F-045

表情に基づく感情推定モデルの基礎検討
A Study of Emotion Estimation Model Based on Facial Expression

福家 穂乃佳† 浦野 昌一†
Honoka Fuke Shoichi Urano

1. はじめに

日本でスマートフォンが普及し始めた 2012 年時点の SNS 利用率は、10 代～30 代が 50%以上と比較的高い傾向にあったのに対し、40 代～60 代は 40%以下と低い傾向にあった[1]。しかし、近年、SNS の利用は日本社会に定着してきており、2022 年の SNS 利用率は、20 代～30 代が 70%以上、40 代～50 代が 60%以上、60 代が 50%以上であり、2012 年と比べ、利用率が増加傾向にある[2]。それに伴い、SNS 上では多くの動画が投稿されており、スポーツの試合のハイライト動画もその一つである。

しかし、ハイライト動画の作成を行う際の課題点として、長時間の動画を短時間の動画に編集するという作業には時間を要する点、また、編集作業を他に委託した場合には高額な発注費用がかかる点が挙げられる。

そこで、筆者らは、ハイライト動画を自動で作成するシステムを構築することを目的とする。「HaiLyts」をはじめとした現在までに開発されているハイライト動画を自動作成するソフトウェアは、選手の動きに着目し、動きの大きなシーンを抽出することができる。しかし、動静の差が少ない動画の場合、ハイライトシーンを正確に抽出することができないという課題点がある。そのため、筆者らは、スポーツ観戦時の観客の表情を基に、スポーツの試合におけるハイライト動画を自動作成するシステムを構築することを目的とする。本稿では、畳み込みニューラルネットワークを用いた表情分析による感情推定モデルの構築を目指す。

2. 感情推定

感情推定とは、表情、声、文章などから人の感情を推定することである。本稿では表情を分析対象とし、畳み込みニューラルネットワークを用いて感情推定を行う。感情には多くの種類があるが、本研究はハイライトシーンを抽出することを目的としているため、中性とそれ以外の感情という 2 種類に関しての感情推定を行う。

表情を基にした感情推定では、畳み込みニューラルネットワークを用いた学習済みモデルである AlexNet を画像の特徴量抽出器として用い、抽出された特徴量をもとにニューラルネットワークによって感情推定を行った研究が報告されている[3]。

本稿では、顔全体の画像に対して畳み込みニューラルネットワークを用いて学習した感情推定モデルの作成と顔のランドマーク検出をオープンソースソフトウェアである dlib により行い、顔のパーツの画像に対して畳み込みニューラルネットワークを用いて学習した感情推定モデルの作成を行う。

† 明治大学先端数理科学研究科 Meiji University
Advanced Mathematical Sciences

3. 分析手法

3.1 畳み込みニューラルネットワーク(CNN)

畳み込みニューラルネットワークは顔認識、文字認識、自動運転など様々な分野で活用されている深層学習アルゴリズムであり、畳み込み層、プーリング層、全結合層で構成される。畳み込み層では、入力データに対し 2×2 や 3×3 などのフィルタを掛け合わせ、特徴マップを取得する畳み込み処理が行われる。プーリング層では、畳み込み処理により生成された特徴マップの代表値を取得することにより特徴マップを生成している。これにより、入力データの特徴を失うことなく、情報を圧縮することができる。

また、畳み込みニューラルネットワークは、学習したパターンは画像内での位置が変化しても認識することができる移動不変性を持つ。畳み込みニューラルネットワークのフロー例を図 1 に示す。

図 1　CNN のフロー例

3.2 顔のランドマーク検出

顔のランドマーク検出は、スマートフォンのロック解除など様々な分野で活用が進む顔認証技術に使用されている。顔認証技術は、画像や映像内から顔を検出し、目、鼻、口などから特徴を捉え、登録された情報との照合を行うことができる技術であり、顔のランドマーク検出が使われている。

本稿では、顔のランドマーク検出にオープンソースソフトウェアである dlib を使用する。dlib は 68 点の顔のランドマークを検出することができる。Pixabay 内の画像に対し、dlib を用いて顔のランドマーク検出を行った例を図 2 に示す。

図 2　ランドマーク検出例

3.3 ヒートマップによる可視化

ヒートマップは、行列型のデータに対し、色付けを行い、濃淡を用いて表現するグラフである。ヒートマップを用いることで、CNN がどの部分に注視して分類を行ったかを可視化することができる。本稿では、Grad-CAM[4]を用いて可視化を行う。Grad-CAM とは、説明可能な AI であるクラス活性化マッピング(Class Activation Map)手法の一つであり、Selvaraju らによって考案された。最初のクラス活性化マッピング手法である CAM[5]は、Global Average Pooling(GAP)を用いたモデルのみに適応するため、モデルに制約がかかる。そこで、様々なモデルに対応させるため CAM を一般化した手法が Grad-CAM である。CAM は、全結合層の重みを各チャンネルの重みとしていたのに対し、Grad-CAM は、特徴マップの勾配から各チャンネルの重みを求めている。つまり、Grad-CAM は、偏微分の値が最も大きい部分が分類の際に注視している部分であることを示す。Grad-CAM を用いて Pixabay 内の画像に対し、ヒートマップによる可視化を行った例を図 3 に示す。図 3 の顔画像は筆者らがこれまでに作成したモデルで検証を行った際、笑顔と分類された画像である。図 3 より、口に注視して笑顔と分類していることが分かる。

図 3　ヒートマップによる可視化例

3.4 GAN

GAN は、Generative Adversarial Network(敵対的生成ネットワーク)と呼ばれる AI 技術の一つであり、Generator と Discriminator という 2 種類のニューラルネットワークで構成される。Generator は偽物の画像を生成し、Discriminator は与えられた画像が、訓練用データ内の本物の画像か Generator が生成した偽物の画像かを判別する。この手順を繰り返し、互いに競い合い精度を高めている。GAN を用いることにより、低解像度の画像を高解像度化すること、文章から画像を生成すること、架空の画像を生成することが可能となる。

4. シミュレーション

本稿では、有名人の顔のカラー画像 202,599 枚で構成される CelebA データセットを使用し、顔全体の画像と顔のパーツの画像を、畳み込みニューラルネットワークを用いてそれぞれ学習した 2 つのモデルの検証を行う。図 4 にシミュレーションの概要を示す。

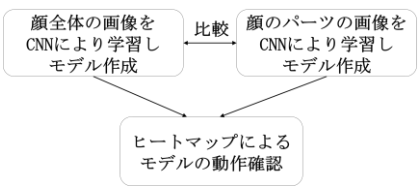

図 4　シミュレーションの概要

4.1 データセット

学習用データセット、検証用データセット、評価用データセットを以下で設定する。

＜学習用データ＞
笑顔の顔画像：28,517 枚
笑顔でない顔画像：28,517 枚

＜検証用データ＞
笑顔の顔画像：9,053 枚
笑顔でない顔画像：9,053 枚

＜評価用データ＞
笑顔の顔画像：4,802 枚
笑顔でない顔画像：4,802 枚

4.2 シミュレーションフロー

シミュレーションの手順を以下に示す。

1　画像データを読み込み、画像サイズを 178×178 ピクセルに変更
2　CNN モデルの作成
　2.1　顔全体の画像を用いた CNN モデルの作成
　2.2　顔のパーツ画像を用いた CNN モデルの作成
　　2.2.1　顔全体の画像からランドマーク検出を用いて顔のパーツ画像の取得
　　2.2.2　パーツ画像のサイズを正方形に変更
3　GAN を用いて、一枚の画像に複数の未知の顔が表示される画像を評価用データから生成
　3.1　生成した画像に対し、各モデルを用いて表情判定
　3.2　ヒートマップを用いて各モデルの動作確認
4　各モデルの精度の比較

5. まとめ

本稿では、スポーツの試合におけるハイライト動画の自動作成に向け、顔全体の画像と顔のパーツの画像を用いた 2 つのモデルの作成と精度検証方法の提案を行った。

今後は、シミュレーションを進めていくと共に笑顔以外の表情によるモデルを作成し、ハイライトシーンを抽出する際に有効となる感情を検討していく。

参考文献

[1] 総務省情報通信政策研究所, ” 平成 29 年情報通信メディアの利用時間と情報行動に関する調査報告書”, (2018).
[2] 総務省情報流通行政局情報通信政策課情報通信経済室, ”国内外における最新の情報通信技術の研究開発及びデジタル活用の動向に関する調査研究の請負成果報告書”, (2022).
[3] 原田 大生, 橋本智昭, “事前学習済み畳み込みニューラルネットワークを用いた表情画像の特徴量抽出に基づく機械学習による感情推定”, 第 65 回自動制御連合講演会, No.2G2-1(2022).
[4] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization”, IEEE, International Conference on Computer Vision (ICCV), pp. 618-626, (2017).
[5] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, Antonio Torralba, “Learning Deep Features for Discriminative Localization”, IEEE, Conference on Computer Vision and Pattern Recognition (CVPR), pp.2921-2929, (2016).