Vision Transformer を用いた顔魅力要因の可視化
Visualization of facial attractiveness factors using Vision Transformer

佐野 貴紀†
Takanori Sano
†慶應義塾大学
Keio University
takanori_sano@keio.jp

概要

本研究では，Vision Transformer を用いて顔の魅力を予測するモデルを構築した．構築したモデルの Attention 機構を用いて，魅力の予測において重要な特徴を可視化した．その結果，魅力度の高い顔画像の場合に特に，目の領域が活性する傾向が見られた．この結果は心理学研究で報告されている知見と整合的であった．本アプローチは，顔魅力に関与する特徴の理解に有効であることが示唆された．

キーワード：Vision Transformer, 顔魅力 (facial attractiveness)

1. 序論

顔の魅力は社会的に重要な要素であるため，その要因については多くの研究が行われてきた．対称性，平均性，性的二型性は顔魅力において重要であることが知られており，これらの要因については既に多数の報告がある [1]．しかしながら，実験的手法だけでは顔の魅力に影響を与える要因を網羅的に検討することは困難であるため，近年ではデータ駆動型アプローチ[2]が注目されている．この手法には，3D モデリング[3]，幾何学的形態測定法による分析[4]，画像統計的分析[5]，そして深層学習を用いた解析[6-9]も始まりつつある．例えば，肌，滑らかさ，明るさの情報を用いた畳み込みニューラルネットワーク(CNN)モデルが魅力予測の精度向上に有効であることや[6]，その予測においては目や口などの特徴が魅力予測に重要であることが報告されている[7]．また，筆者は以前の研究においては，CNN と gradient-weighted class activation mapping (Grad-CAM)[8]を用いることで，性的二型性に関連する特徴が重要な特徴として抽出された[9]．

このように，これまでの深層学習を用いた先行研究では，CNN に基づくモデルが主流であったが[6-9]，近年の画像認識の分野では，CNN に代わって Vision Transformer(ViT)モデル[10]が注目されている．CNN は比較的画像のテクスチャ等の情報に注目する一方で[11]，Vision Transformer はより形態的な特徴に注目することが知られている[12]．顔は，肌やテクスチャなどの情報に加え，顔のパーツ，輪郭といったよりまとまりのある特徴によって構成されているため，ViT での顔魅力要因の探索を行うことは有用であると考えられる．そこで本研究は，ViT モデルを用いて顔魅力に重要な特徴を抽出し，心理学的知見との対応を調査することを目的とする．

2. 方法

データセットは SCUT-FBP5500[13]を用いた．このデータセットには，アジア人男性画像 2000 枚，アジア人女性画像 2000 枚，白人女性画像 750 枚，白人男性画像 750 枚が含まれており，各画像には，魅力的かどうかを示す美しさスコアが付与されており，本研究ではこれを魅力の指標として扱った．これらの画像を用いて，ViT モデルを構築し，ベンチマークとして単純な CNN モデルと比較して精度を向上させることができるかどうかを確認した．画像性別による違いを確認するために，画像は性別ごとに分けてモデルを構築した．精度の指標として，画像ラベル値と予測結果のピアソン相関を使用し，3 分割のクロスバリデーション法により検証を行った．ベンチマークモデルには 6 層の CNN モデルを構築し，CNN 層の直後に Batch normalization 層，2 層目ごとに Max pooling 層を設置した．Vit モデルは vit-keras を使用して構築した．各モデルにおいて，最適化関数は SGD に，バッチサイズは 160 に，エポックは 500 に，学習率は 0.005 に固定した．精度検証の後，ViT モデルを用いて，モデルがどの顔特徴を重要と判断しているかについてモデル内の Attention 機構を用いて可視化を行った．可視化は，各男女画像の魅力ラベルを用いて，上位 50 点を高魅力度画像，中位 50 点を中魅力度画像，下位 50 点を低魅力度画像として抽出し，各顔画像を重ね合わせ平均化した．平均化の際，ランドマーク情報が欠落している男性画像を 1 枚除外した．分析は Keras/TensorFlow (version 2.8.0) ，Python (version 3.9.7)によって行われ，モーフィングは Python ライブラリの facemorpher を使用した．

3. 結果・考察

検証の結果，男性画像の場合は，ViT モデルのピアソン相関平均値は 0.796，CNN モデルのピアソン相関平均値は 0.765 となり，女性画像の場合は，ViT モデルのピアソン相関平均値は 0.788，CNN モデルのピアソン相関平均値は 0.769 となり，ViT モデルの精度が高いことが示された．次に，男女画像それぞれにおいて，全データを学習した ViT モデルの Attention 機構により顔魅力に重要な特徴の可視化を行った(図 1，図 2)．その結果，男性画像，女性画像のどちらも，特に魅力得点の高い顔画像の場合は目の領域が重要な領域であることが確認できた．目が顔全体の魅力度評価に大きな影響を与えること [14]や，目と皮膚の間の輝度コントラストが性的二型性や魅力に関連すること[15, 16]が心理学研究により知られている．また，本研究の結果は，CNN ベースの手法を用いた先行研究の結果と概ね一致する[9]が，先行研究と比較して目の活性領域はより形態部分に絞られているように見える．CNN は比較的画像のテクスチャ等の情報に注目する一方で[11]，Vision Transformer は形態情報に注目する[12]といったモデルによる特性が影響しているのかもしれない．これらの結果が，人の実際の注意の傾向とどのように対応づいているかについては，心理実験の組み合わせ等による詳細の調査が今後必要となる．

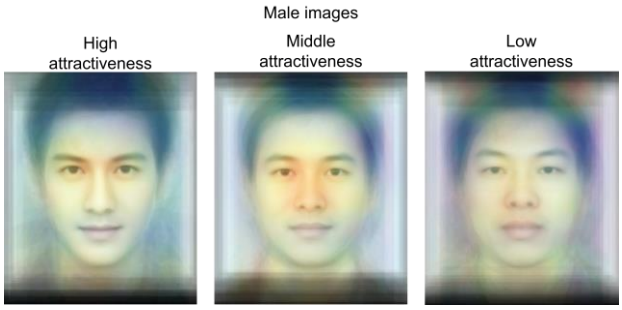

図 1 男性画像の魅力予測に重要な特徴の可視化結果

図 2 女性画像の魅力予測に重要な特徴の可視化結果

文献

[1] Rhodes, G. (2006). "The evolutionary psychology of facial beauty." Annu. Rev. Psychol., Vol. 57, No. 1, pp. 199-226.
[2] Todorov, A., Dotsch, R., Porter, J. M., Oosterhof, N. N., & Falvello, V. B. (2013). "Validation of data-driven computational models of social perception of faces." Emotion, Vol. 13, No. 4, pp. 724.
[3] Nakamura, K., & Watanabe, K. (2020). "A new data-driven mathematical model dissociates attractiveness from sexual dimorphism of human faces." Scientific Reports, Vol. 10, No. 1, pp. 1-11.
[4] Nakamura, K., Ohta, A., Uesaki, S., Maeda, M., & Kawabata, H. (2020). "Geometric morphometric analysis of Japanese female facial shape in relation to psychological impression space." Heliyon, Vol. 6, No. 10, e05148.
[5] Otaka, H., Shimakura, H., & Motoyoshi, I. (2019). "Perception of human skin conditions and image statistics." JOSA A, Vol. 36, No. 9, pp. 1609-1616.
[6] Xu, J., Jin, L., Liang, L., Feng, Z., Xie, D., & Mao, H. (2017, March). "Facial attractiveness prediction using psychologically inspired convolutional neural network (PI-CNN)." In 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 1657-1661). IEEE.
[7] Xu, J., Jin, L., Liang, L., Feng, Z., & Xie, D. (2015). "A new humanlike facial attractiveness predictor with cascaded fine-tuning deep learning model." arXiv preprint arXiv:1511.02465.
[8] 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).
[9] SANO, T. (2022). "Visualization of Facial Attractiveness Factors Using Gradient-weighted Class Activation Mapping to Understand the Connection between Facial Features and Perception of Attractiveness." International Journal of Affective Engineering, Vol. 21, No. 2, pp. 111-116.
[10] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929.
[11] Baker, N., Lu, H., Erlikhman, G., & Kellman, P. J. (2018). "Deep convolutional networks do not classify based on global object shape." PLoS computational biology, Vol. 14, No. 12, e1006613.
[12] Tuli, S., Dasgupta, I., Grant, E., & Griffiths, T. L. (2021). "Are convolutional neural networks or transformers more like human vision?". arXiv preprint arXiv:2105.07197.
[13] Liang, L., Lin, L., Jin, L., Xie, D., & Li, M. (2018, August). "SCUT-FBP5500: A diverse benchmark dataset for multi-paradigm facial beauty prediction." In 2018 24th International conference on pattern recognition (ICPR) (pp. 1598-1603). IEEE.
[14] Terry, R. L., & Davis, J. S. (1976). "Components of facial attractiveness." Perceptual and motor skills.
[15] Russell, R. (2009). "A sex difference in facial contrast and its exaggeration by cosmetics." Perception, Vol. 38, No. 8, pp. 1211-1219.
[16] Jones, A. L., Russell, R., & Ward, R. (2015). "Cosmetics alter biologically-based factors of beauty: Evidence from facial contrast." Evolutionary Psychology, Vol. 13, No. 1, 147470491501300113.