原著論文 Original Paper

畳み込みニューラルネットワークを用いた表情表現の獲得と顔特徴量の分析

Feature Acquisition and Analysis for Facial Expression Recognition Using Convolutional Neural Networks

西銘 大喜 Taiki Nishime
琉球大学大学院 理工学研究科 情報工学専攻
Graduate School of Information Engineering, University of The Ryukyus
taiki_one@eva.ie.u-ryukyu.ac.jp

遠藤 聡志 Satoshi Endo
琉球大学 工学部 情報工学科
School of Information Engineering, University of The Ryukyus
endo@ie.u-ryukyu.ac.jp, https://ie.u-ryukyu.ac.jp/

當間 愛晃 Naruaki Toma
(同　上)
tnal@ie.u-ryukyu.ac.jp

山田 孝治 Koji Yamada
(同　上)
koji@ie.u-ryukyu.ac.jp

赤嶺 有平 Yuhei Akamine
(同　上)
yuhei@ie.u-ryukyu.ac.jp

keywords: facial expression, convolutional neural networks

Summary

Facial expressions play an important role in communication as much as words. In facial expression recognition by human, it is difficult to uniquely judge, because facial expression has the sway of recognition by individual difference and subjective recognition. Therefore, it is difficult to evaluate the reliability of the result from recognition accuracy alone, and the analysis for explaining the result and feature learned by Convolutional Neural Networks (CNN) will be considered important. In this study, we carried out the facial expression recognition from facial expression images using CNN. In addition, we analysed CNN for understanding learned features and prediction results. Emotions we focused on are “happiness”, “sadness”, “surprise”, “anger”, “disgust”, “fear” and “neutral”. As a result, using 32286 facial expression images, have obtained an emotion recognition score of about 57%; for two emotions (Happiness, Surprise) the recognition score exceeded 70%, but Anger and Fear was less than 50%. In the analysis of CNN, we focused on the learning process, input and intermediate layer. Analysis of the learning progress confirmed that increased data can be recognised in the following order “happiness”, “surprise”, “neutral”, “anger”, “disgust”, “sadness” and “fear”. From the analysis result of the input and intermediate layer, we confirmed that the feature of the eyes and mouth strongly influence the facial expression recognition, and intermediate layer neurons had active patterns corresponding to facial expressions, and also these activate patterns do not respond to partial features of facial expressions. From these results, we concluded that CNN has learned the partial features of eyes and mouth from input, and recognise the facial expression using hidden layer units having the area corresponding to each facial expression.

1. はじめに

コミュニケーションにおいて, 表情は自分や相手の感情を表し, 発する言葉と同等に重要な情報の 1 つであると考えられる. 表情と感情の結びつきは普遍的なものであり, 笑顔なら喜び, 幸せ, 安堵, というような感情の認識, 推定が可能である [Ekman 87]. しかし, 表情は表出の強弱や主観的な判断から認識に個人差が存在し, 喜びから安堵, 満足へと多様な分類が可能であることから, 人の表情認識でも一意に判断できない表情の存在がすると考えられる. 表情認識では, Ekman ら [Ekman 87] により定義された普遍性を持つ基本 6 感情 (怒り, 嫌悪, 恐怖, 喜び, 悲しみ, 驚き) に無表情を加えた 7 表情を扱うことが一般的であるが, 対象とする表情の選択は多様であり, 表情認識で広く用いられている JAFFE データセット [Lyons 90], MSFDE データセット [Beaupré 05] などでも, “Happiness”, “Joy”のように似た意味の異なる定義がされる場合や, ”Shame”のようなデータセット固有に定義された表情も存在している.

機械学習による表情認識では, Facial Action Coding System(FACS) のように人手で定義される特徴量

を用いた手法が一般的であったが, 最近では畳み込みニューラルネットワーク (Convolutional Neural Networks:CNN)[LeCun 89] を用いた手法が広く用いられている. これは, 物体認識で高い性能を示すことや画像生成 [Gatys 15], 自動運転技術 [Bojarski 16] など幅広い問題への応用などで成果を上げていることに起因すると考えられる. しかし, その一方で CNN には高い精度が得られる原因や予測説明などに対する疑問も残っている. 表情認識は, 認識に個人差が存在する問題であることから, 認識精度だけで結果の信頼性を評価することは難しく, 結果に対する説明や CNN モデルの性質を明らかにすることは重要である. 予測モデルと人の認識結果が異なる場合に得られる予測結果に加えて予測に有効な特徴, モデル性質の情報などがあれば, 人と異なった認識結果でも間違いではなく, 解釈の違いとして扱うことで意思決定において有意な情報を示すことができる.

以上を踏まえ, 本論文では表情画像と CNN を用いて表情特徴の学習と精度評価, 学習済み CNN の分析を行い, CNN によって学習された有効な表情特徴と CNN 内部での表情特徴の処理を明らかにすることを目的とする. 表情認識の精度評価に学習済み CNN の分析結果を加味して, 人の表情認識との違いや CNN を用いた表情認識の解釈性について議論する.

2. 関 連 研 究

野宮 [野宮 11] らは, FACS 特徴量で定義される眉と目の端点と鼻と口の周囲の顔特徴点 (Action Unit:AU) から距離や領域内の画素値などの特徴量を抽出し Support Vector Machine を用いて表情認識を行い, CK[Kanade 00], MMI[Bartlett 06], JAFFE[Lyons 90] データセットに対し 89%, 59%, 61%の精度を得ている. また, Mengryi[Liu 13] らは, AU に基づく顔の部分的な特徴量を抽出し, それらの情報を入力としボルツマンマシンを用いて表情認識を行い, CK+[Lucey 10], MMI[Bartlett 06], SFEW[Dhall 11] に対し 92%, 74%, 26%の精度を得ている. これら 2 つの研究では, FACS 特徴量に基づかない自然環境で撮影された表情画像データセット (JAFFE, SFEW) での認識精度が低下してしまう点と, FACS を用いて各表情画像データに表情ラベルを付与する際に専門的知識が必須であることから, 大量のラベル付きデータセットの用意が困難である点が問題となる. また, 特徴量を学習する手法を用いた表情認識の研究も存在する. Victor[Neagoe 13] らは, CNN と JAFFE[Lyons 90] データセットを用いて, 被験者に注目した表情認識を行った. この研究では, 学習データに含まれていた被験者の表情では約 95%, 学習データに含まれていない被験者の表情では約 65%の認識精度が報告されている. この研究では, CNN が学習した特徴量や性質などに対する分析はなされていない. 学習された特徴量の検討に関連した研究として, 中間層出力から入力を再現し可視化する Simonyan[Simonyan 13] らの研究が挙げられる. 彼らの可視化手法は, 中間層出力に現れた特徴量を元に入力を再現するため, 最適に学習が行えず認識精度が低い場合では, 適切な特徴量が含まれない中間層出力になってしまい, 良い可視化結果が得られない. 表情では, 認識に個人差が存在することから全ての表情で同程度の高精度で認識することは難しく, 各表情が区別できるレベルの可視化結果を得ることは難しいと考えられる.

本研究では, 学習済み CNN の分析を行う点だけでなく, より大きなデータセットを使用する点で Victor らと異なり, Simonyan らの研究では, 層全体を 1 つとして分析するのに対し, 本研究では, 1 つの層の各ユニットに注目する点で異なる.

3. 実 験

始めに人の表情認識と CNN による表情認識実験, 精度評価, 学習経過の分析を行い, これらの結果から, 人と CNN の表情認識の共通点, 有効な特徴量について考察し, CNN による表情認識に有効な特徴量についての仮説を述べる.

3·1 人の表情認識

人の表情認識実験では, Facial Expression Recognition 2013(FER-2013) データセット [Goodfellow 13] を利用する. FER-2013 データセットは, blissful, enraged など基本 6 感情と無表情に関連する 184 単語を用いた画像検索で得られた顔画像で構成され, 各画像の教師ラベルは検索に用いた単語を元に対象となる 7 種類に関連が強い表情が付与される. また各画像はグレイスケール化され, 48×48 のサイズで顔部分がトリミングされている. 実験では, 各表情からランダムに 10 枚ずつ選び計 70 枚の画像を用いた. 被験者は研究室の学生 10 名である. 紙面に 70 枚の画像をランダムに配置し, 被験者は各表情にラベルを付与する. 表 1 は, 被験者 10 人の全表情の認識精度を表し, 表 2 は各表情の正解件数を表している. 表 1 から全体平均 66%の精度が確認できる. 表 2 の各表情の平均認識件数に注目すると, 喜びと驚き, 無表情では 90%を超え, 怒りと悲しみは約 50%, 嫌悪と恐怖は

表 1 認識精度

回答者	認識率
1	64.3%
2	78.6%
3	72.9%
4	67.1%
5	61.4%
6	62.9%
7	72.9%
8	68.6%
9	50%
10	61.4%
平均	66.01%

表 2 正解件数

回答者	怒り	嫌悪	恐怖	喜び	悲しみ	驚き	無表情
1	5	6	3	8	4	10	9
2	7	7	2	10	9	10	10
3	5	5	3	9	5	9	10
4	6	4	2	9	6	10	10
5	5	3	4	9	4	10	8
6	4	3	1	10	6	10	10
7	4	5	4	10	8	10	10
8	5	5	2	9	7	10	10
9	4	2	0	9	4	6	8
10	6	2	2	9	6	10	8
平均	5.1	4.2	2.3	9.2	5.9	9.5	9.3

表 3 FER-2013 データセットの詳細 (枚)

	怒り	嫌悪	恐怖	喜び	悲しみ	驚き	無表情	合計
Training	3993	2616	4096	7212	4828	3171	4692	30878
Test	466	56	496	895	653	415	607	3588

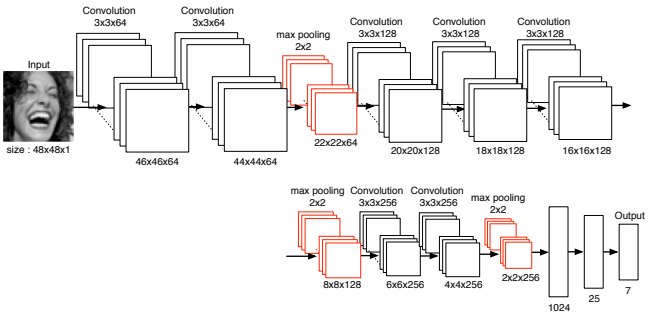

図 1 ネットワーク構成

45%を下回る結果となった．この結果より, 表情によって異なる認識の難しさがうかがえ, 特に恐怖の表情は難しいことが確認できる．以上から, 人の表情認識でも表情毎に異なる認識の難易度が確認され, 認識の個人差が存在していると考えることができる．次に CNN を用いた表情認識を行い, 人の表情認識と比較する．

3·2 CNN を用いた表情認識

§1 データセット

本実験では, 人の表情認識実験と同様に FER-2013 データセットを用いる．実験で使用した表情データセットの詳細を表 3 に示す．データセットは学習用, テスト用に分割されている．嫌悪の件数が少ないので前処理として, かさ増し処理を行い嫌悪画像を 436 枚から 2616 枚に増加させた．ここで, かさ増し処理には, ガンマ変換を用い輝度の異なる画像を生成した, 元画像を含め反転処理を行い, 元データを 6 倍にした．加えて, かさ増し処理後にデータセット全てに GCN を行った．このデータセットでは顔の位置と向きが統一されていないが, CNN では畳み込み処理により顔位置のズレを吸収できるので, これらのズレに対する前処理は行わない．また, CNN による表情認識では, 被験者に向けた表情認識実験とは使用する画像の総数が異なる．

§2 実験環境

表情認識実験に使用したネットワーク構成を図 1 に示す．図中の各層の上数字は処理名とサイズ (縦 × 横 × チャネル), 下数字はその層での処理後の出力サイズを表す．入力層ユニット数を画素数に対応した 48×48 に, 出力層を表情数に合わせて設定した．各プーリング層の直前に Batch Normalization を, 直後に Dropout の処理をそれぞれ行い, 全結合層では各層で Dropout を行っている．活性化関数は ReLU, softmax 関数を使用し, ミニバッチサイズは 128 で固定して学習を行った．また学習は, 300Epoch 行い, 10Epoch 毎でテストデータを用いて精

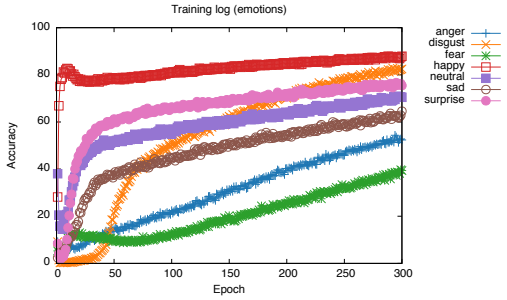

図 2 各表情の精度推移 (学習データ)

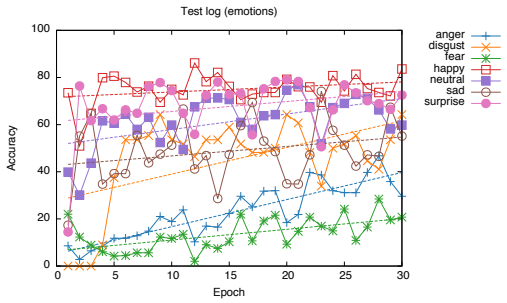

図 3 各表情の精度推移 (テストデータ)

表 4 認識精度 (%)

Actual class \ Predicted class	怒り	嫌悪	恐怖	喜び	悲しみ	驚き	無表情
怒り	29.61	10.3	4.7	14.16	24.67	2.57	13.94
嫌悪	5.35	64.28	3.57	7.14	8.92	0.0	10.71
恐怖	4.03	5.04	20.76	9.67	31.45	8.66	20.36
喜び	0.89	0.22	1.34	83.57	4.13	1.78	8.04
悲しみ	4.13	2.6	4.59	9.8	55.15	0.91	22.81
驚き	1.2	0.72	7.46	6.5	4.09	72.53	7.46
無表情	1.97	1.97	2.63	9.88	22.57	1.15	59.8

度評価を行う．さらに学習終了後に, CNN に対する分析を行う．

3·3 実 験 結 果

§1 精度評価

図 2, 図 3 に学習時の各表情の精度推移 (図 3 の破線は 1 次近似), 表 4 に認識精度を示す．図 2, 図 3 から, 学習による認識精度の向上が確認できる．

CNN の表情認識では, 全体で約 57.1%の認識精度が得られた．表 4 より, 喜びと驚きは 70%を超える精度が得られた．一方, 悲しみについては恐怖や怒り, 無表情として誤認識される結果も目立つ結果となった．また, 前処理を行わなかった顔位置のズレによる影響は確認できなかった．これは CNN の畳み込みとプーリングにより位置に依存しない有効な特徴量が学習された結果だと思われる．

3·4 認識可能データの推移とクラスタ

表情の認識難易度について 300Epoch の学習過程の可視化から考察する．可視化法には, 全結合層 (25 次元) の中間層出力に対し t-SNE[Maaten 08] を用いる．また, 以

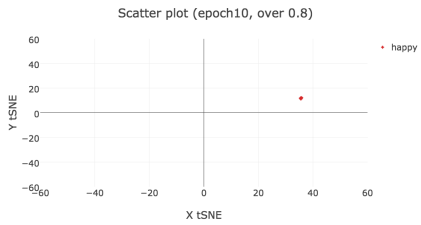

図 4 10 Epoch (確定的正解)

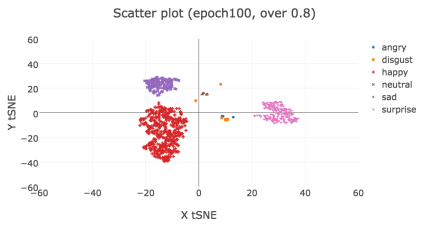

図 5 100 Epoch (確定的正解)

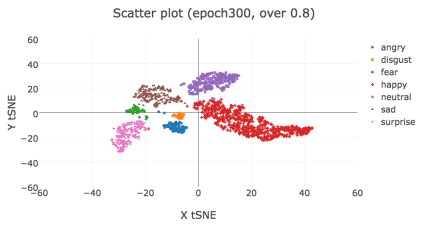

図 6 300 Epoch (確定的正解)

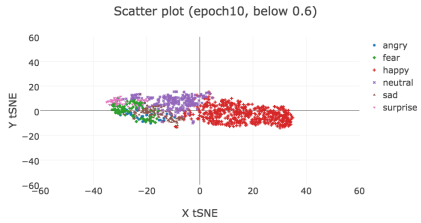

図 7 10 Epoch (暫定的正解)

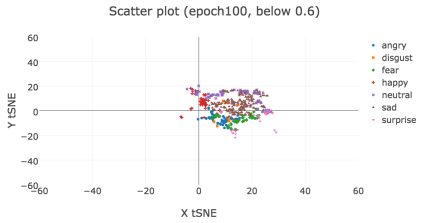

図 8 100 Epoch (暫定的正解)

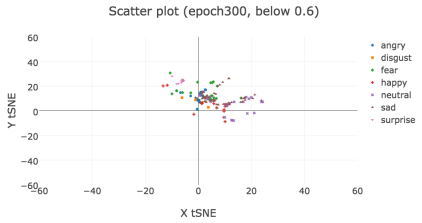

図 9 300 Epoch (暫定的正解)

降, 最終層の Softmax 関数による出力を信頼度と定義し, その出力値が 0.8 以上の正解事例を確定的正解とし, 出力値が 0.6 以下の正解事例を暫定的正解と呼ぶ. 始めに, 確定的正解データの結果を図 4〜図 9 に示す. 図 4, 図 5 より, 100 Epoch で喜び, 驚き, 無表情を学習していることが確認できる. これらの表情は人の表情認識においても精度が高い. 以降は嫌悪, 悲しみ, 怒り, 恐怖の順にプロットされるデータが増加しており, 図 6 から, 学習後半の 100 Epoch 以降に学習した怒りや嫌悪, 恐怖, 悲しみのいわゆるネガティブな表情はプロットされるデータが少なく互いに近くにプロットされていることが確認できる. また, 暫定的正解データの結果を確認すると, 学習初期 (図 7) は, 喜び, 無表情などの固まりが確認できるが, 図 8, 図 9 のように学習によって, プロットされるデータが減少していくことが確認できる. これは, 暫定的正解として判断された表情が学習により, 確定的正解に変化した結果である. 以上より, CNN の表情認識では, 今回扱う 7 表情に対して特徴量の獲得に難易度の差があり, 実験においては, 確定的正解に至るための学習回数の差として表れると考えられる.

3·5 人の表情認識との比較

人と CNN の認識結果を比較すると以下の 2 点が共通した.

- 喜び, 驚き, 無表情は認識しやすい
- 怒り, 嫌悪, 恐怖, 悲しみの認識は難しい

これら 2 つの共通した性質は, 人と CNN の認識精度の比較や学習過程の可視化からも確認することができる. すなわち, 人と CNN の表情認識には相関性があると考えられる. ここで, CNN が獲得した特徴量の意味を考える時, 人が重視する目元や口元といった部分的な顔特徴の獲得が, CNN でも実現されているのではないかと予想される. 実際, FACS 特徴量では目元, 口元の特徴量が極めて重要である. それらを考慮すると, “CNN は入力から目元や口元の部分的特徴を学習し, 各部分的特徴に対応した領域を持つ中間層ユニットを用いて表情認識を行う” という仮説が立てられる.

4. 学習済み CNN の分析と理解

CNN で学習された特徴量を解釈するために, 中間層出力を復元する手法 [Zeiler 14] や分類器を用いた分析手法 [Ribeiro 16] など, 様々な可視化に関する手法が提案されている. 学習した特徴量に対する分析として, 入力層からの情報を抽出している中間層や入力データの特徴量に注目することが一般的である. 本論文では中間層の各ユニットに注目し, それらの出力から学習された特徴量や入力画像毎に異なるであろう CNN の発火パターンを確認する.

4·1 中間層出力の性質と分析

中間層では入力された情報から問題に対し有効な特徴量を抽出しながら出力層に向けて値を伝播すると考えられるが, 実際には表情毎の各ユニットの役割やその性質は未知である. 表情認識を考えた場合, 仮に, ユニット毎に目や口の特徴量を分散的に捉えているのなら, ある表情ではユニットの発火パターンを持つと考えられる. もし中間層がそのような状態であるなら, それらのユニットに大きく反応する入力画像を収集することで, 特定の発火するユニットの役割を明らかにすることができると考えた. そこで本節では, 各ユニットの役割とその性質を明らかにすることを目的とし, 以下の流れで分析を行う.

(1) 各表情を入力し得られる中間層出力からヒートマップを作成し, 発火パターンの有無を確認する
(2) 表情毎に対応した発火パターンが発見できれば, 対応ユニットに大きく反応する入力画像を収集する
(3) 集めた画像データを観察し, 発火するユニットの役割, 性質について考察する

分析に使用する入力データは, 3·4 節と同様に確定的正解データを使用する.

4·2 分 析 結 果

§1 発火パターンの有無について

始めに 25 次元の全結合層の出力を対象にヒートマップを作成した. ヒートマップでは抜粋した 20 件の確定的正解データを縦軸, ユニット番号を横軸とする. 図 10 に結果を示す.

図の赤列は縦軸で表される表情画像を入力した場合に発火したユニットを示し, 青列は発火していないユニットを示す. 実験に用いた CNN は ReLU を使用しているため負の値は次層へ伝播しないので, 青列の値の大きさは考慮する必要がないと考えられる. 赤列の発火ユニットに注目すると, 発火の大きさは異なるが, 各表情が中間層出力に発火パターンを持つことが確認できる. このような発火パターンは他の表情でも確認できる. 表 5 に表情毎にまとめたものを表す. 表 5 のユニット番号に対応した × 印は発火を表し, 空欄は発火しないユニットを示す. 表情毎の発火パターンを比較すると, 喜び, 驚き, 無表情では発火パターンに重なりが少なく, 嫌悪, 悲しみ, 怒り, 恐怖の表情内では共通した発火ユニットが存在しており発火パターンに重なる部分が多い. この結果より, 喜び, 驚き, 無表情では学習した部分的特徴に対応する独立した領域を持ち, 反対に嫌悪, 悲しみ, 怒り, 恐怖では共通した部分的特徴に対応した領域を持つと考えられる. したがって, 表情により差はあるが, 部分的特徴に対応した領域を持つ中間層ユニットを用いて表情認識を行う, という仮説に合った内容であると考えられる. 次に各発火ユニットが対応する部分的特徴を分析する.

§2 最大発火を示す表情の収集

表情毎の発火ユニット群で捉えている部分的特徴について, 特定ユニットの発火を最大化する入力データを用いて分析する.

図 11 に喜びの各ユニットの発火を最大化する入力の上位 20 件を示した. 図の縦軸のユニット番号は, 先のヒートマップの発火パターンにそれぞれ対応しており, 各表情画像上の数値は対応するユニットの出力値を示している. 各出力値は発火が大きい順に左から右に並んでいる. また, ユニット毎に発火の大きいデータを選択しているので, 複数ユニットで発火が大きい場合は, 複数行に表れる. 確定的正解データのみを対象としているので, 人目でも表情が判断しやすいデータが並んでいることが確認できる. しかし, 各行毎に異なるパターンが確認できず, 例えば, 開口度合いに対応した喜びや驚き, 眉や目の形状に対応した無表情のように, 行毎すなわち, ユニット毎で部分的特徴に対応していないことがわかる.

中間層出力に注目して各表情に特有のユニット発火パターンの分析, 中間層ユニットの発火を最大化する入力画像の収集を行った. その結果, 学習済みネットワークの中間層出力では表情毎に異なる発火パターンを有するが, 各ユニットの出力値が部分的特徴に対応しないことが確認された. これは, 学習に用いた教師信号が部分的特徴ではなく, 正解表情であるか否かで表現されていることが影響していると考えられる. 以上より, CNN は入力から目元や口元の部分的特徴を学習し, 部分的特徴に対応した領域を持つ中間層ユニットを用いて表情認識を行うのではなく, “入力から目元や口元の部分的特徴を学習し, 各表情毎に対応した領域を持つ中間層ユニットを用いて表情認識を行う”と仮説の後半部分を見直す.

また, 畳み込み層に対しては, 発火の大きいユニットを選択して分析を行ったが, 同様の結果であった.

4·3 入力に注目した分析手法

前節では中間層の状態について分析と考察を行ったが, 表情認識のためにネットワークが有効であると判断した特徴は不明なままである. 本節では, 入力画像から得られる特徴量について, 各表情の信頼度を表す出力値の変化に注目し, 評価実験の結果も踏まえて検討する. また, 前節と同じく確定的正解データを使用して分析を行う. 分析手法を以下に示す.

(1) 入力画像を n 個の分析領域に分ける
(2) 分析領域から 1 カ所選択し, マスク処理として,0.0, 0.2, 0.4, 0.6, 0.8, 1.0 の輝度値で初期化する. これは, 初期化する値によりネットワークの出力値が異なるためである.
(3) 上記処理を行った画像を入力し, CNN の該当表情の信頼度を示すユニットの出力値平均の変化を調べる

始めに, 認識精度が最も高い結果となった喜びの表情について考察する. 図 12 右に喜び画像の例, 対応する領域番号を左にそれぞれ示す. 図 13 に喜びの分析結果を示す. グラフには 0.0 から 1.0 の輝度値で分析領域をマスクした結果が異なる折れ線で表現されており, それぞれ信頼度を示す出力値の変化である. さらに縦軸は出力値, 横軸は各分析領域の番号を示している. この分析では, 部分的特徴の定義を恣意的に決定することを避けるため一律の分割を用いている. また, マスク処理において, マスクする領域が小さすぎるとネットワークの出力に変化が確認できず, 逆に大きすぎるとマスク領域に複数の部分的特徴が含まれてしまい, 単一の部分的特徴の検討を行うことが困難になるため 4×4 の領域を使用した. この分析結果から, 分析領域の 10 番, 11 番をマスクした場合に, 喜びの出力値が最大 0.6 まで減少することが確認できる. 加えて, 分析領域 14 番と 15 番をマスクした場合にも 0.2 ほど出力値が減少していることから, 喜びの認識には口の形状とほうれい線辺りの口元の特徴が強く影響していると考えられる. 図 14 を分析対象とした場合の結果を図 15 に示す. 図 13 の結果とは異なり, 分析領域 9 番, 10 番で出力値が 0.0 を示していることがわかる. これは顔

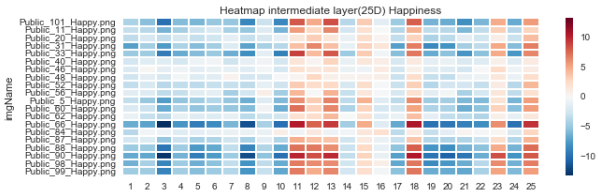

図 10 中間層ヒートマップ (喜び)

表 5 表情毎の発火パターン

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
喜び											x	x	x					x					x		x
驚き	x			x	x	x	x			x			x	x	x		x	x	x					x	
無表情		x	x	x			x	x		x		x		x		x					x	x	x	x	x
嫌悪	x	x	x	x				x	x		x		x				x		x	x					x
悲しみ		x	x		x	x		x	x	x					x	x				x	x	x			x
怒り		x	x		x	x		x	x		x		x			x			x	x		x	x	x	
恐怖			x		x	x		x		x			x	x	x	x			x	x	x			x	

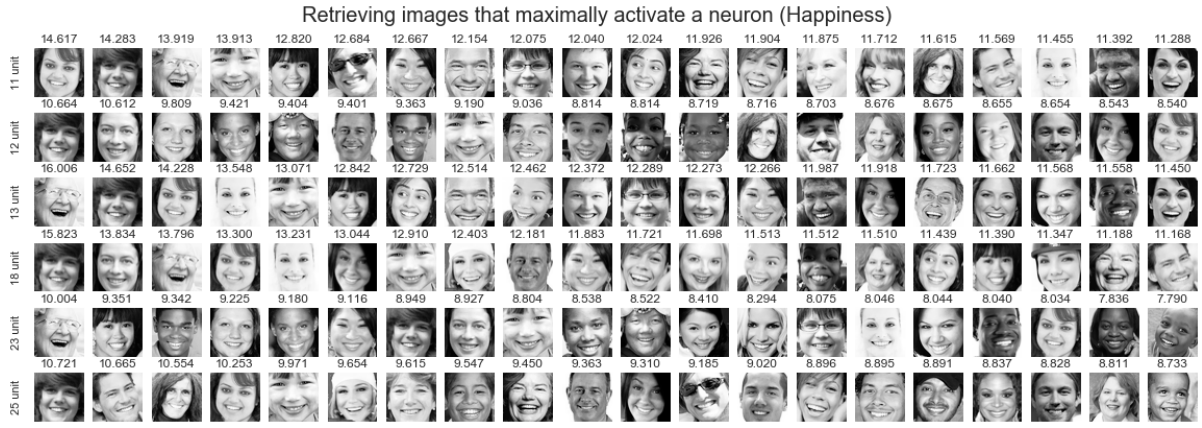

図 11 ユニット発火を最大化する入力集合 (喜び)

が斜めを向いているために, 口の位置が図 12 と図 14 で異なることが影響していると考えられる. この結果から, 画像内の特定位置の画素値に強く影響されているわけではなく, 口元の特徴が喜びの認識に影響を与えていることが確認できた. このことからも喜びの表情認識において, 口元の特徴を学習していることが確認できる. 次に, 喜び以外の驚き, 無表情, 嫌悪, 恐怖, 怒り, 悲しみについての結果を確認する.

図 17, 図 19 に, それぞれ驚き, 無表情の分析結果を示す. 驚きの結果では, 喜びと同様に分析領域 10 番と 11 番のマスク時に出力値が下がっている. 加えて, 分析領域 6 番と 7 番での出力値の減少傾向も確認できるが, 0.8 程度であることから認識に強く影響しないと考えられる. これに対し, 無表情ではマスクを行った場合の出力値の減少が全体的に少ない. 表情の表出が少ない無表情では妥当な結果と思われるが, 分析領域 6 番と 7 番, 10 番, 11 番を同時にマスクし, 顔全体を隠さない限りは無表情だと認識できる結果である.

図 21 にネガティブな表情の中では, 認識精度が高かった嫌悪の結果を示す. グラフは, 喜び, 驚き, 無表情とは異なり, マスクによって出力値が大きく減少している. 特に分析領域 10 番と 11 番をマスクした場合の値の減少は大きく, 嫌悪と判断されなくなるほどに影響を与えることが確認できる. 今回の結果から, 口元や目元単体の特徴で嫌悪を読み取ることが難しいことがわかる.

最後に, 恐怖, 怒り, 悲しみの分析結果を示す. 図 23, 図 25 に示す恐怖, 怒りでは嫌悪と同様な結果がうかがえるが, 怒りは嫌悪よりもマスクに対して非常に敏感であり, 分析領域 6 番や 7 番の目元をマスクした場合でも出力値が 0.4 以下に変化していることが確認できる. 図 27 に示す悲しみでは, これまでの結果と異なり, 分析領域 10 番と 11 番それぞれを 0.6 以上の値でマスクした場合に, わずかではあるが出力値が上昇している. 以上のことより, 口元や目元の情報は表情認識に重要であるが, 表情毎に認識に必要な各特徴の影響度が異なることがわかる. また, 全ての分析結果を通して顔の中心部以外の入力はほとんど影響が無かったことから, 目元と口元の組み合わせを考慮する場合には, 顔の輪郭内部で複数箇所のマスク処理を行う分析が有効だと思われる.

4·4 学習済み CNN 分析の考察

本章では, 3 章の精度評価と学習経過の分析の結果からの仮説を元に, 中間層出力と入力に注目した学習済みネットワークの分析を行った.

入力に注目したマスク処理による分析では, 目元や口元の部分的特徴が認識に大きく影響していることがわかる. また, マスクによる出力値の変化幅から, 各部分的特徴が認識に与える影響度が表情毎に異なることを確認した. ヒートマップによる中間層分析では, ユニット発火の可視化により, 各表情に対応する発火パターンの存在と, 各発火パターンで共通して発火するユニットの存在を確認した. さらに, 発火を最大化する表情画像の分析から, 各発火パターンのユニットは部分的特徴に対応するのではなく, 各表情に対応していることを確認した. 各発火パターンのユニットが表情に対応した領域で構成されることは, 学習に用いた表情の部分的特徴が教師信号として陽に与えられないことが影響していると考えられる.

以上の結果より, CNN は入力から目元や口元の部分的

図 12 分析画像例 (喜び)

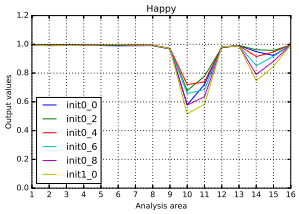

図 13 分析結果 (喜び)

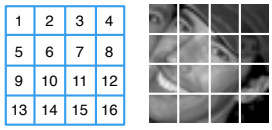

図 14 分析画像例 (非正面)

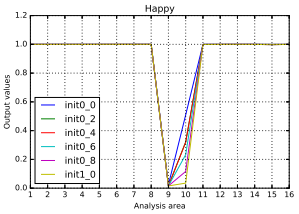

図 15 分析結果 (非正面)

図 16 分析画像例 (驚き)

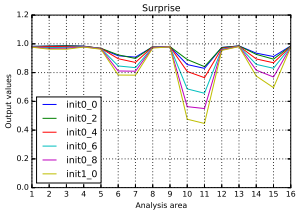

図 17 分析結果 (驚き)

図 18 分析画像例 (無表情)

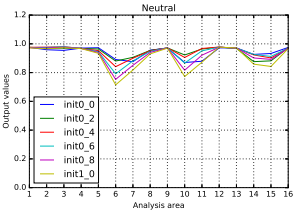

図 19 分析結果 (無表情)

特徴を学習し, 各表情に対応した領域を持つ中間層ユニットを用いて表情認識を行うと結論づけることができる.

5. お わ り に

本論文では, CNN と表情画像データセットを用いて表情表現の学習と精度評価, 人の表情認識との比較, 学習経過の分析を行った. また, 有効な特徴と CNN 内部での各特徴の処理を明らかにすることを目的とした入力と中間層に注目した学習済み CNN の分析を行った.

始めに CNN による学習と精度評価を行い, 全体で約 57%の認識精度が得られた. 各表情に注目すると, 喜び, 驚きの精度は 70%, 嫌悪, 無表情, 悲しみの精度は 50%を超える結果が得られたが, 怒り, 恐怖の表情では 30%に満たない結果となった.

学習経過に注目した分析からは, 喜び, 驚き, 無表情, 嫌悪, 悲しみ, 怒り, 恐怖の順で認識可能データの増加傾向が見られ, 学習の進行に応じた暫定的正解から確定的正解への量的な変化が確認できた. 認識可能なデータの増加を CNN の表情学習の難易度として捉えると, 人の表情認識の結果に沿った傾向であると考えられる. さらに認識しやすい喜び, 驚きと認識しにくい恐怖の表情という人と CNN に共通した傾向も確認できた. また, 入力に注目したマスク処理による分析から, 表情画像に含まれる目元や口元の部分的特徴が認識に大きく影響しており, マスク処理による出力値の変化幅から部分的特徴は表情毎に影響度が異なることを確認した. 中間層の分析でのヒートマップによるユニット発火の可視化からは, 各表情に対応する発火パターンの存在と各発火パターンで共通した発火ユニットの存在することを示した. さらに各ユニットの発火を最大化する表情画像の分析より, 顔の部分的特徴ではなく各表情に対応した領域を持つことを明らかにした. 以上の結果を踏まえ, CNN は入力から目元や口元の部分的特徴を学習し, 各表情に対応した領域を持つ中間層ユニットを用いて表情認識を行う, と結論付けることができる.

今後は, 学習経過でのプロット位置による各表情の関係性の分析とヒートマップでの発火の大きさに注目した分析, 複数の層の発火パターンの分析が課題として挙げられる. また, マスク処理による分析では, 分割数を増やし複数箇所マスクすることで, より細かな特徴を扱うことができ, 部分的特徴を組み合わせた複雑な分析が可能になると考えられる.

◇ 参 考 文 献 ◇

[Bartlett 06] Bartlett, M. S., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I., and Movellan, J. Fully automatic facial action recognition in spontaneous behavior. IEEE 7th International Conference on Automatic Face and Gesture Recognition (2006)

[Beaupré 05] Beaupré, M. G., and Hess, U Cross-cultural emotion recognition among Canadian ethnic groups. Journal of Cross-Cultural Psychology Vol.36, No.3 pp.355-370 (2005)

[Bojarski 16] Bojarski, M., Del Testa, D., et al. End to end learning for self-driving cars. arXiv:1604.07316 (2016)

[Dhall 11] Dhall, A., Goecke, R., Lucey, S., and Gedeon, T. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. IEEE International Conference on Computer Vision Workshops (ICCV Workshops) (2011)

[Ekman 87] Ekman, P., Friesen, W. V., 工藤力 (訳) 表情分析入門. 誠信書房 (1987)

[Gatys 15] Gatys, L. A., Ecker, A. S., and Bethge, M. A neural algorithm of artistic style. arXiv:1508.06576 (2015)

[Goodfellow 13] Goodfellow, I. J., Erhan, D., Courville, A., et al. Challenges in representation learning: A report on three machine learning contests. International Conference on Neural Information Processing. Springer Berlin Heidelberg (2013)

[Huang 16] Huang, G., Liu, Z., Weinberger, K. Q., and van der Maaten, L. Densely connected convolutional networks. arXiv:1608.06993 (2016)

[Kanade 00] Kanade, T., Cohn, J. F., and Tian, Y. Comprehensive database for facial expression analysis. Proceedings. Fourth IEEE International Conference on Automatic Face and Gesture Recognition (2000)

図 20 分析画像例 (嫌悪)

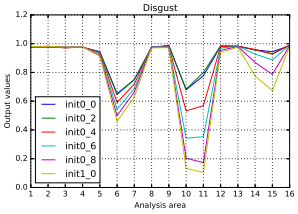

図 21 分析結果 (嫌悪)

図 22 分析画像例 (恐怖)

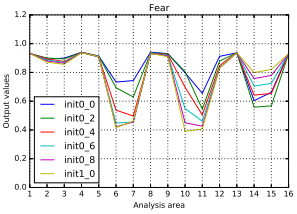

図 23 分析結果 (恐怖)

図 24 分析画像例 (怒り)

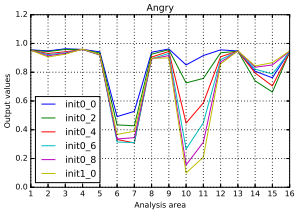

図 25 分析結果 (怒り)

図 26 分析画像例 (悲しみ)

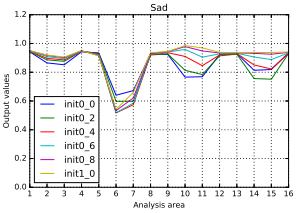

図 27 分析結果 (悲しみ)

[LeCun 89] LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. Backpropagation applied to handwritten zip code recognition. Neural Computation Vol.1 No.4 pp.541-551 (1989)

[Liu 13] Liu, M., Li, S., Shan, S., and Chen, X. Au-aware deep networks for facial expression recognition. 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (2013)

[Lucey 10] Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., and Matthews, I. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. Computer Society Conference on Computer Vision and Pattern Recognition Workshops (2010)

[Lyons 90] Lyons, M., Akamatsu, S., Kamachi, M., and Gyoba, J. Coding facial expressions with gabor wavelets. Proceedings of Third IEEE International Conference on Automatic Face and Gesture Recognition (1998)

[Maaten 08] Maaten, L. V. D., and Hinton, G. Visualizing data using t-SNE. Journal of Machine Learning Research 9.Nov pp.2579-2605 (2008)

[野宮 11] 野宮浩揮, 宝珍輝尚. 顔特徴量の有用性推定に基づく特徴抽出による表情認識. 知能と情報, Vol.23 No.2 pp.170-185 (2011)

[Ribeiro 16] Ribeiro, M. T., Singh, S., and Guestrin, C. Why Should I Trust You?: Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM (2016)

[Simonyan 13] Simonyan, K., Vedaldi, A., and Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv:1312.6034 (2013)

[Neagoe 13] Neagoe, V. E., Barar, A. P., Sebe, N., and Robitu, P. A deep learning approach for subject independent emotion recognition from facial expressions. Recent Advances in Image, Audio and Signal Processing pp.978-960 (2013)

[Vincent 08] Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P. A. Extracting and composing robust features with denoising autoencoders. Proceedings of the 25th International Conference on Machine learning. ACM (2008)

[Zeiler 14] Zeiler, M. D., and Fergus, R. Visualizing and understanding convolutional networks. European Conference on Computer Vision. Springer International Publishing (2014)

〔担当委員：福井 健一〕

2017 年 3 月 24 日 受理

著 者 紹 介

西銘 大喜(学生会員)

2015 年 琉球大学工学部情報工学科卒業. 現在, 同大学院理工学研究科情報工学専攻在学中.

遠藤 聡志(正会員)

1990 年北海道大学大学院工学研究科 電気工学専攻修士課程修了. 同年, 北海道大学工学部助手. 1995 年琉球大学工学部情報工学科講師. 1996 年同助教授. 2004 年同教授. 複雑系工学に関する研究に従事. 計測自動制御学会, 日本知能情報ファジイ学会各会員. 博士 (工学).

當間 愛晃(正会員)

2003 年琉球大学大学院理工学研究科総 合知能工学専攻 (博士後期課程) 修了. 博士 (工学). 2004 年琉球大学工学部情報工学科助手. 2007 年同大学助教. 2015 年同大学准教授. 複雑系工学, データ/テキスト/Web マイニング, 人工知能に従事. 自然言語処理学会, 日本認知科学会各会員.

山田 孝治(正会員)

1995 年北海道大学大学院工学研究科情報工学専攻修了. 博士 (工学). 同年, 琉球大学工学部情報工学科助手, 1996 年 同講師, 1999 年 同准教授. 2014 年 同教授. マルチエージェント, 知能ロボットに関する研究に従事. 情報処理学会, 機械学会, ロボット学会各会員.

赤嶺 有平(正会員)

2004 年琉球大学大学院理工学研究科 博士課程総合知能工学専攻修了. 博士 (工学). 同年, 日本学術振興会特別研究員. 2006 年琉球大学工学部情報工学科助手. 2007 年から同助教. 交通システム, 複合現実感の研究に従事. 地理情報システム学会会員.