H-001

顔画像を用いた個人特徴の減算によるユーザの曖昧な内部状態推定
Ambiguous User States Estimation by Subtraction of Individual Features from Facial Images

朝枝 彩夏[1)]　武村 紀子[1)]
Ayaka Asaeda　Noriko Takemura

1　はじめに*

新型コロナウイルス感染症をきっかけとして，教育現場では自宅で学習が行える e-Learning が利用される機会が増加している．また，仕事現場においてはオフィスに出社せず自宅から遠隔で業務が行える在宅勤務を導入する企業が増加している．しかし，オンラインでの作業は，長時間の着座や動画の視聴による疲労や眠気，また，周りに人がおらず緊張感がなくなることによる集中力の低下などを招く恐れがある．こうした事態を未然に防ぐために，近年，ユーザの状態を推定し注意を促すようなシステムが注目を集めている．このようなシステムの開発のためには，ユーザの正確な内部状態の把握が求められる．

内部状態の推定には表情や心拍，脳活動などさまざまなセンシングデータが用いられている．その中でも，表情は内部状態が表れやすく，データの収集も容易である点から，数多くの状態推定の研究に用いられてきた [1][2]．しかし，この表情を用いて状態推定を行う表情認識には問題点がある．人の顔に共通して備わる眉，目，鼻，口といったパーツは，大きさや間隔，角度，形などは人によって異なる．そのような顔の構造における個人差は個人識別を行う上では不可欠な要素となる一方，表情認識を行う上では精度に悪影響を及ぼす可能性がある．また，顔の構造の他にも文化的差異によって表情表出の仕方やその強弱などに個人差が生じることが知られている [3][4][5]．このような個人差の問題を解決するために考えられる方法として，年齢や性別，文化の異なるなるべく多くの人の顔画像データを収集することが挙げられる．しかし，顔画像には個人を特定するための要素が数多く含まれており，プライバシーの問題から多くのデータを収集することは容易ではない．また，機械学習による状態推定モデルを構築するためには，各データに対して内部状態のラベルが必要となるが，アノテーションにかかる時間的，金銭的コストは大きく，データセットが大規模になるほどそのコストも比例して大きくなってしまう．そのため，本研究では少量のデータセットで個人差を考慮した表情認識の手法の開発を目指す．

個人差を考慮した表情認識の研究は既にいくつか行われている．Xie ら [6] はある顔画像から抽出した表情を任意の顔画像に埋め込むことで新たな表情画像を生成する Two-branch Disentangled Generative Adversarial Network を用いて，個人差を考慮した表情認識を行った．また Liu ら [7] は特徴空間において異なる被写体の同一表情の距離が同一の被写体の異なる表情の距離よりも小さくなるように学習を行うことで，個人差を考慮した表情認識を行った．他にも 3D 顔モデルを生成して表情特徴を取り出す手法や CNN において個人特徴と表情特徴を分離する手法など表情認識における個人差の影響を抑えるために様々な試みが行われてきた [8][9][10]．これらの従来研究の多くは比較的推定が容易な基本感情を推定の対象としている．しかし，実システムへの応用を考えると，単純で明確な感情の推定だけではなく，集中や眠気，疲労といった曖昧な内部状態の推定が重要となる．そこで本研究では個人差を考慮した表情認識の中でも，曖昧な内部状態を対象とした状態推定を目指す．曖昧な内部状態は表情に表れにくく，変化も微小であるため，基本感情の推定よりも個人差による影響が大きい．例えば，覚醒度の推定であれば，眠っている状態と起きている状態は単純に瞼の開閉動作を見れば状態の判別は容易であるが，眠たい状態は瞼だけではなく，眉の角度や口の開き具合など全体的な様子から微妙な変化を読み取らなければならないため判別が難しい．

本研究では，現存の基本感情を推定の対象とした表情認識手法で高い精度が得られている Deviation Learning Network(以下，DLN)[10] に基づいて，曖昧な内部状態推定を行う．具体的には，DLN で用いられている偏差モジュールを用いて，顔画像特徴から個人特徴を減算することで個人によらない顔画像特徴を抽出し，内部状態の推定を行う．

評価実験では，e-Learning 時の学習者の顔画像データを用いて，学習者の 3 段階の覚醒度（Awake/Drowsy/Asleep）を推定し，本手法の有用性を検証する．

2　提案手法

人物が映った画像から顔領域を抽出し，顔画像を DLN に基づく状態推定モデルに入力することで内部状態の推定を行う．以下に各手法の詳細を述べる．

2.1　顔検出手法

Multi-task Cascaded Convolutional Neural Networks(以下， MTCNN)[11] を用いて画像データ中の学習者の顔領域のみを 160 × 160 で抽出し， 顔画像データを作成した．MTCNN とは，顔領域の検出を行う Proposal Network(P-Net)，P-Net の出力をもとに顔でない領域部分を候補から削除する Refine Network(R-Net)，R-Net の出力をもとに目・鼻・口部分を検出し，最終的に顔領域を出力する Output Network(O-Net) の 3 段階の CNN から構成される顔検出手法である．顔の最小検知サイズは 90 に設定し，学習者の背後に他の人が映っていても学習者以外は顔検知の対象から外れるよう処理を行った．なお，MTCNN は一定の角度以上学習者が下を向いている場合は顔検知を行うことができない．画像データを MTCNN に適用した例を図 1 に示す．

1）　九州工業大学 Kyushu Institute of Technology

＊　本論文は画像の認識・理解シンポジウム MIRU2023 においてコンセプト論文として発表した内容に基づく

切り取り前

切り取り後

図 1: MTCNN 適用例

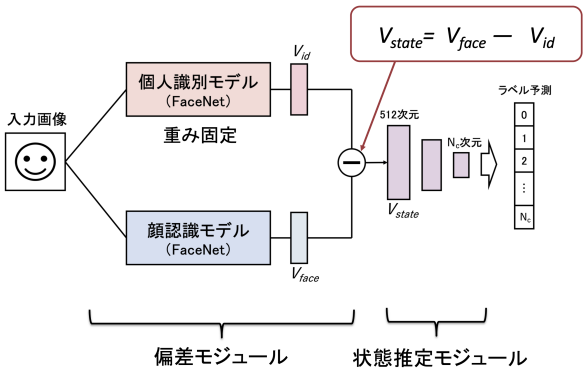


図 2: 提案手法の概要

表 1: アノテーション基準

ラベル	基準
Asleep	- 1 秒以上目を閉じている
Drowsy	- まぶたが常に開いている状態でない - 瞳孔が動かない - 目を閉じている時間が 1 秒以内である - 体や頭の動きが制御できていない
Awake	- まぶたを大きく開けている状態 - 瞳孔が左右に動く - 体や頭の動きが制御できている

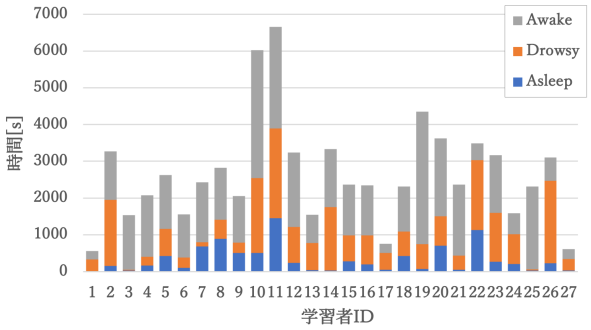


図 3: ラベルの内訳

2.2 状態推定手法

状態推定手法の概要を図 2 に示す．まず，Google が開発した FaceNet[12] と呼ばれる顔認証用のモデルを用いて個人識別モデルを学習する．FaceNet とは，顔画像から抽出された特徴をユークリッド空間へ最適な埋め込みができるよう学習を行い，生成した空間での顔画像間の距離を計算することで，顔の類似度を求めることができる手法である．個人識別モデルの学習には，様々な年齢や人種を含む 9,131 人の顔画像データセット（約 331 万枚）である VggFace2[13] を用い，FaceNet を構成する CNN としては Inception Resnet (v1)[14] を用いる．次に，個人識別モデルと全く同じ構造，同じ重みの顔認識モデルを用意し，並列ネットワークを構築する．本ネットワークを学習させる際は，個人識別モデルの重みは固定し，顔認識モデルのみを学習させる．顔認識モデルから出力される顔画像特徴 V_{face}（512 次元）から個人識別モデルから出力される個人特徴 V_{id}（512 次元）を要素ごとに減算することで，個人によらない 512 次元の顔画像特徴 V_{state} を得る．最後に，全結合層からなる状態推定モジュールにおいて内部状態の推定を行う．最終層の出力（N_c 次元，N_c は識別状態クラス数）にソフトマックス関数を適用し，損失関数は交差エントロピー誤差を用いる．また各層において，活性化関数 Rectified Linear Unit および Dropout（選出率 0.4）を適用する．

3 評価実験

提案手法の有効性を検証するために，講義動画視聴時の学習者の顔画像データを用いて覚醒度の推定を行う．

3.1 データセット

大学生 53 名の被験者に対し，講義動画（スライド＋音声，情報学に関する内容）を視聴している様子をカメラで撮影した．講義動画の視聴はノートパソコンを用いて行うものとし，ノートパソコン内蔵のカメラを用いて図 1 左のような上半身の画像を撮影する．データ収集実験は 4 回に分けて行い，被験者は 1 回の実験あたり講義動画（約 10 分）を 1〜3 本視聴する．ただし，各被験者が参加可能な回のみデータ収集を行ったため，被験者により総データ数は異なる．撮影した画像サイズは 640 × 480 画素で，フレームレートは 30fps である．また，1 秒ごとに覚醒度のラベル（Asleep（眠っている），Drowsy（眠そう），Awake（起きている））を付与している．覚醒度のアノテーションはアノテータ間での一貫性を持たせるため，表 1 に示すアノテーションの基準を設けている．ここでは，すべてのラベル (Asleep/Drowsy/Awake) についてデータが存在する被験者 27 名について評価実験を行う．各被験者のデータにおけるラベルの内訳を図 3 に示す．

3.2 比較手法

個人差を考慮した状態推定モデルである提案手法の有効性を検証するため，偏差モジュールの代わりに顔認識モデルのみで覚醒度の推定を行う手法を用いて比較実験を行う．ただし，比較手法を学習する際の顔認識モデルの重みの初期値は提案手法と同じものを用いる．

比較手法の概要を図 4 に示す．

また，本実験では状態推定モジュールにおける全結合層の数を 1 層（512 → 3 次元），2 層（512 → 128 → 3 次元），3 層（512 → 128 → 32 → 3 次元）とする 3 つのパターンについて推定精度の比較を行った．

3.3 評価手法

27 名分の被験者を 3 名ずつ，9 グループに分割し，Leave one-group out 交差検証を行う．1 グループをテストデータ，他の 1 グループを検証データ，残りの 7 グループを学習データとし，すべてのグループが 1 回ずつ

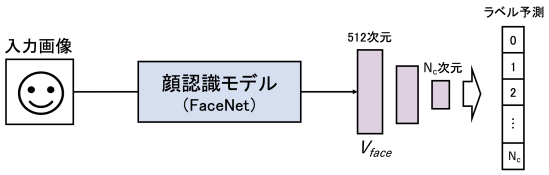


図 4: 比較手法の概要

表 2: 各グループのデータ

	Asleep	Drowsy	Awake
group1	5430	5427	5381
group2	18630	19446	19698
group3	12752	17881	19405
group4	46152	49583	49587
group5	6621	10426	10448
group6	14457	14244	14330
group7	12810	15897	16033
group8	27114	27583	26976
group9	8400	8324	8390

テストデータになるように計 9 回実験を行い，9 回分の評価値の平均により性能評価を行う．表 2 に各グループにおける各ラベルの内訳を示す．なお，データに対しては被験者ごとにアンダーサンプリングを行っている．評価指標には多クラス分類においてクラスごとに計算した F1 score の平均を取る macro-F1 score を用いる．

以下，本実験で使用したハイパーパラメータの設定について述べる．本実験ではエポック数は 3 に設定した．また，ミニバッチ学習を用いており，ミニバッチサイズは 128 に設定した．学習率の初期値は 0.001 に設定し，1 エポック目と 2 エポック目が終わった段階で学習率を 0.1 ずつ乗算するように設定した．加えて，過学習抑制の手法のひとつである weight decay を導入し，値は 0.001 と設定した．本実験では 100 バッチごとに検証データを用いてモデルの性能を測り，その中で macro-F1 score の値が最も高かったモデルを用いて，テストデータによる評価を行った．Optimizer には Stochastic Gradient Descent(SGD) を用いた．

3.4 実験結果

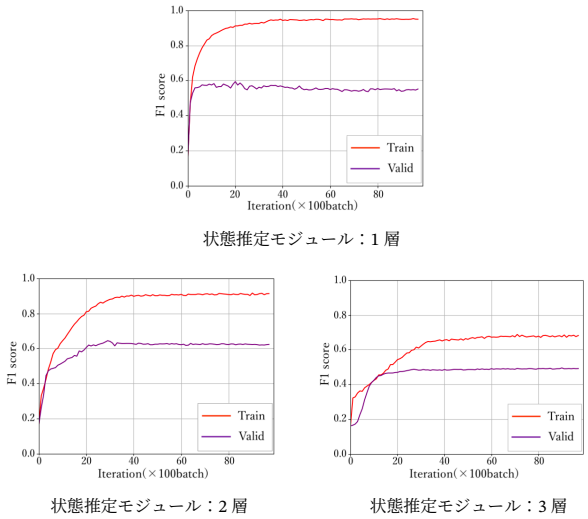
表 3 に提案手法及び比較手法における macro-F1 score の結果を示す．提案手法と比較手法の macro-F1 score を比較すると，すべてのパターンで提案手法が比較手法の精度を上回った．このことから，偏差モジュールにより顔画像特徴から個人の特徴を減算し，個人によらない顔画像特徴を抽出することが曖昧な内部状態推定に有効であることがわかる．

状態推定モジュールの層の数についての考察

状態推定モジュールにおける層の数による精度の違いについて比較する．図 5 に学習及び検証データにおける学習曲線を示す．各次元削減パターンを比較すると，提案手法と比較手法ともに 512 次元から 128 次元に削減した後，3 次元に削減する 2 層の場合が最も高い値を示した．一般に，ニューラルネットワークは層を深くすることでパラメータ数が多くなり，モデルの表現力が上がる

表 3: 提案手法及び比較手法の macro-F1 score

状態推定モジュール	提案手法	比較手法
512 → 3	**0.526**	0.487
512 → 128 → 3	**0.555**	0.505
512 → 128 → 32 → 3	**0.509**	0.474



状態推定モジュール：1 層

状態推定モジュール：2 層

状態推定モジュール：3 層

図 5: 学習曲線（状態推定モジュールの層数を変化）

ことが知られている．しかしパラメータ数が多いほど，最適化を行う際の探索空間は複雑になり，学習が難しくなる傾向にある．層の数が 1 層，2 層，3 層における学習曲線をそれぞれ比較すると，学習時における精度推移は 1 層や 2 層の場合は 90%近くで収束しているのに対して，3 層の場合では 70%程と低い傾向にあった．このことから，3 層は上手く学習を行うことができず，結果 1 層よりもパラメータ数が多く表現力のある 2 層が最も精度の良いモデルとなったと考えられる．

予測値の内訳についての考察

状態推定モジュールを 2 層にした場合における提案手法の混合行列を表 4 に示す．ただし，本混合行列は 9 回分の交差検証のテストデータの結果を全て足し合わせたものである．正解ラベルと予測ラベルが一致している箇所を見ると，Asleep と Awake は比較的正しく推定できているのに対し，Drowsy はあまり正しく推定できていないことがわかる．さらに詳細な内訳を見ると，“Asleep/Awake” および “Asleep/Drowsy” は高い精度で識別できているが，“Drowsy/Awake” の識別を失敗しているケースが多い．これは，Asleep は完全に瞼を閉じた状態であるのに対し，Drowsy および Awake はどちらも瞼は開いた状態であり，微妙な瞼の開き具合や眉の角度，口形の微妙な違いに基づいて判断しなければならないためである．また，このことは同時にアノテーションの難しさにも繋がる．Drowsy についてはアノテータにより判断が異なるケースが存在し，このアノテーションの曖昧さが推定精度の低下をもたらしている可能性も考えられる．

表 4: 提案手法の混同行列

		正解ラベル		
		Asleep	Drowsy	Awake
予測ラベル	Asleep	110499	22778	10972
	Drowsy	23512	63907	56070
	Awake	18355	82126	103206
計		152366	168811	170248

4 終わりに

表情認識において，個人の顔の作りや表情の表出方法の違いが認識精度に影響を及ぼすことが問題として挙げられている．そこで，本研究では個人差の影響を低減させる偏差モジュールを用いて，基本感情よりも推定が難しい曖昧な内部状態の推定を行う手法を提案した．e-Learning 中の学習者の顔画像データを用いて覚醒度推定問題について評価実験を行った結果，提案手法の有効性を示すことができた．

しかしながら，問題点として眠そうな状態と起きている状態を正確に識別することが難しい点が挙げられる．これは，眠そうな状態と起きている状態の判別が難しく，アノテーションに曖昧性があることも原因の一つであると考えられる．そこで今後の展望として，曖昧なアノテーションをソフトラベルとして扱うなど，アノテーションの曖昧性の問題に対処していく．また，現在は状態推定の際に 1 枚の画像のみを入力としているが，動きの情報は曖昧な状態を推定する上で有用であると考えられるため，時系列データを入力できるように本手法を拡張する．

参考文献

[1] Ji-Hae Kim, Byung-Gyu Kim, Partha Pratim Roy, and Da-Mi Jeong. Efficient facial expression recognition algorithm based on hierarchical deep neural network structure. *IEEE Access*, Vol. 7, pp. 41273–41285, 2019.

[2] Hongli Zhang, Alireza Jolfaei, and Mamoun Alazab. A face emotion recognition method using convolutional neural network and image edge computing. *IEEE Access*, Vol. 7, pp. 159081–159089, 2019.

[3] Hillary Elfenbein, Martin BeauprÃ©, Manon LÃ©vesque, and Ursula Hess. Toward a dialect theory: Cultural differences in the expression and recognition of posed facial expressions. *Emotion (Washington, D.C.)*, Vol. 7, pp. 131–46, 03 2007.

[4] Wallace V. Friesen. Cultural differences in facial expressions in a social situation: An experimental test on the concept of display rules. 1973.

[5] Paula Niedenthal, Magdalena Rychlowska, Fangyun Olivia Zhao, and Adrienne Wood. Historical migration patterns shape contemporary cultures of emotion. *Perspectives on psychological science : a journal of the Association for Psychological Science*, Vol. 14, p. 1745691619849591, 06 2019.

[6] Siyue Xie, Haifeng Hu, and Yizhen Chen. Facial expression recognition with two-branch disentangled generative adversarial network. *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 31, No. 6, pp. 2359–2371, 2021.

[7] Xiaofeng Liu, B.V.K. Vijaya Kumar, Ping Jia, and Jane You. Hard negative generation for identity-disentangled facial expression recognition. *Pattern Recogn.*, Vol. 88, No. C, pp. 1–12, apr 2019.

[8] Zibo Meng, Ping Liu, Jie Cai, Shizhong Han, and Yan Tong. Identity-aware convolutional neural network for facial expression recognition. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pp. 558–565, 2017.

[9] Mohammad Rami Koujan, Luma Alharbawee, Giorgos Giannakakis, Nicolas Pugeault, and Anastasios Roussos. Real-time facial expression recognition “in the wild” by disentangling 3d expression from identity. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pp. 24–31, 2020.

[10] Wei Zhang, Xianpeng Ji, Keyu Chen, Yu Ding, and Changjie Fan. Learning a facial expression embedding disentangled from identity. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6755–6764, 2021.

[11] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, Vol. 23, No. 10, pp. 1499–1503, 2016.

[12] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 815–823, 2015.

[13] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2018)*, pp. 67–74, 2018.

[14] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI’17, pp. 4278–4284. AAAI Press, 2017.