H-030

畳み込みニューラルネットワークによる笑顔評価の推定

Estimation for Evaluation of Smile by Convolutional Neural Networks

糸洲 昌隆† 遠藤 聡志† 當間 愛晃† 山田 孝治† 赤嶺 有平†
Masataka Itosu Satoshi Endo Naruaki Toma Koji Yamada Yuhei Akamine

1. はじめに

人間のコミュニケーションにおけるメッセージ伝達は、視覚情報が全体の 55%を占めるという結果が報告されている[1]。このことから、表情や態度、ジェスチャーなどの視覚から得られる情報は、対人コミュニケーションにおいて、相手が受ける印象を決定する重要な要素であることがわかる。井上らの研究によると[2]、同じ笑顔でも口を開けた笑顔と閉じた笑顔では、受け手の性別や年齢によって印象に差異が見られたという結果が報告されている。また、笑顔の形状と印象に関する研究では[3][4]、顔パーツの形状や位置によって異なる印象を受けると結論づけている。以上のことから、人間は笑顔の中にも異なる印象を表す複数の笑顔を持つことがわかる。これは笑顔に限定されるものではなく、他の表情でも同様であると考えられる。従って、表情を印象によって分類することで、感情で分類されていた表情を細分化することができるのではないかと考えた。

表情認識の研究において、畳み込みニューラルネットワーク(CNN)が使用された例は、数多く存在する。その一つの例として、VICTOR らが Support Vector Machine(SVM)と CNN をそれぞれ用いた表情認識を行っている研究が挙げられ、[5]CNN に関しては認識精度が SVM よりおよそ 6%上回る、約 65%の精度が得られている。本研究では、機械学習の一つである CNN を用いて笑顔表情の良し悪しを学習させることで、好評価を持たれるような笑顔を推定する。また、笑顔評価の推定に影響を与える要素を分析するために、逆畳み込みニューラルネットワーク[6]を使用した画像生成による学習部分の可視化を行う。

2. 畳み込みニューラルネットワークと逆畳み込みニューラルネットワーク

Deep Learning の一つである CNN は、人の脳の視覚野における神経回路を模した順伝播型のネットワークである。CNN の特徴は、畳み込み層とプーリング層と呼ばれる部分的に結合された層で構成されていることにある。畳み込み層は入力値に対してフィルタ処理を行うことで、入力から特徴を抽出した特徴量マップを出力する層である。プーリング層では、畳み込み層で抽出された特徴量をフィルタで計算、出力し、画像内での微小な位置の変化にも同じ出力で対応することができる。これら二つの層は入力画像から得られた特徴量を圧縮する動作を行うことで、高い汎化性能を実現している。一般的に CNN はこれらの層の後に、出力層付近に一層以上の全結合層を接続した構造となる。

逆畳み込みニューラルネットワークは CNN の機能である畳み込みやプーリングの逆の処理を行うことで、畳み込んだ特徴量を元の画像に復元するように学習するモデルである。逆畳み込みネットは主に、逆畳み込み層とアンプーリング層を交互に接続するような構造となっており、アンプーリング層で特徴量サイズを拡大し、逆畳み込み層で元の入力値に復元する処理を行っている。

3. 提案手法

笑顔の良し悪しを CNN に学習させるために以下の 5 つのプロセスの方法(図 1)を提案する。

(1) 笑顔とその他表情で識別する学習器 A を構築する。
(2) ラベルなしデータ群から笑顔画像の抽出を行う。
(3) 笑顔ラベルが付与された画像群でアンケートを行う。
(4) 笑顔印象データセットを作成する。
(5) 笑顔評価を推定する学習器 B を構築する。

プロセス 1 では、表情識別用データセットを「笑顔」と「その他」で構成し、学習器 A で学習を行う。プロセス 2 では学習器 A を用いてラベルなしデータ群から笑顔画像を収集する。プロセス 3 ではプロセス 2 で得た笑顔データ群に対して 5 点満点のアンケートを行う。プロセス 4 では評価点が高い画像と低い画像を取得し、評価点が高い画像に「良い笑顔(Good)」ラベル、評価点が低い画像に「悪い笑顔(Bad)」ラベルを割り当てることで笑顔印象データセットを作成する。プロセス 5 では笑顔印象データセットを学習器 B で学習させる。

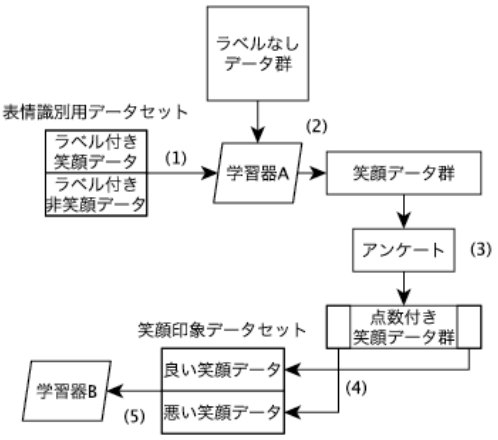

図 1:提案手法の流れ

4. 実験

4.1 実験概要

学習実験では提案手法に則り、Good ラベルと Bad ラベルが付与された笑顔印象データセットを作成し、学習器 B を用いて笑顔評価の推定を行う。学習器 A と学習器 B で使用するネットワークモデルは、ILSVRC 2012 で Krizhevsky らが提案したネットワーク構造(図 2)を用いた[7]。本実験ではプロセス 1 から 4 までを、学習器 B を構成するまでの前処理として扱う。プロセス 1 の表情識別用データセットとして、The Japan Female Facial Expression Database(JAFFE)[8], Cohn-Kanade database(CK)[9], Montreal

† 琉球大学 University of The Ryukyu,

Set of Facial Displays of Emotion(MSFDE)[10], Karolinska Directed Emotional Faces Database(KDEF)[11]を使用し、プロセス 2 のラベルなしデータ群として、Happy People Images(HAPPEI)[12]を使用する。本実験ではプロセス 1 から 4 までを、学習器 B を構築するまでの前処理として扱う。プロセス 1 の結果、学習器 A の全体精度は 98.33%となった。プロセス 2 の笑顔画像抽出工程では、7425 枚のラベル無しデータ群から 1052 枚の笑顔画像を取得することができた。プロセス 3 では 6 人(男性 5 人、女性 1 人)にアンケートを行って、笑顔データ群に点数をつけた。プロセス 4 では、評価点が高い方から 80 枚の画像を Good ラベル、評価点の低い方から 80 枚の画像を Bad ラベルと割り当てを行い、笑顔印象データセットを作成した。また、学習器 B の精度確認手法として 2 分交差検定を行った。

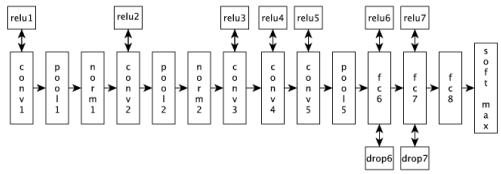

図 2:使用するネットワーク

4.2 実験結果

学習器 B の精度を表す混同行列を図 3 に示す。図 3 より、学習器 B の全体精度は約 70.94%となった。また、プロセス 4 で Good ラベルと Bad ラベルが付与されなかった 892 枚の点数付き笑顔データを学習器 B に識別させることで、中間的な評価のデータに対しての笑顔評価の傾向を分析した(図 4)。図 4 より、評価点が約 3.8 点以上は Good と識別された画像が多く、特に約 4.0 以上では Good と Bad 画像数の差が明確に出ており、この傾向は人の感性と一致していると考えられる。しかし、約 3.8 点以下は Bad 画像の方が Good 画像よりも多いが、その差は少ない傾向にある。このような結果になった理由は、低評価の笑顔は評価側の個人差が出やすいため、特徴が捉えづらいからだと考える。

これらの結果より、学習器 B の笑顔評価の学習は一定の成果を得られているが、Bad 画像の推定が Good と比べてうまくいっていないことが確認できる。

図 3:学習器 B の精度
(行:正解ラベル、列:予測のラベル)

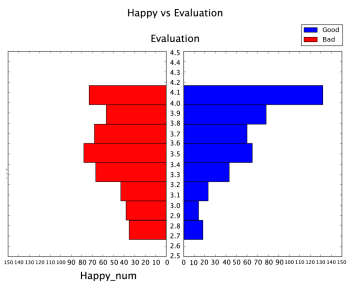

図 4:点数付き笑顔データ群の識別結果
(横軸:画像数、縦軸:アンケート評価点)

5. 笑顔の特徴部位の分析

笑顔評価の推定に影響を与える要素を分析するために、学習器 B の pool5 層(図 2 参照)から出力した値で学習した逆畳み込みニューラルネットワークを使って特徴量の可視化を行う。逆畳み込みニューラルネットの構造は図 2 の畳み込みネットとは逆の構造となるように構築した。学習後の逆畳み込みニューラルネットから生成された画像を図 5 に示す。図 5 より、Good 画像と Bad 画像の視覚的な違いを見ることはできないが、目や口の部分が強調されているように見ることができる。よって、学習器 B は目や口の部分の特徴を捉えて笑顔評価を推定していることが確認できる。

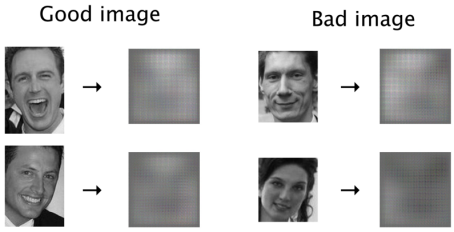

図 5:生成した画像の例

6. まとめ

本研究では、CNN による笑顔の好感度推定を行った結果、約 70%の認識精度が得られた。また、作成した学習器 B は逆畳み込みによる特徴量の可視化により、目と口の特徴が認識に影響を与えていることを確認した。今後の課題として、アンケート対象者を増やすことと、印象による微妙な笑顔表情の違いを学習することができるアルゴリズムを検討する。

参考文献

[1] Mehrabian Albert,"Silent messages: Implicit communication of emotion and attitude."Belmont, CA: Wadsworth (1981).
[2] 井上 清子,"表情が初対面の相手に与える印象" 生活科学研究 36(2014):183-194
[3] 井口 竹喜,"魅力的な笑顔に表れる幾何学的特徴" Konica Minoruta technology report 4 (2007): 91-96.
[4] 菅原 徹,et al. "笑顔の多様性と印象の関係性分析" 感 性工学研究論文集 7.2(2007): 401-407.
[5] NEAGOE, VICTOR-EMIL, et al. "A Deep Learning Approach for Subject Independent Emotion Recognition from Facial Expressions." Recent Advances in Image, Audio and Signal Processing (2013): 93-98.
[6] Zeiler, Matthew D., et al. "Deconvolutional networks." Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, 2010.
[7] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems. 2012.
[8] Lyons, Michael, et al. "Coding facial expressions with gabor wavelets." Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on. IEEE, 1998.
[9] Kanade, Takeo, Jeffrey F. Cohn, and Yingli Tian. "Comprehensive database for facial expression analysis." Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on. IEEE, 2000.
[10] Beaupré, Martin G., and Ursula Hess. "Cross-cultural emotion recognition among Canadian ethnic groups." Journal of Cross-Cultural Psychology 36.3 (2005): 355-370.
[11] Lundqvist, Daniel, Anders Flykt, and Arne Öhman. "The Karolinska directed emotional faces (KDEF)." CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet (1998): 91-630.