機械学習は不気味の谷を理解するか：FaceNet を例にして
Does Machine Learning Understand the Uncanny Valley? An Example using FaceNet

今泉 拓†，李 璐†，植田 一博†
Taku Imaizumi, Lu Li, Kazuhiro Ueda
†東京大学
The University of Tokyo
taku-imaizumi605@g.ecc.u-tokyo.ac.jp

概要
近年機械学習を用いた顔認識の精度が向上しているが，機械学習は人と同様に不気味の谷を再現することができるのだろうか。本研究では FaceNet アルゴリズムを用いて検討を行った。その結果，ヒトらしい形状の評価について機械学習と人間の間で強い相関が見られたものの，一部の対象において評価が著しく異なったため，不気味の谷の一部のみが再現された。さらに，機械学習の注目領域を可視化すると，口やあごといった局所的領域が判断の根拠になっていることが示唆された。これらの結果は，人間と機械学習で注目領域が異なる可能性，および不気味の谷研究における分類曖昧性仮説を支持している。

キーワード：不気味の谷, 顔認識, 機械学習, ディープラーニング,FaceNet, Grad-CAM, ロボティクス

1. はじめに

人工物の外見がヒトに類似するにつれて人は好感度を抱くようになるが，実物の一歩手前まで近づくと途端に嫌悪感を抱くようになり，実物と見分けがつかないほどになると一転して好印象を抱くようになる現象は「不気味の谷現象」として知られている[1]。この現象は，実験室実験でも再現されてきた(図 1)[2][3]。

不気味の谷現象が生じる原因として，ヒトと人工物のカテゴリーの境界にあるような対象について嫌悪感を抱くために不気味の谷が生じるという，分類曖昧性(categorization ambiguity)仮説が検討されている[4]。ただし，不気味の谷を実証する先行研究は，提示されたヒトやアンドロイドの顔画像に対して人が好感度等を質問紙で回答するという実験パラダイムで行われている。そのため，好感度を評定する段階で，画像のヒトとの類似度に対する主観評価の影響を受けている可能性を否定できない。

そこで，本研究では主観評価によらないヒトとの類似度の評価方法として，深層学習を用いた顔認識アルゴリズム(FaceNet)[5]を用いることで，分類曖昧性仮説をより厳密に検討した。すなわち，ヒトとの類似度を顔認識アルゴリズムで評価した場合，先行研究と同様に不気味の谷は再現されるのだろうか。具体的には次の 3 点を検討した。

1. 人間による形状の評価と機械学習によるヒトとの類似度の評価に相関が見られるのか。
2. 横軸に FaceNet による形状の評価，縦軸に人による好感度の評価をプロットした場合，不気味の谷は見られるのか。
3. FaceNet の注視領域を可視化した場合，人と FaceNet とでは分類の根拠が異なるのか。

特に,人間と FaceNet で評価基準が異なるために不気味の谷が再現されない可能性が考えられるため，機械学習の判断根拠として注目領域の可視化を行った。

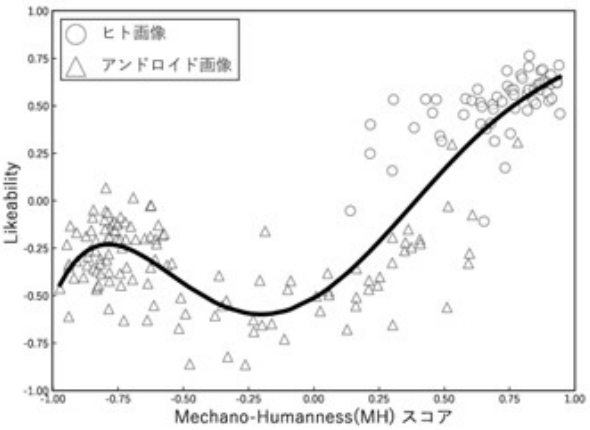

図 1 Mathur et al. (2020) による不気味の谷の再現
フィッティングと図表作成は筆者らが行った。

2. 手法

2.1. 顔認識アルゴリズム

顔認識アルゴリズムとして，FaceNet を用いた。FaceNet は各画像について 512 次元の特徴量ベクトルを算出するため，複数画像間のユークリッド距離の計算を行うことが可能である。本研究では，各画像の特徴量ベクトルを取得した後，主成分分析によって次元圧縮を行った。そのうち，第一主成分を FaceNet スコアと定義した。

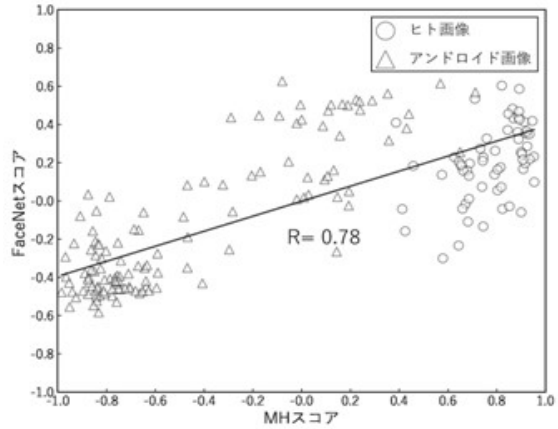

図 2 MH スコアと FaceNet スコアのプロット

データセットのうちヒトの顔画像を三角形で，アンドロイドの顔画像を丸でプロットした(今後の図表でも同様)。MH スコアは，-1 に近いほど機械らしく，1 に近いほどヒトらしいと評価されていることを示す。

2.2. 画像データセット

FaceNet を用いて Validated face corpus(顔画像データセット)を評価した。このコーパスは 182 枚の顔画像(ロボット:122 枚，ヒト 60 枚)と，各画像のヒトとの類似度(MH スコア)と好感度(Likability)の主観評価値が記録されている。

2.3. 注視領域の可視化手法

FaceNet の注目領域の可視化手法として Grad-CAM[6]を用いた。Grad-CAM は CNN ベースの画像認識モデルに対して，ある入力とその予測に対してヒートマップを用いた局所的な説明を与える手法である。

3. 結果

3.1. FaceNet と人間による評価の相関

FaceNet とスコアと MH スコアの散布図を図 2 に示す。FaceNet スコアと MH スコアとの間に強い相関が見られたことから(スピアマンの順位相関係数$R = 0.78$)，人間による形状の評価と FaceNet スコア(第一主成分)は概ね一致していたと言え，FaceNet スコアはヒトとの類似度を表していると解釈できる。

3.2. 不気味の谷の再現

続いて，FaceNet スコアと Likability のフィッティングの結果を図 3 に，本研究と Mathur et al. (2020)のフィッティングの比較を図 4 に示した。先行研究[2][3]や森の仮説[1]では二峰性のグラフになっていたが，本研究のフィッティング曲線は単峰性であった。また，先行研究と同様に，最小値から最大値に向けて単調増加にな

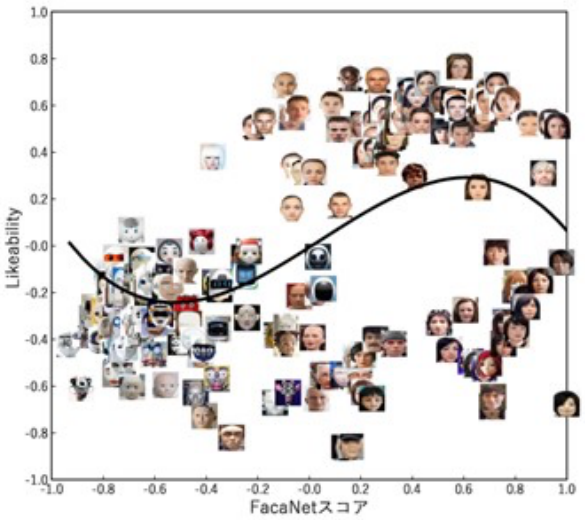

図 3 FaceNet スコアと Likability のフィッティング

Mathur et al. (2020) と同一の方法を用いてフィッティングを行った

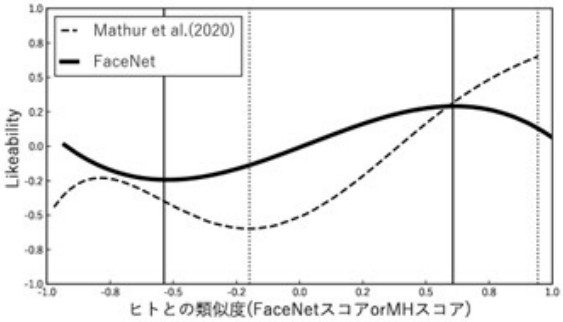

図 4 不気味の谷のフィッティングの比較

本研究のフィッティングを実線で，Mathur et al. (2020) の結果を点線で示す。縦線はフィッティングの最小値と最大値を示す。

ったものの，最小値や最大値は先行研究と比べ小さい値となった。以上から，不気味の谷の一部の特徴のみが再現されたと判断できる。

3.3. 注視領域の可視化

事後分析として，シルエット法に基づき K-means 法を用いて 4 クラスタに分類し(図 5，表 1)，各クラスタについて Grad-CAM による注目領域の可視化を行った(図 6)。図 6 から，FaceNet スコアが高い群 3,4 は，FaceNet スコアが低い群 1,2 と比べて，画像の下方が判断の根拠となっていることがわかる。

4. 考察

注視領域を可視化した結果は，FaceNet が鼻や口，あごを判断根拠としてヒトとの類似度を評価していることを示している。一方で，人が顔を見る際には目や鼻を注視することが知られており[7]，人と機械学習とで

表 1 各クラスタの情報

群	アンドロイド画像(%)	FaceNet スコア		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

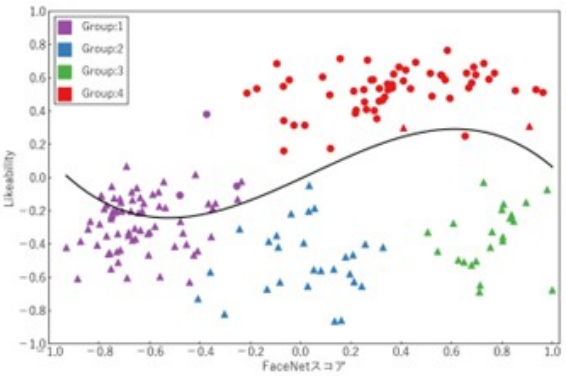

図 5 クラスタ分類

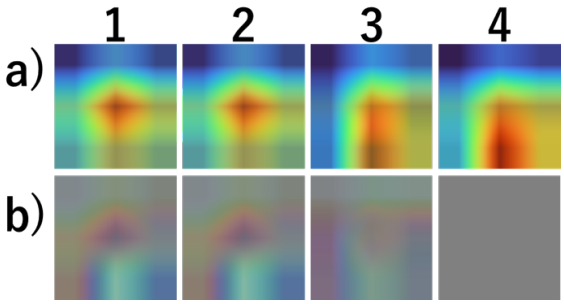

図 6 クラスタごとの注目領域の平均

a)では赤い色ほど判断根拠として強く，青い色ほど判断根拠として弱いことを示している。b)は，a)の各群における群 4 を基準にした差分を示している。具体的には，赤い色が濃いほど群 4 に比べて注目していることを示し，青い色が濃いほど群 4 に比べて注目していないことを示す。

は判断の根拠となる情報が異なるために，不気味の谷が再現されなかった可能性が考えられる。

さらに，アンドロイド画像から構成されている群 1,2,3 において，群 1 は群 2,3 に比べて FaceNet スコアが低いものの，Likability の面ではやや高い値が記録された。このことは，機械学習からみてヒトらしいと評価されるアンドロイド画像は，人からの好感度が低いことを意味している。この結果は，画像特徴量の視点からみてヒトらしいが実際は人工物である画像に対して人は嫌悪感を覚える可能性を示唆しており，機械学習と人間による分類が異なる対象に嫌悪感を抱くという点で分類曖昧性仮説をサポートする結果となっていると考えられる。

文献

[1] Mori, M., (1970). The uncanny valley. *Energy, 7*(4), 33-35.
[2] 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.
[3] 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.
[4] 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.
[5] 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).
[6] Hsiao, J. H. W., & Cottrell, G. (2008). Two fixations suffice in face recognition. *Psychological science, 19*(10), 998-1006.
[7] 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).