Does Machine Learning Replicate the Uncanny Valley? An Example using FaceNet

Taku Imaizumi (taku-imaizumi605@g.ecc.u-tokyo.ac.jp)
Graduate School of Interdisciplinary Information Studies, The University of Tokyo
7-3-1, Hongo, Bunkyo-Ku, Tokyo 113-0033, Japan

Lu Li (2000lilu0317@g.ecc.u-tokyo.ac.jp)
Graduate School of Interdisciplinary Information Studies, The University of Tokyo
7-3-1, Hongo, Bunkyo-Ku, Tokyo 113-0033, Japan

Kazuhiro Ueda (ueda@g.ecc.u-tokyo.ac.jp)
Graduate School of Arts and Sciences, The University of Tokyo
3-8-1, Komaba, Meguro-Ku, Tokyo 153-0902, Japan

Abstract

Androids that strongly but imperfectly resemble humans in shape can elicit negative emotions in people, a phenomenon known as the "uncanny valley," which has been replicated in laboratory experiments. Recently, the accuracy of face recognition utilizing machine learning has increased, raising the question of whether machine learning can replicate the uncanny valley effect. In this study, using FaceNet as a representative face recognition algorithm, we examined the similarity of face recognition to human evaluation and its replication of the uncanny valley. The results revealed a strong correlation between machine learning and human evaluation of human-like shapes. However, because the evaluations recorded were significantly disparate for some objects, it is evident that only certain aspects of the uncanny valley were replicated. Furthermore, visualization of the activation maps suggests that localized regions, such as the mouth and chin, acted as the basis for judgment. These findings support the idea that human and machine learning have distinct areas of attention, as well as the categorization ambiguity hypothesis, and perceptual mismatch hypothesis in the study of the uncanny valley effect.

Keywords: face recognition; machine learning; FaceNet; uncanny valley; Grad-CAM

Introduction

Uncanny Valley

When an artifact, such as a robot or agent, resembles but does not fully emulate the human form, the viewer may have a negative perception of it. This phenomenon, referred to as the "uncanny valley" (Mori, 1970; Mori, MacDorman & Kageki, 2012) has been considered a challenge to be overcome in facilitating communication between humans and robots (MacDorman et al., 2005). Understanding the mechanism of the uncanny valley contributes not only to the domains of robotics and human-agent interaction by fostering the creation of favorable agents, but also to cognitive science by facilitating laboratory experiments using such agents (Piwek, McKay & Pollick, 2014; de Borst & de Gelder, 2016).

The uncanny valley has been chiefly investigated from two perspectives: the categorization ambiguity hypothesis, and the perceptual mismatch hypothesis (Kätsyri, Förger, Mäkäräinen & Takala, 2015). The categorization ambiguity hypothesis posits that the uncanny valley arises due to an aversion to objects that straddle the boundary between human and artifact, while the perceptual mismatch hypothesis contends that negative affinity is caused by an inconsistency between the human-likeness levels of specific sensory cues.

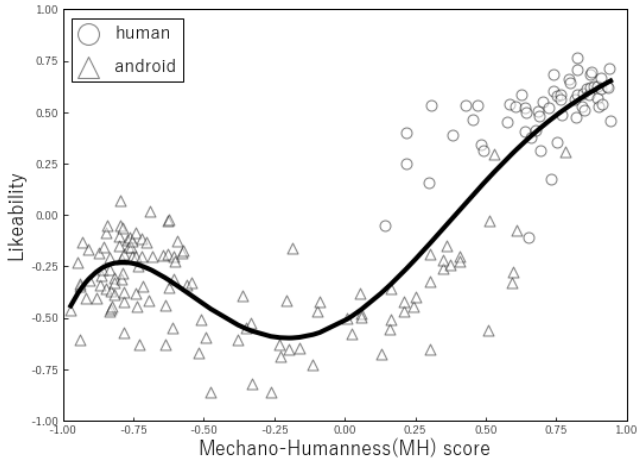

Figure 1: Reproduction of the uncanny valley as reported by Mathur et al. (2020). Fitting and charting were performed by the authors.

Although the uncanny valley theory is not experimentally proposed, several empirical studies have confirmed its existence (Seyama & Nagayama, 2007; Mathur & Reichling, 2016; Mathur et al., 2020). Mathur et al. (2020) assessed the similarity to humans and likability of 182 facial images (robot:122, human:60) in a questionnaire by plotting the similarity to humans on the horizontal axis, and likability on the vertical axis (Figure 1). The horizontal axis of the figure shows that the closer to 1, the more human-like an image was judged to be, and the closer to -1, the more machine-like it was judged to be. When the approximation curves were plotted, a valley effect was clearly observed and notably more pronounced on the mechanical than on the human-like side.

Previous studies that have demonstrated the uncanny valley effect were conducted using an experimental paradigm in which participants were asked to respond to the degree of

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

human similarity, and then provide assessments of the image's likability. Consequently, it is possible that the level of likability assigned to these images was influenced by their degree of similarity to humans. Therefore, eliciting separate responses based on the degree of similarity to humans and the degree of likability is a clear limitation of previous studies.

Furthermore, because similarity to humans is a subjective evaluation, it is possible that factors other than shape, such as knowledge, could influence the outcome. Hence, the categorization ambiguity hypothesis can be examined more rigorously by measuring the degree of similarity to humans, without involving subjective evaluation.

Machine Learning and Face Recognition

In this study, a face recognition algorithm was employed to evaluate similarity of images of robots or agents to humans without involving subjective evaluation. A face recognition algorithm is a technique used for determining whether an object in an image or video is a human face and is utilized in cameras, security, and other applications. In recent years, deep learning-based algorithms, such as FaceNet (Schroff, Kalenichenko & Philbin, 2015), have achieved high accuracy levels.

Therefore, the questions arise, when face recognition algorithms are utilized to evaluate the degree of similarity to humans; is it possible to replicate the uncanny valley effect, as demonstrated in previous studies? Face recognition algorithms learn and make judgments based on a vast number of images. In what respects do their judgments concur with those made by humans, and in which do they diverge?

Hypothesis

This study investigates two hypotheses. The first examines whether the evaluation of shapes by humans and face recognition algorithms are congruent. Given that recent face recognition algorithms have proven to be highly accurate, it is hypothesized that this consistency of evaluation will be attained.

The second hypothesis is whether an uncanny valley effect can be observed when the shape evaluation by the face recognition algorithm is plotted on the horizontal axis, and the likability evaluation by humans is plotted on the vertical axis on the graph. While this remains speculative, the uncanny valley may not be replicated because of the presence of certain images in which the human evaluation and face recognition algorithm's evaluation differ significantly.

If this discrepancy exists, the human and algorithm's evaluation criteria may diverge. Therefore, if the uncanny valley is not replicated, visualization of the activation map of the face recognition algorithm (a detailed description of the visualization method will be provided in the Method section) will be implemented to investigate the possibility that the focus area differs between human and machine learning. By visualizing the activation map, it is possible to examine the differences in cues for judging robots and humans, without intervening in subjective evaluation, thus contributing to the discussion of the perceptual mismatch hypothesis.

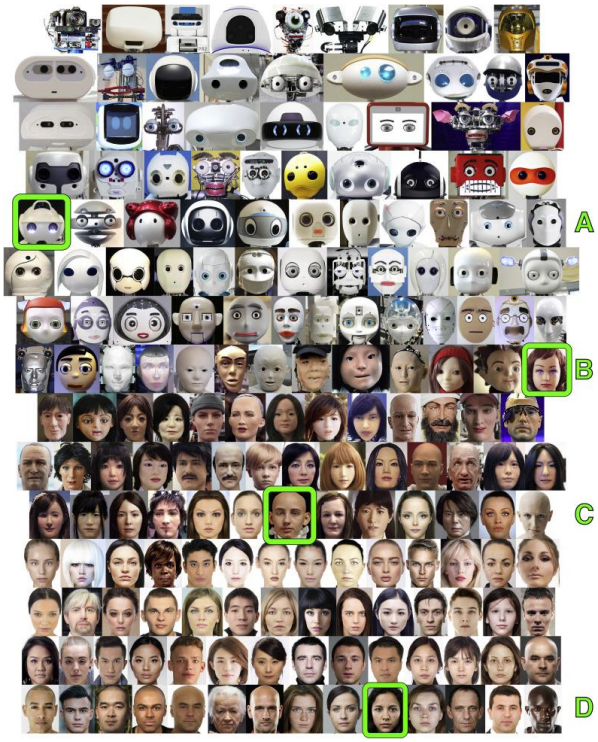

Figure 2: All face images of Validated Face Corpus in ascending order of mean MH score. Boxed faces are those with MH scores closest to the MH scores associated with: (A) the initial likability apex of the Uncanny Valley curve (estimation described in Section 3); (B) the likability low point of the Uncanny Valley; (C) the robot/human category boundary (estimation described in Section 4); and (D) the final apex of likability.

Figure 3: Sample output from Grad-CAM. Redder colors are more important for the decision, bluer colors are less important.

Method

Face Recognition Algorithm : FaceNet

In this study, the face recognition algorithm, FaceNet (Schroff et al., 2015), which is trained using the VGGFace2 dataset (Cao et al., 2018) comprising over 3 million face images, was utilized. FaceNet outputs 512-dimensional feature vectors, which enable the calculation of Euclidean distances between multiple images, making it an apt choice for this study.

Subsequently, dimensionality compression was executed through principal component analysis, with the first principal component defined as the FaceNet score.

Image Data Sets

The Validated Face Corpus (Mathur et al., 2020) was utilized as the image dataset to be evaluated by FaceNet. This corpus comprised 182 face images (122 robot images and 60 human images, as depicted in Figure 2), for which the similarity to humans (mechano-humanness (MH) score, with -1 being machine-like and +1 human-like) and likability (on a scale of -100 to +100, standardized with 0 as the mean) scores were recorded. The images and ratings were identical to those used by Mathur et al. (2020). In addition to the image, their MH score and likability were also analyzed in this study.

Although the size of the images in the Validated Face Corpus is not constant, in this study, the images were resized to 160 pixels by 160 pixels by using the OpenCV resize function to facilitate visualization through the following steps.[1]

Activation Mapping Algorithm: Grad-CAM

The Grad-CAM method (Selvaraju, 2017) was employed to visualize the basis of the decisions made by the FaceNet algorithm. Grad-CAM is a heatmap-based local explanation technique for CNN (convolutional neural network) based image-recognition models and their respective inputs. As shown in Figure 3, the closer the color is to red, the more it influences the prediction. Because of the ability of CNNs to extract features while preserving location information, the data in the final layer could be used to determine which regions of the image were most influential in the prediction.

The color of each pixel $C_{ij}(1 \leq i \leq 160, 1 \leq j \leq 160)$ was computed using the following formula, which applied the feature vector $k(1 \leq k \leq 512)$ dimensions, factor loadings of the first principal component a_{1k}, and Grad-CAM output value g_{kij}.

$$C_{ij} = \sum_{k=1}^{512} a_{1k} g_{kij}$$

Result

Correlation between MH Score and FaceNet Score

The correlation of each principal component with the MH score is illustrated in Figure 4, with the MH score and FaceNet score (first 23 principal components) for each image depicted in Figure 5. Because some principal components did not satisfy the normality assumption, the Spearman's rank correlation coefficient was used.

The FaceNet score (the first principal component) exhibits a notably robust correlation with the MH score in comparison

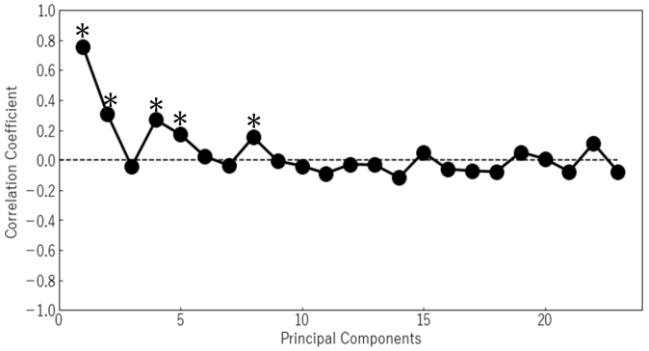

Figure 4: Spearman's rank correlation coefficient between each principal component and MH score. Principal components are illustrated up to the 23rd component, which has a cumulative contribution of 80%. Significant differences are shown, *p<.05.

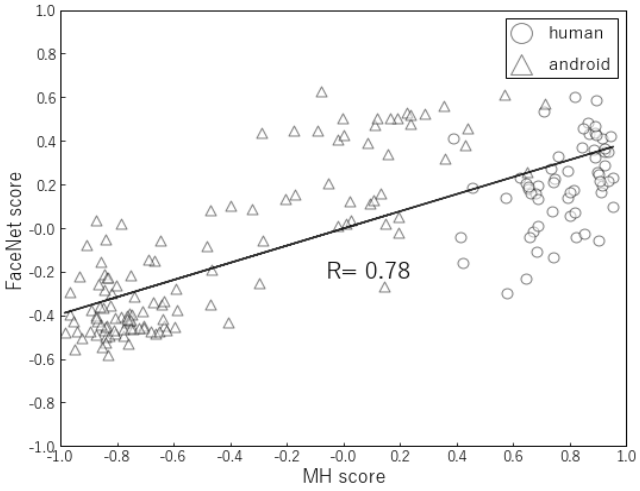

Figure 5: A plot of the MH score and FaceNet score (first principal component). The solid line is the regression line with all images as independent variables. Within the dataset, human face images are plotted as triangles, and android face images are plotted as circles (as in subsequent figures and tables).

to the other principal components (as evidenced by a Spearman's rank correlation coefficient of R = 0.78), indicating that the first principal component can be understood as a measure of similarity with humans in terms of shape.

The strong correlation between the MH score and FaceNet score suggests that the shape evaluations conducted by humans and FaceNet are congruent.

Reproduction of the Uncanny Valley

The trajectory of the Uncanny Valley was determined using ordinary least squares models to regress likability onto polynomial terms of the FaceNet score. Utilizing Akaike's

[1] Resizing images is an established technique for improving the accuracy of face recognition (Dharavath, Talukdar & Laskar, 2014). This preprocessing method is consistent to those used in Sequeira et al. (2021) and Kim, Yun & Ro (2022). These studies, comparable to this study, examined face recognition algorithms (CNN) trained on the VGGFace2 dataset as training data and employed Grad-CAM.

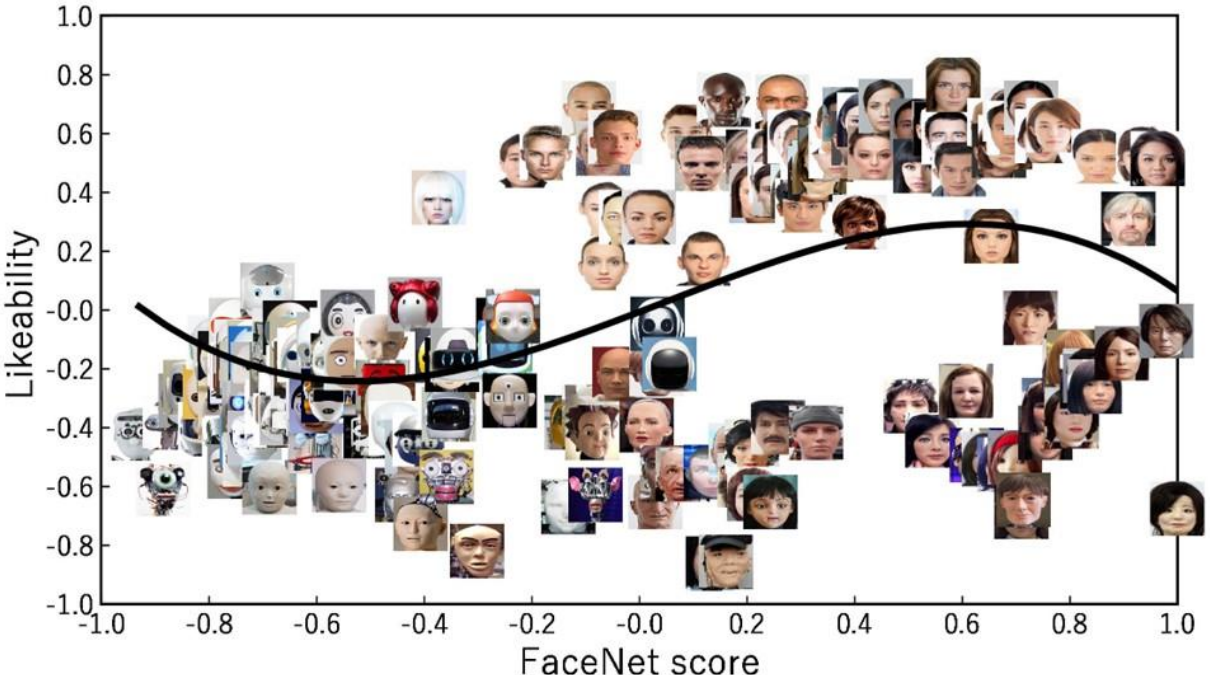

Figure 6: Fitting FaceNet score and Likability. Since the target face images are plotted, some of the plots overlap each other (see Figure 9 for detailed locations). The fittings of this study are illustrated as solid lines.

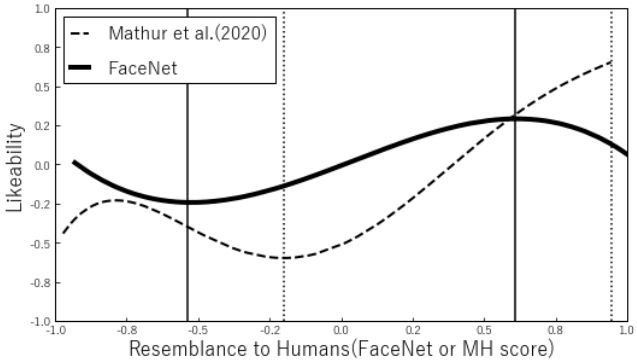

Figure 7: Comparison of the fittings of the uncanny valley. The fittings of this study are illustrated as solid lines and the results of Mathur et al. (2020) are illustrated as dotted lines. The vertical lines indicate the minimum and maximum values of the fittings.

Information Criterion (AIC),[2] the most suitable, lowest-order polynomial model was selected. The graph illustrating the relationship between the FaceNet score and likability is presented in Figure 6. A comparison of the fitting curves from this study with those of Mathur et al. (2020) is shown in Figure 7, and the minimum and maximum values of the fitting curves are presented in Table 1.

The fitting curve in this study diverges from that of Mathur et al. (2020) in the following three respects. First, the absolute minimum and maximum values are relatively insignificant (resulting in a relatively flat graph); second, the right

Table 1: Comparison of two fittings. The second and fourth lines show the x-coordinates of Figure 7, where the maximum and minimum values were recorded, respectively.

	This study	Mathur et al. (2020)
Minimum Likability	-0.243	-0.598
- FaceNet or MH score	-0.536	-0.198
Maximum	0.291	0.652
- FaceNet or MH score	0.607	0.943

endpoint is not the maximum value; and third, this fitting curve exhibits a single peak, whereas Mathur & Reichling (2016) and Mathur et al. (2020) posited the presence of two peaks.[3] The sole point of convergence between this fitting curve and the prior study is the monotonous progression from the minimum to the maximum value. From this analysis, it can be inferred that only a subset of the features associated with the uncanny valley were replicated by utilizing the FaceNet score.

To investigate the reason for the partial replication of the uncanny valley, clustering was conducted as a post-hoc analysis. The results of the silhouette method indicated that four clusters were optimal (as depicted in Figure 8); subsequently, the K-means method was employed for classification (Figure 9 illustrates the clustering, and Table 2 provides the characteristics of each cluster). As a result, a group with a high FaceNet score and low MH score (group 3) was identified. Group 3 is composed of android images and is a group that was not included in the fitting of Mathur

[2] This methodology is consistent with that Mathur et al. (2020).

[3] Of course, the first hypothesis by Mori (1970) also proposed two peaks graph.

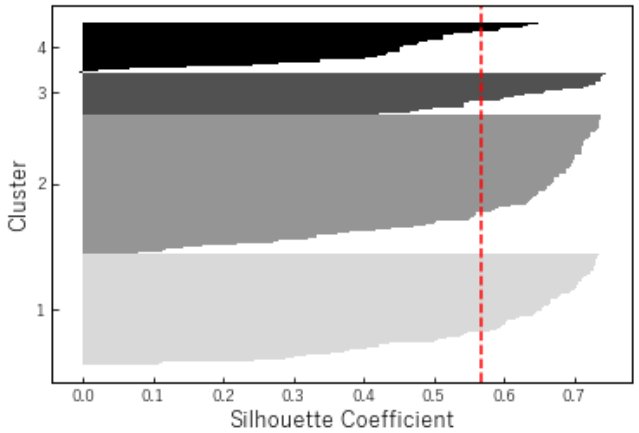

Figure 8: Selection of number of clusters by silhouette method. The dotted line indicates the silhouette coefficient. Since several samples in each group show higher values than the silhouette coefficient, four groups are chosen as appropriate.

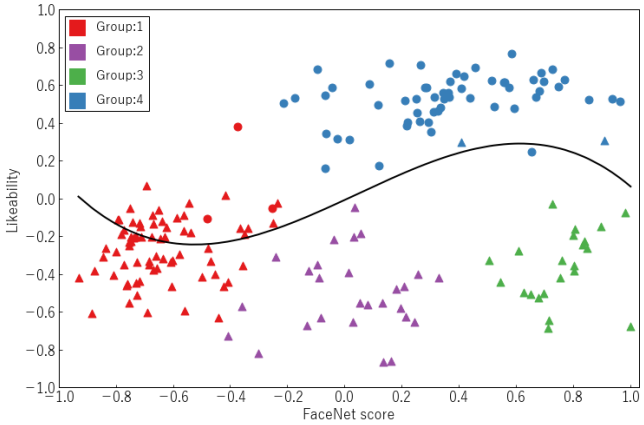

Figure 9: Cluster separation by K-means method. The position of the plots and the fitting curves are the same as in Figure 6. Within the dataset, human face images are plotted as triangles, and android face images are plotted as circles.

et al. (2000) (as shown in Figure 2). In addition, this group is not found in Mathur & Reichling (2016) and Seyama & Nagayama (2007), which also reproduced the uncanny valley.

This suggests that only a portion of the uncanny valley was replicated in some of the android images due to the discrepancy between the FaceNet ratings and MH scores.

Activation map

The average activation map for each cluster is presented in Figure 10. Groups 1 and 2 share a commonality in their focus on the central portion of the display, whereas Groups 3 and 4 share a commonality in the utilization of the lower section of the display as a basis for their decision-making. Groups 3 and 4, which possess high FaceNet scores (see Table2), exhibit a greater likelihood as compared to Groups 1 and 2, which have low FaceNet scores. Since each face image in the dataset was all taken from the front view, the lower part of the screen corresponds to the nose, mouth, and chin. This finding

Table 2: Information on each group

group	Android Image Percentage	FaceNet score		Likability	
		M	*SD*	*M*	*SD*
1	95.9	-0.632	0.16	-0.257	0.18
2	100	0.018	0.19	-0.504	0.21
3	100	0.754	0.13	-0.354	0.19
4	3.39	0.374	0.29	0.520	0.13

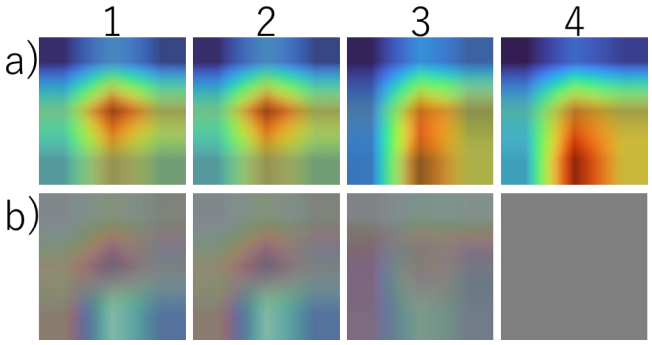

Figure 10: Average activation map for each cluster.

Figure 10a illustrates that redder colors are more important for the decision, and bluer colors less important. Figure 10b illustrates the differences for Figure 10a with respect to group 4. The darker red color indicates more attention paid compared to group 4, and the darker blue color indicates less attention paid compared to group 4.

suggests the possibility that FaceNet utilizes the nose, mouth, and chin as the foundation for its evaluation of similarity to humans.

Discussion

In this study, two hypotheses were investigated: whether similar results could be attained between human and machine learning in judging the similarity of shapes and whether the uncanny valley could be replicated using machine learning evaluations. A significant correlation was observed between the MH score, which is a questionnaire-based evaluation, and the FaceNet evaluation (FaceNet score), suggesting that the evaluations were congruent between humans and machine learning. However, as high FaceNet scores were recorded for some images that were rated as dissimilar to humans in terms of the questionnaire, the uncanny valley was only partially replicated for some features when FaceNet scores were plotted on the horizontal axis and likability by the questionnaire on the vertical axis. Post-hoc analysis of the activation map indicated that the images rated with high FaceNet scores were based on the lower part of the face (e.g., the nose, mouth, and chin).

It is well established that humans tend to direct their gaze towards the eyes and nose when viewing faces (Hsiao & Cottrell, 2008). Therefore, it is plausible that the uncanny valley effect was not replicated because of the dissimilarity in the information used by humans and machine learning. Specifically, machine learning may rely on localized parts of the lower face, such as the mouth and chin, to make

judgments, whereas humans employ wide-area facial information by positioning the vantage point at the center of the face. This may lead to a discrepancy between the human-like shape inferred from machine learning with huge number of images and that inferred by humans. Future research should endeavor to investigate the types of information that humans utilize to judge human-like features.

Contributions to the Uncanny Valley Study

Previous studies have put forth the classification ambiguity hypothesis and the perceptual mismatch hypothesis as potential causes of the uncanny valley (Kätsyri, Förger, Mäkäräinen & Takala, 2015).

As demonstrated in Table 2, among groups 1, 2, and 3, which comprise android images, groups 2 and 3 possess higher FaceNet scores than group 1, yet lower likability. This suggests a potential aversion to images that can be classified as human-like from the perspective of machine learning trained with huge number of images, thus lending support to the classification ambiguity hypothesis.

Based on the activation map, it can be inferred that group 3 comprises robots that possess human-like cues (nose, mouth, and chin). The low likability of group 3, where the machine learning categorization does not align with the actual categories, supports the perceptual mismatch hypothesis, which posits that the inconsistency between human-likeness levels of specific sensory cues produces a negative inclination towards the robot.

However, it is important to exercise caution when interpreting these findings, as the FaceNet algorithm used in this study is based on a pre-trained model utilizing a dataset of human face images.

Differences between Machine Learning and Human

While a strong correlation was observed between the MH and FaceNet scores, the activation map generated by Grad-CAM suggests that the basis for judgment may diverge substantially between machine learning and humans. This implies that even if the outputs of human and machine learning are similar, the judgment process may not necessarily be the same. In addition to this study, it is expected that more cognitive science research utilizing machine learning as a method that does not involve subjective evaluation will be conducted in the near future. Therefore, it is important to exercise some caution with regard to different criteria for decision making between machine learning and humans.

It is also worth noting that some studies have pointed out that Grad-CAM and other explanatory methods have certain limitations (e.g., Heo, Joo & Moon, 2019). In this study, Grad-CAM was employed as a prominent algorithm; however, future research is needed to determine whether similar results can be obtained using other explanatory methods.

Conclusion

In this study, a machine learning algorithm (FaceNet) was employed to evaluate an android face dataset and investigate whether the uncanny valley could be replicated. Although there was a strong correlation between the shape evaluation by machine learning and that by humans, the evaluations were significantly disparate for some images, resulting in a partial reproduction of the uncanny valley effect. Visualization of the activation map of FaceNet revealed that the nose, mouth, and chin were utilized as decision criteria, suggesting that humans and machine learning may have different areas of focus, which can be considered as a topic for future research.

Acknowledgments

This study was supported by JST CREST [JPMJCR19A1], JSPS KAKENHI [JP22H03911], and the Telecommunications Advancement Foundation.

References

Cao, Q., Shen, L., Xie, W., Parkhi, O. M., & Zisserman, A. (2018). Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)* (pp. 67-74). IEEE.

de Borst, A. W., & de Gelder, B. (2015). Is it the real deal? Perception of virtual characters versus humans: an affective cognitive neuroscience perspective. *Frontiers in psychology*, *6*, 576.

Dharavath, K., Talukdar, F. A., & Laskar, R. H. (2014). Improving face recognition rate with image preprocessing. *Indian Journal of Science and Technology*, *7*(8), 1170-1175.

Kätsyri, J., Förger, K., Mäkäräinen, M., & Takala, T. (2015). A review of empirical evidence on different uncanny valley hypotheses: support for perceptual mismatch as one road to the valley of eeriness. *Frontiers in psychology*, *6*, 390.

Kim, H. I., Yun, K., & Ro, Y. M. (2022). Face Shape-Guided Deep Feature Alignment for Face Recognition Robust to Face Misalignment. IEEE Transactions on Biometrics, Behavior, and Identity Science, 4(4), 556-569.

Heo, J., Joo, S., & Moon, T. (2019). Fooling neural network interpretations via adversarial model manipulation. *Advances in Neural Information Processing Systems*, *32*.

Hsiao, J. H. W., & Cottrell, G. (2008). Two fixations suffice in face recognition. *Psychological science*, *19*(10), 998-1006.

MacDorman, K. F., Minato, T., Shimada, M., Itakura, S., Cowley, S., & Ishiguro, H. (2005). Assessing human likeness by eye contact in an android testbed. *Proceedings of the XXVII annual meeting of the cognitive science society*, (pp. 21-23).

Mathur, M. B., & Reichling, D. B. (2016). Navigating a social world with robot partners: A quantitative cartography of the Uncanny Valley. *Cognition*, *146*, 22-32.

Mathur, M. B., Reichling, D. B., Lunardini, F., Geminiani, A., Antonietti, A., Ruijten, P. A., Levitan, C. A., Nave, G., Manfredi, D., Bessette-Symons, B., Szuts, A., & Aczel, B. (2020). Uncanny but not confusing: Multisite study of perceptual category confusion in the Uncanny Valley. *Computers in Human Behavior*, *103*, 21-30.

Mori, M., (1970). The uncanny valley. *Energy*, *7*(4), 33-35.

Mori, M., MacDorman, K. F., & Kageki, N. (2012). The uncanny valley [from the field]. *IEEE Robotics & automation magazine*, *19*(2), 98-100.

Piwek, L., McKay, L. S., & Pollick, F. E. (2014). Empirical evaluation of the uncanny valley hypothesis fails to confirm the predicted effect of motion. *Cognition*, *130*(3), 271-277.

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 815-823).

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618-626).

Seyama, J., Nagayama, R. S., (2007). The Uncanny Valley: Effect of Realism on the Impression of Artificial Human Faces. *Presence: Teleoperators and Virtual Environments*, *16* (4). 337–351.

Sequeira, A. F., Gonçalves, T., Silva, W., Pinto, J. R., & Cardoso, J. S. (2021). An exploratory study of interpretability for face presentation attack detection. *IET Biometrics*, *10*(4), 441-455.