaluate several regression models with the data (cf. **Table 2**).
3) Select the best model for being used in (2).
4) Derive the minimum number of votes to make the selected model stable (cf. **Fig. 6**).

5) Derive the number of votes per pair to switch from the improved deep-feature similarity, obtained from the trained model, to the similarity based on the votes received so far (cf. **Fig. 7**).

VI. CONCLUSION

This study proposed using a regression model built with human-perceived similarity data (pre-collected from other images in the same image category) to improve the existing similarity calculation of images, used in a recently developed APGWAP called Lottery. We then investigated a question of how much human-perceived similarity data should be obtained to make the model stable. We also found that the regression model in use can only make the calculated similarity close to the human-perceived similarity, but there is still noticeable difference. This indicates the importance of Lottery in its purpose of collection of human-perceived similarity data of images.

Nevertheless, this article does not discuss the aspect of gaming experience of human audiences. This will be the direction of our future work. We will also apply the proposed approach for improving the calculation of image similarity in other applications.

REFERENCES

[1] S. Glickman, N. McKenzie, J. Seering, R. Moeller, and J. Hammer, "Design challenges for livestreamed audience participation games," in *Proc. Annu. Symp. Comput.-Hum. Interact. Play Extended Abstr. (CHI PLAY)*, Oct. 2018, pp. 187–199.

[2] Y. Jiang, P. Paliyawan, R. Thawonmas, and T. Harada, "An audience participation angry birds platform for social well-being," in *Proc. 19th Annu. Eur. Game-On Conf. Simulation AI Comput. Games (GAMEON)*, 2018, pp. 1–6.

[3] M. Bell, S. Reeves, B. Brown, S. Sherwood, D. MacMillan, J. Ferguson, and M. Chalmers, "EyeSpy: Supporting navigation through play," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst. (CHI)*, Apr. 2009, pp. 123–132.

[4] S. Frazier and M. Riedl, "Persistent and pervasive real-world sensing using games," in *Proc. 2nd AAAI Conf. Hum. Comput. Crowdsourcing*, 2014, pp. 16–17.

[5] N. C. Nguyen, R. Thawonmas, P. Paliyawan, and H. V. Pham, "JUSTIN: An audience participation game with a purpose for collecting descriptions for artwork images," in *Proc. IEEE Conf. Games (CoG)*, Aug. 2020, pp. 344–350.

[6] Z. Wei, P. Paliyawan, and R. Thawonmas, "A card game for collecting human-perceived similarity data of artwork images," *IEEE Access*, vol. 10, pp. 8103–8111, 2022.

[7] Z. Wei, S. Wang, P. Paliyawan, and R. Thawonmas, "Difference in perceived similarity between humans and machines: A case study on Ukiyo-e artworks," *ART Res.*, vol. 22, p. 8, 2022.

[8] B. David, "Ukiyo-e explained," Univ. Michigan, Kent, Global Oriental, Folkestone, U.K., Tech. Rep., Oct. 2021, pp. 1–403. [Online]. Available: https://brill.com/view/title/19342

[9] S. Kikuchi, *A Treasury of Japanese Wood Block Prints (Ukiyo-e)*. New York, NY, USA: Crown Publishers, 1969, p. 31.

[10] K. Li, B. Batjargal, A. Maeda, and R. Akama, "Artwork information embedding framework for multi-source Ukiyo-e record retrieval," in *Digital Libraries at Times of Massive Societal Transition*. Cham, Switzerland: Springer, 2020, pp. 255–261. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-64452-9_23#citeas

[11] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2414–2423.

[12] W. T. Chu and Y. L. Wu, "Deep correlation features for image style classification," in *Proc. ACM Multimedia Conf. (MM)*, 2016, pp. 402–406.

[13] S. Matsuo and K. Yanai, "CNN-based style vector for style image retrieval," in *Proc. ACM Int. Conf. Multimedia Retr. (ICMR)*, Jun. 2016, pp. 309–312.

[14] Z. Wei, L. Xiong, K. Mori, T. D. Nguyen, T. Harada, R. Thawonmas, K. Suzuki, and M. Kidachi, "Deep features for image classification and image similarity perception," in *Proc. Jpn. Assoc. Digit. Hum. Conf. (JADH)*, 2017, pp. 60–62.

[15] Y. Zhang and T. Yamasaki, "Style-aware image recommendation for social media marketing," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 3106–3114.

[16] N. Kayhan and S. Fekri-Ershad, "Content based image retrieval based on weighted fusion of texture and color features derived from modified local binary patterns and local neighborhood difference patterns," *Multimedia Tools Appl.*, vol. 80, nos. 21–23, pp. 32763–32790, Sep. 2021.

[17] D. Pathak and U. S. N. Raju, "Content-based image retrieval for super-resoluted images using feature fusion: Deep learning and hand crafted," *Concurrency Comput., Pract. Exp.*, pp. 1–25, Jan. 2022. [Online]. Available: https://onlinelibrary.wiley.com/action/doSearch?SeriesKey=15320634&sortBy=Earliest&startPage=8&pageSize=20, doi: 10.1002/cpe.6851.

[18] X. Lu, L. Zhang, L. Niu, Q. Chen, and J. Wang, "A novel adaptive feature fusion strategy for image retrieval," *Entropy*, vol. 23, no. 12, pp. 1–15, 2021, doi: 10.3390/e23121670.

[19] N. Fumo and M. A. R. Biswas, "Regression analysis for prediction of residential energy consumption," *Renew. Sustain. Energy Rev.*, vol. 47, pp. 332–343, Jul. 2015.

[20] R. Pino-Mejías, A. Pérez-Fargallo, C. Rubio-Bellido, and J. A. Pulido-Arcas, "Comparison of linear regression and artificial neural networks models to predict heating and cooling energy demand, energy consumption and CO_2 emissions," *Energy*, vol. 118, pp. 24–36, Jan. 2017.

[21] S. M. Ayyoubzadeh, S. M. Ayyoubzadeh, H. Zahedi, M. Ahmadi, and S. R. N. Kalhori, "Predicting COVID-19 incidence through analysis of Google trends data in Iran: Data mining and deep learning pilot study," *JMIR Public Health Surveill.*, vol. 6, no. 2, pp. 1–6, 2020. [Online]. Available: https://publichealth.jmir.org/2020/2/e18828/

[22] Y. Chen, G. Peng, Z. Zhu, and S. Li, "A novel deep learning method based on attention mechanism for bearing remaining useful life prediction," *Appl. Soft Comput.*, vol. 86, pp. 1–11, Jan. 2020.

[23] S. Jiao, Y. Gao, J. Feng, T. Lei, and X. Yuan, "Does deep learning always outperform simple linear regression in optical imaging?" *Opt. Exp.*, vol. 28, no. 3, pp. 3717–3731, 2020.

[24] J. Seering, S. Savage, M. Eagle, J. Churchin, R. Moeller, J. P. Bigham, and J. Hammer, "Audience participation games: Blurring the line between player and spectator," in *Proc. Conf. Designing Interact. Syst.*, Jun. 2017, pp. 429–440.

[25] D. Ramirez, J. Saucerman, and J. Dietmeier, "Twitch plays Pokemon: A case study in big G games," in *Proc. Digit. Games Res. Assoc. (DiGRA)*, 2014, pp. 3–12.

[26] L. von Ahn and L. Dabbish, "Designing games with a purpose," *Commun. Assoc. Comput. Mach.*, vol. 51, no. 8, pp. 58–67, 2008.

[27] L. Barrington, D. Turnbull, and G. Lanckriet, "Game-powered machine learning," *Proc. Nat. Acad. Sci. USA*, vol. 109, no. 17, pp. 6411–6416, Apr. 2012.

[28] E. Law and L. von Ahn, "Input-agreement: A new mechanism for collecting data using human computation games," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst. (CHI)*, Apr. 2009, pp. 1197–1206.

[29] C. J. Lintott, K. Schawinski, A. Slosar, K. Land, S. Bamford, D. Thomas, M. J. Raddick, R. C. Nichol, A. Szalay, D. Andreescu, P. Murray, and J. Vandenberg, "Galaxy zoo: Morphologies derived from visual inspection of galaxies from the sloan digital sky survey," *Monthly Notices Roy. Astron. Soc.*, vol. 389, no. 3, pp. 1179–1189, Sep. 2008.

[30] T. Kathleen, B. Rahul, S. Noah, P. Jovan, and P. Zoran, "PointCraft: Harnessing players' FPS skills to interactively trace point clouds in 3D," in *Proc. 10th Int. Conf. Found. Digit. Games*, 2015, pp. 1179–1189.

[31] K. Tuite, N. Snavely, D.-Y. Hsiao, N. Tabing, and Z. Popovic, "PhotoCity: Training experts at large-scale image acquisition through a competitive game," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst. (CHI)*, May 2011, pp. 1383–1392.

[32] E. Law and L. von Ahn, "Human computation," *Synth. Lectures Artif. Intell. Mach. Learn.*, vol. 5, no. 3, pp. 1–121, 2011.

[33] K. Siu, A. Zook, and M. O. Riedl, "A framework for exploring and evaluating mechanics in human computation games," in *Proc. 12th Int. Conf. Found. Digit. Games (FDG)*, Aug. 2017, pp. 1–4.

[34] Z. Wei, Y. Ma, S. Wang, N. C. Nguyen, P. Paliyawan, R. Thawonmas, T. Harada, K. Suzuki, and M. Kidachi, "Establishment of an Ukiyo-e similarity database by a game with a purpose," in *Proc. Jpn. Assoc. Digit. Hum. Conf. (JADH)*, 2019, pp. 55–58.

ZHENAO WEI is currently pursuing the Ph.D. degree with the Graduate School of Information Science and Engineering, Ritsumeikan University, Japan. His research interests include game AIs and recommender systems.

PUJANA PALIYAWAN received the D.Eng. degree from the Graduate School of Information Science and Engineering, Ritsumeikan University, Japan. He is currently a Senior Researcher with the Research Organization of Science and Technology, Ritsumeikan University. His research interests include AI, HCI, and health monitoring systems.

RUCK THAWONMAS (Senior Member, IEEE) received the B.Eng. degree in electrical engineering from Chulalongkorn University, Bangkok, Thailand, in 1987, the M.Eng. degree in information science from Ibaraki University, Hitachi, Japan, in 1990, and the Ph.D. degree in information engineering from Tohoku University, Sendai, Japan, in 1994. He is currently a Full Professor with the College of Information Science and Engineering, Ritsumeikan University, Japan, where he is leading the Intelligent Computer Entertainment Laboratory, with more than 40 laboratory graduates, working in game industry. He has published more than 250 peer-reviewed papers in both Japanese and English. His current research interests include games for health and for humanities. He is a member of the Ritsumeikan Center for Game Studies. He was the General Chair of the 2020 IEEE Conference on Games. He is currently the Founding Chair of Entertainment and Gaming Technical Committee, IEEE Consumer Technology Society; and an Associate Editor of IEEE TRANSACTION ON GAMES and *Games for Health Journal*.

• • •