

ディープニューラルネットワークによる画像からの表情表現の学習

Emotion Acquisition From Facial Image Using Deep Neural Network

西銘 大喜*¹
Taiki Nishime遠藤 聡志*²
Satoshi Endo當間 愛晃*²
Naruaki Toma山田 孝治*²
Koji Yamada赤嶺 有平*²
Yuhei Akamine*¹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

*²琉球大学工学部情報工学科

School of Information Engineering, University of The Ryukyus

In this study, we carried out the emotion acquisition from facial expression image using a Deep Neural Network(DNN). Emotions we focused on are "happiness", "sadness", "surprise", "anger", "disgust", "fear" and "neutral". By using Auto-encoder and Denoising Auto-encoder approaches, we have obtained an emotion recognition rate of about 68%; five emotions (happiness, surprise, disgust, anger and neutral) recognition rate was about 75%. However, the sadness and fear emotions recognition from only this data seemed to be difficult. This result was the same tendency as preliminary experiment of human recognition. After DNN training, eyes and mouths were confirmed to be recognized effectively from facial expressions. The results corroborate that DNNs work effectively for learning facial expression representations. This paper details these experiments and investigations regarding the influence of DNN learning from facial expressions.

1. はじめに

表情はコミュニケーションを行う上で、相手の感情を表し、発せられる言葉と同等に重要な情報であると考えられる。現在では、携帯電話やインターネットの発達により音声、テキストだけを用いてコミュニケーションを行う機会が増加しつつあり、直接顔を会わせるよりも、何らかのデバイスを介したコミュニケーション方法が多いと考えられる。表情と感情の結びつきは普遍的なものであり[1]、笑顔なら喜び、幸せという感情の認識、推定が可能であると考えられる。表情認識が可能であるなら、表情から感情を推定し、直接相手に知らせることにより、テキスト、音声よりも早い感情の伝達が可能になることも考えられる。

表情認識では、目や眉、口などの顔パーツを特徴量として設計し、機械学習を行う手法が有効であり、約67.5%の認識率が報告されている[2]。DNNを用いて表情認識を行う研究も報告されている[3]。Neagoeらの研究では、CNNを用いて7層と8層のニューラルネットワークを構築し、平均63.13%の認識率が報告されている。しかしながら学習を終えたDNNに関する考察が無く、DNNが何を学習したのかという議論はなされていない。そこで本研究では、人の感覚情報の推定の一例である表情表現の学習を行い、DNNで得られた特徴量、表情推定での有効な特徴量の検討を行う。

2. Deep Neural Network

DNNとは、ニューラルネットワークにおいて複数の中間層を用いて学習を行う手法である。何層も重ねた深い構造を持つニューラルネットワークによる学習はDeep Learning(深層学習)とも呼ばれている。DNNでは、各層の段階的な処理により入力信号を抽出、圧縮された出力層の出力を特徴量として用いる。DNNは中間層が増えることにより、誤差の勾配が分散してしまい入力層に近い層の重みに対して、誤差逆伝播法によ

る学習が適切に行えないという問題があった。その為DNNを構築するための様々な手法が提案されている。

2.1 Stacked Autoencoder

本研究では、多層ニューラルネットワークの構成法としてStacked Autoencoderを用いる[4]。図1にAutoencoder(AE)のネットワークを示す。AEでは、入力信号自体を教師信号として入力と出力が同じ値($X_n = X'_n$)となるよう、誤差逆伝播法で重みの調整(学習)を行う[5]。活性化関数 s にはシグモイド関数を用いる。式(1)、(2)で表される入力と重みの和を利用する。

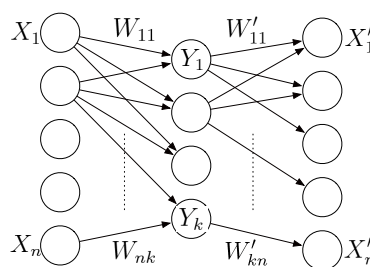


図1: Autoencoder

$$Y_k = s \left(\sum_{i=1}^n W_{ki} X_i \right) \quad (1)$$

$$X'_n = s \left(\sum_{i=1}^k W'_{in} Y_i \right) \quad (2)$$

Stacked Autoencoderでは、AEを積み重ねて階層型のニューラルネットワークを構成する。始めにAEの方法で学習を行う(図2-1)。学習を終えたAEから出力層を取り除くことで、入

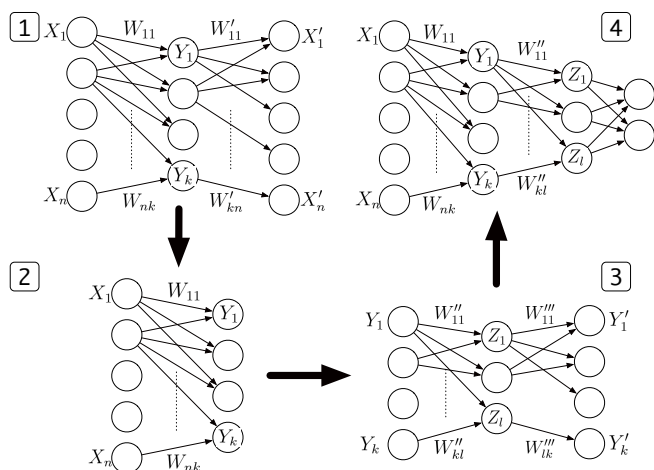


図 2: Stacked Autoencoder の構成法

力信号を圧縮するネットワークが得られる (図 2-2). 圧縮された出力信号を入力信号として、さらに小さな AE の学習を行う (図 2-3). これらの処理を再帰的に行い、出力層を除いた AE を積み重ねることによって Stacked Autoencoder が得る事ができる (図 2-4). Stacked Autoencoder では、AE による階層ごとの学習を終えた後にネットワーク全体を教師信号を用いて学習を行うことが一般的である。

2.2 Denoising Autoencoder

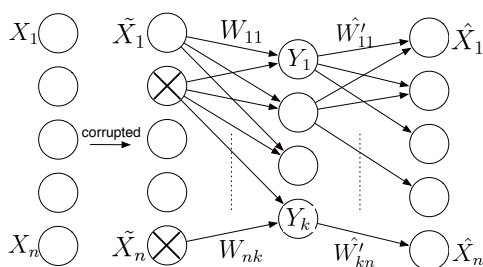


図 3: Denoising Autoencoder

本研究では、AE に加え、Denoising Autoencoder (DAE) も合わせて使用する [5]. DAE を図 3 に示す. DAE では、入力データに対し、ノイズとしてランダムにユニットの値を 0 にする. 出力層への値の伝達にはノイズを加えた入力 ($\tilde{X}_n$) を使用し、AE と同じように出力層まで出力と重みの積の和をとる. ノイズを加えていない入力 (X_n) とノイズを加えた値から作られる出力層出力 ($\hat{X}_n$) の値が同じに値になるように重みの調整を行う ($X_n = \hat{X}_n$). この手法はノイズを加えた入力を使用して学習を行うことで過学習を防ぎ、より汎化能力の高い重みの初期値を与える仕組みである。

3. 実験

学習実験では、Stacked Autoencoder と Stacked Denoising Autoencoder を用いて DNN を構成し、表情表現の学習を行う. 学習終了後の DNN が得られた特徴量についても検討する. 表情に関しては、喜び、悲しみ、驚き、怒り、嫌悪、恐怖、無表情の 7 種類を対象とする. また DNN との表情認識の比較のために、10 人を対象とした表情認識テストを行う。

3.1 データセットについて

実験では The Japan Female Facial Expression Database (JAFPE) [6]、Cohn-Kanade database (CK) [7]、Montreal Set of Facial Displays of Emotion (MSFDE) [8]、Karolinska Directed Emotional Faces database (KDEF) [9] の 4 種類のデータセットの顔が正面を向き、顔全体が確認できる画像を 64×64 のサイズのグレースケールにして使用している. 詳細な情報と画像例を以下の表 1、図 4 に示す。

データセット	喜び	悲しみ	驚き	怒り	嫌悪	恐怖	無表情	合計
JAFPE	40	34	41	26	39	8	25	213
CK	69	28	83	44	57	24	0	305
MSFDE	31	32	0	32	32	32	0	159
KDEF	138	130	134	134	133	136	129	934
合計	278	224	258	236	261	200	154	1611

表 1: データセットと各表情の枚数

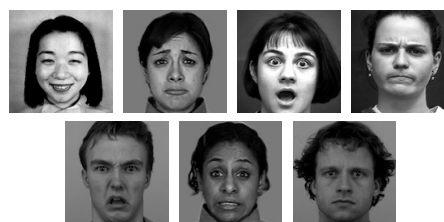


図 4: 実験に使用する画像例

今回の実験では、各表情から 10 枚ずつランダムに選んだ 70 枚をテスト用データとし、残りの 1541 枚を学習用データにした。

3.2 実験に使用した DNN について

入力ユニット数は画像データの各ピクセル値、出力ユニット数は推定する感情数に設定した。

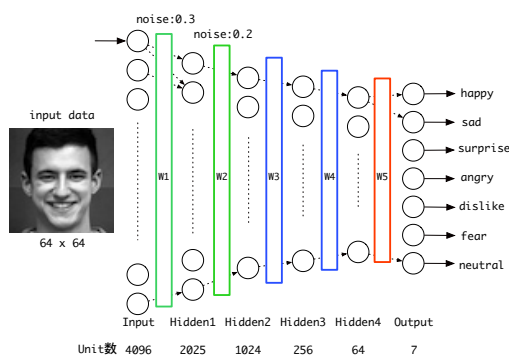


図 5: DNN モデル

DNN には、画像データの各ピクセル値を入力する. 出力層の各ユニットの値が各感情を表す数値となり、最も値の大きいユニットの出力を推定結果として使用する. 図 5 の $W1$ から $W5$ はユニット間を繋ぐ重みを表す. $W1, W2$ は DAE、 $W3, W4$ は AE を用いて重みの初期値を決定した. これはユニット数が少ないネットワークに対してノイズを加えると学習誤差が単に増加する結果となったので、今回は $W3, W4$ は AE を使用している. DAE、AE の学習終了時に DNN 全体を誤差逆伝播法で学習を行う. $W5$ は教師データを使用した全体の学習でのみ重みの調整を行う。

3.3 人間による表情認識テスト

DNN との表情認識を比較するために、10 人を対象に表情認識のペーパーテストを行った。学習用データとして各表情 1 枚ずつ計 7 枚を与え、DNN での実験と同じく各表情 10 枚ずつ、計 70 枚に対し、表情を判断してもらった。

3.4 実験結果

seed	認識率	怒り	嫌悪	恐怖	喜び	無表情	悲しみ	驚き
0	72.9%	6	10	6	8	8	6	7
10	64.3%	5	7	5	8	6	4	10
20	75.7%	9	8	5	10	8	4	9
30	64.3%	6	9	5	7	4	5	9
40	60%	8	9	2	9	4	3	7
50	72.9%	6	8	5	10	7	8	7
平均	68.35%	6.7	8.5	4.7	8.7	6.2	5	8.2

表 2: DNN による認識率 正解事例数

回答者	認識率	怒り	嫌悪	恐怖	喜び	無表情	悲しみ	驚き
1	64.3%	5	6	3	8	9	4	10
2	78.6%	7	7	2	10	10	9	10
3	72.9%	5	5	3	9	10	5	9
4	67.1%	6	4	2	9	10	6	10
5	61.4%	5	3	4	9	8	4	10
6	62.9%	4	3	1	10	10	6	10
7	72.9%	4	5	4	10	10	8	10
8	68.6%	5	5	2	9	10	7	10
9	50%	4	2	0	9	8	4	6
10	61.4%	6	2	2	9	8	6	10
平均	66.01%	5.1	4.2	2.3	9.2	9.3	5.9	9.5

表 3: 人による認識率 正解事例数

表 2、表 3 は DNN と人との表情認識の実験結果を示す。表 2、表 3 の 3 列目から 9 列目の数値それぞれ 10 枚中での正しく認識できた事例数を表す。DNN、人の両方で約 68% 程度の認識率が得られた。表情別だと喜び、驚きは両方で高い認識精度が得られている。この 2 表情は人間にとって、口の形が特徴的で他との区別が容易である表情であり、DNN の結果にも大きく影響していると考えられる。人間にとって嫌悪は、口や眉の動きが小さく恐怖や悲しみとの区別が難しいと考えられるが、DNN では嫌悪の特徴を捉え高い認識精度が得られている。また、表 2、表 3 から恐怖、悲しみは区別しづらく精度が低いことが確認できる、これより画像だけの表情では人間と DNN の表情認識に共通している部分が存在することが確認できる。

3.5 DNN で得られた特徴量

今回の実験で用いた DNN が表情を認識する為に表情画像のどの部分に反応しているのかを重みの値から検討する。学習を終えた DNN のユニット間の結合重みは正負の値を持ち、重みの値が大きければユニットの出力は有効な特徴量として次層ユニットへの出力に大きく影響することが考えられる [10]。よって、学習を終えた DNN の重みに閾値を設け、その値を超える重みに繋がれたユニットを表情認識を行う上で有効なユニットだと判断することができる。また、入力層のユニットは入力画像の各ピクセルに対応しているため、入力層ユニットの中で有効なものを選別することは、入力画像中どの部分が有効であると DNN が判断したかを確認することになる。入力画像における有効な部分を抽出した結果を図 6 に示す。抽出結果より目や口に有効とされる特徴が集まっており、背景部分には特徴が全くないことを学習している様子が確認できる。

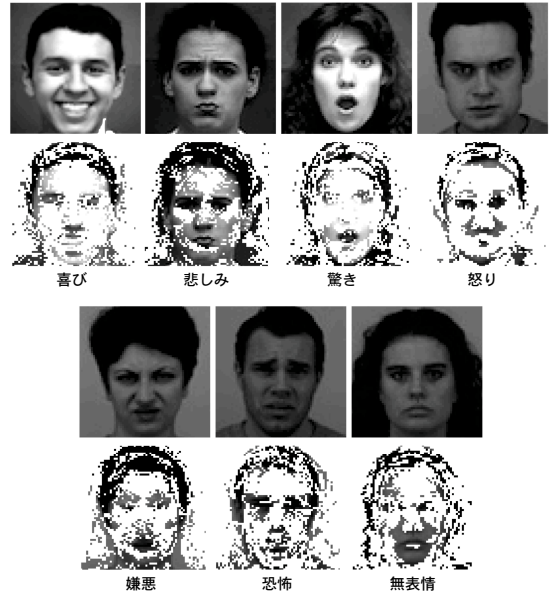


図 6: 特徴抽出結果

ここで、人手により設計された特徴量との比較を行ってみる。表情認識を行う為の特徴量として、Action Unit(AU) と呼ばれる顔の特徴点を用いた手法が提案されている [11]。この研究では、顔パーツの最大可動域を 100 として、動く幅と顔パーツを 1 つのセットとして AU 番号と数値で特徴量が明記されている。表 4、表 5 に AU の一例と無表情を除く 6 表情の AU を示す。

表 4、表 5 より、表情ごとの顔の特徴点を把握することができる。今回の実験で、DNN と人による認識の両方で精度が悪かった、恐怖と悲しみの表情に注目してみる。この二つの表情の AU で最も特徴が大きく現れるものは、どちらも”AU4:眉を下げる”になる。他にも、”AU1:眉の内側を上げる”、”AU15:唇両端を下げる”がどちらの表情にも含まれている。これより、表情を認識する為の顔特徴量が大きく現れる部分が同じであることが、DNN を用いた表情認識の結果にも影響を与えていると考えられる。

特徴抽出結果 (図 6) と AU との特徴量の比較に関して、図 6 による有効なユニットの可視化では AU との比較が難しい。図 6 では、目や口などに特徴があることが確認できるが、AU では口と言っても唇両端、下唇、上唇とより細かいポイントで定義されているのが理由として挙げられる。ただ、目や口元に有効な特徴量があるという DNN の学習によって得られた特徴は人手で設計された特徴量と共通していることが確認できる。

表情	AU 番号 (可動具合を示す値)
驚き	AU1-(40),2-(30),5-(60),15-(20),16-(25),20-(10),26-(60)
恐怖	AU1-(50),2-(10),4-(80),5-(60),15-(30),20-(10),26-(30)
嫌悪	AU2-(60),4-(40),9-(20),15-(60),17-(30)
怒り	AU2-(30),4-(60),7-(50),9-(20),10-(10),20-(15),26-(30)
喜び	AU1-(65),6-(70),12-(10),14-(10)
悲しみ	AU1-(40),4-(50),15-(40),23-(20)

表 4: 基本 6 感情の Action Unit パラメータ

AU No.	AU の内容	AU No.	AU の内容
1	眉の内側を上げる	20	唇両端を横に引っ張る
2	眉の外側を上げる	23	唇を固く閉じる
4	眉を下げる	24	唇を押し付ける
5	上脣を上げる	25	顎を下げて唇を開く
6	頬を持ち上げる	26	顎を下げて唇を開く
7	脣を緊張させる	27	口を大きく開く
8	唇を互いに接近させる	28	唇を吸い込む
9	鼻にしわを寄せる	29	下あごを突き出す
10	上唇を上げる	30	顎を左右にずらす
11	鼻唇溝を深める	32	唇を噛む
12	唇両端を引き上げる	35	頬を吸い込む
13	頬を膨らませる	41	上脣を下げる
14	えくぼを作る	43	閉眼
15	唇両端を下げる	44	細目にする
16	下唇を下げる	45	まばたく
17	オトガイを上げる	46	ウィンクする
18	唇をすぼめる		

表 5: Action Units(AU) の一例

4. おわりに

本研究では、DNN と表情画像を用いて表情表現の学習を行い、DNN が表情表現の学習において有効に機能していることを確認し、約 68% の認識率が得られた。得られた特徴量に関して、学習を終えた DNN の重みから検討を行ったが、中間層の出力自体の意味理解については更なる実験と考察、DNN 自体の構造理解が必要である。

参考文献

- [1] Paul Ekman, W.V.Friesen, 工藤 力 (訳), "表情分析入門", (1987)
- [2] 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", (2011)
- [3] VICTOR-EMIL NEAGOE, ANDREI-PETRU BRAR, NICU SEBE, PAUL ROBITU, "A Deep Learning Approach for Subject Independent Emotion Recognition from Facial Expressions", Recent Advances in Image, Audio and Signal Processing, pp.93-98 (2013)
- [4] Y. Bengio, P. Lamblin, D. Popovici and H. Larochelle, "Greedy Layer-Wise Training of Deep Networks," in Advances in Neural Information Processing Systems 19 (NIPS'06), pp.153-160, (2007)
- [5] P. Vincent, H. Larochelle Y. Bengio and P.A. Manzagol, "Extracting and Composing Robust Features with Denoising Autoencoders", Proceedings of the Twenty-fifth International Conference on Machine Learning (ICML'08), pp.1096 - 1103, (2008)
- [6] M.J.Lyons, S.Akamatsu, M.Kamachi and J.Gyoba, "Coding Facial Expressions with Gabor Wavelets", Proc. of 3rd IEEE International Conference on Automatic Face and Gesture Recognition, pp.200-205, (1998)
- [7] T. Kanade, J. F. Cohn, and Y. Tian. "Comprehensive database for facial expression analysis", In FG, pages 46—53, (2000)
- [8] Beaupré, M.G., & Hess, U. "Cross-cultural emotion recognition among Canadian ethnic groups", Journal of Cross-Cultural Psychology 36, pp.355-370, (2005)
- [9] Lundqvist, D., Flykt, A., & Öhman, A. The Karolinska Directed Emotional Faces - KDEF, CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet, ISBN 91-630-7164-9. (1998)
- [10] 久保宏一郎, "階層型ニューラルネットの学習とその応用に関する研究", pp.8-9, (1998)
- [11] 森島繁生, "表情の認識 工学の立場から", 日本医用画像工学会誌, Vol.12, No.6, pp.688-693, (1994)