Vol.2022-CG-185 No.10
2022/3/12

顔の特徴点を利用した人物顔画像の感情推定

金森 透有[†1,a)]　金森 由博[†1,b)]　遠藤 結城[†1,c)]　三谷 純[†1,d)]

概要：人物の顔画像からの感情推定は、ニューラルネットワークを用いた教師あり学習手法により一定水準の精度に達している。本研究では、入力として顔画像だけでなく顔の特徴点を利用することで、さらなる精度向上を目指す。顔の特徴点のみを入力とした推定結果を、顔画像のみで学習済みのネットワークによる推定結果にマージすることによって精度向上を図るが、単体では前者の精度は後者に劣り、単純に合算しても精度は向上しない。そこで本研究では、顔画像のネットワークの出力は固定し、顔の特徴点のみの結果に対して、マージ時の重みの調整方法を 2 つ提案する。1 つ目は、推定の不確実さを利用する。具体的には、顔の特徴点による感情推定時に得られる各感情カテゴリの確率のうち最大のものを確信度と見なし、重みとして用いる。2 つ目は、顔の特徴点のみの推定結果が顔画像のみの推定結果と異なっている場合に重みを小さくする。以上により、提案手法が顔画像のみを入力としたベースライン手法による推定精度を上回ることを示す。

1. はじめに

人間とコンピュータとの相互のやりとりにおいて、人物画像を入力とした感情推定は、ロボット工学の分野や自動車の運転アシストなど様々な場面で有用である。人物画像の中でも顔画像に注目した感情推定の研究として、人物顔画像を入力とした感情推定の手法 [4], [16], [17] が提案されている。これらの手法では、畳み込みニューラルネットワークを用いた教師あり学習によって感情を推定する。すでに一定水準の精度には達しているが、まだ精度向上の余地があると考えられる。既存手法では人物の顔の RGB 画像のみを入力としていているが、顔の部位 (鼻や目など) の位置情報を入力に加えることで、顔をより正確に認識でき、精度向上につながる可能性がある。

本研究では、人物顔画像だけでなく、別途推定された顔の特徴点の情報も入力することで、感情推定の精度向上を目指す。具体的には、顔の RGB 画像を入力とした既存手法 [17] による感情推定の結果と、顔の特徴点を入力とした新たなネットワークによる感情推定の結果をマージする。このとき、後者の推定精度は前者の推定精度に劣り、単純に両者の推定結果を合算すると、むしろ顔の RGB 画像のみを入力とした推定精度よりも劣化してしまう。また、前者の学習済みネットワークは単独ですでに高い精度を達成できており、fine tuning などさらなる学習を試みると、むしろ精度が低下することを確認した。そこで本研究では、精度を維持するため前者の学習済みネットワークのパラメータを固定し、後者のネットワークから得られる特徴量を、推定の信頼度に基づいて重み付けし、精度向上を図る。まず、顔の特徴点を入力としたネットワークの出力である、各感情カテゴリの確率を得る。その中で最も高い確率を、顔の特徴点に基づく推定結果の確信度だと見なし、重み付けに利用する。さらに、もし RGB 画像を入力としたネットワークと顔の特徴点を入力としたネットワークの感情推定の結果が食い違っていた場合、顔の特徴点を入力としたネットワークの信頼度が低いものと見なし、その特徴量の重みを減らす。これらの重み付けを施した上で、RGB 画像を入力としたネットワークの出力と、顔の特徴点を入力としたネットワークの出力を連結し、最終的な推定結果を得る。これにより、RGB 画像のみを入力とした既存手法では推定に失敗するような顔画像に対して提案手法では正しく推定できることを、定量的および定性的に示す。

2. 関連研究

2.1 コンテキストを考慮した感情推定

人物画像を入力とした既存手法の多くは、人物の顔、場合によっては人物の姿勢など、人物のみに着目して感情推定を行なってきた。しかし心理学の研究 [2], [3] によると、シーンのコンテキスト (例えば、テニス選手がテニス場で

†1　現在，筑波大学
Presently with University of Tsukuba
a)　michiari123@gmail.com
b)c)d)　{kanamori,endo,mitani}@cs.tsukuba.ac.jp

テニスをしているという状況) が人物の感情に重要な影響を与えていることを示唆している。そこで、Kosti らの研究 [9] では、人物だけでなく、その周りの情報も学習させて感情推定を行なっている。Mittal らの研究 [12] では、心理学のフレーゲの文脈原理 [15] に基づいて、1) 顔や人物の姿勢、2) 感情推定を行いたい人物をマスクした画像全体、3) 人物同士や人物と物体間の物理的な距離を測るための深度マップ、の 3 つを入力としている。

本研究でも当初、Mittal らの手法およびそのデータセットを利用して、コンテキストを考慮した感情推定を検討した。しかし、Mittal らの手法でも精度が低く、データセットを確認すると人間でも判断の難しい画像が多数見つかった。現時点ではコンテキストの考慮は難しいと考え、感情が直接的が現れると考えられる、人物の顔に特化した手法を検討した。

2.2 CNN を用いた感情推定

顔画像を入力とした最近の手法は、畳み込みニューラルネットワーク (CNN) を用いている [4], [16]。例えば Burkert らの研究 [4] では、MMI データセット [13] で 98.4 %、CKP データセット [14] で 99.6 %と高い精度を達成している。しかし、これらのデータセットの画像は、実験室内の青い背景で人物が撮影されていたり [13]、正面顔に限定して背景が完全に除去されていたり [14] と、現実的でない条件設定となっている。Shi と Zhu の研究 [17] では、*Amend Representation Module* (ARM) ブロックを導入し、背景、横顔やグレースケール画像など、多様な画像を含む RAF-DB [10] というデータセットで高精度を達成した。本研究は彼らの手法をベースラインとし、彼らと同じく RAF-DB を用いて実験を行った。彼らの手法と RAF-DB については、それぞれ 3 節と 4.4 項で説明する。

2.3 Transformer を用いた感情推定

深層学習に基づく既存手法の多くは上述の CNN を用いているが、自然言語処理の分野で提案された Transformer を画像処理に適応させた *Vision Transformer* (ViT) [6] が提案され、ViT に基づく感情推定の手法も登場している。例えば最新の手法の 1 つである Aouayeb らの研究 [1] が挙げられるが、必ずしも CNN ベースの手法より優れている訳ではなく、本研究のベースライン [17] よりも精度は劣っている。

3. ベースライン手法

本研究でベースラインとして採用した、CNN を用いた最新手法の 1 つである Shi と Zhu の手法 [17] について説明する。彼らの研究では ResNet [7] をバックボーンネットワークとして利用しているが、ResNet を用いた教師あり学習では精度が十分でないとして、その理由を 2 つ挙げている。まず、1) 元々 ResNet は顔以外の物体 (建物、動物など) を含む画像データセットで訓練されているため、顔に特化した推定を行いづらいという点である。次に、2) 畳み込み層やプーリング層などでのゼロパディングによって、特徴マップの端の画素にゼロの値が意図せず導入されてしまい、精度を下げている点である。彼らの研究では、1) への対策として、ミニバッチごとに指数平滑移動平均 (EMA) を取り、一般物体認識用に学習された ResNet を、顔に特化した分類タスクである感情推定に転用した。2) への対策としては、ゼロで埋められた画素を除外して畳み込みを行う *Amend Representation Module* (ARM) ブロックを導入した。

本研究でもこれら 2 つの工夫は踏襲する。さらに本研究では、顔特徴点を入力としたネットワークに、新たに設計した CNN に加えて ARM ブロックを採用している。

4. 提案手法

提案手法の概要を説明する。入力は、ベースライン手法 [17] の入力である人物の顔画像 I および、人物の顔の特徴点を表すヒートマップ $I_{heatmap}$ である。特徴点は、検出器で自動抽出されたものを用いる (4.4 項参照)。顔画像 I をベースラインネットワーク (**図 2** 上段)、ヒートマップ $I_{heatmap}$ を、本研究で新たに導入した顔特徴点ネットワーク (図 2 下段) にそれぞれ入力し、両者から得られた特徴マップをマージして感情推定を行う。当初の実験の結果、学習済みベースラインネットワークを fine tuning したり、ベースラインネットワークを含めネットワーク全体をゼロから学習したりしても、むしろベースラインよりも精度が悪化することがわかった。そこで、ベースラインの精度を維持するために、ベースラインネットワークのパラメータは固定し、顔特徴点ネットワークの特徴マップに重み付け (4.2 項参照) を行ってからマージ処理を行う。以下、それぞれの詳細を述べる。

4.1 ネットワークモデル

提案手法のネットワークモデルについて説明する。まず、人物顔画像のネットワーク (図 2 の “RGB-based Network”) では、人物顔画像 I を入力として、ResNet18 とベースライン手法の ARM ブロック (3 節) を使用して人物顔画像の特徴マップ F_{RGB} を得る。ここで図 2 中の “RGB-based Network” については、学習済みモデルの精度を保つためにパラメータ更新を行わない。次に、人物の顔の特徴点のネットワーク (図 2 の “Landmark-based Network”) では、人物顔画像 I から、RAF-DB で提供されている目や鼻などの顔の特徴点の座標を白い点で表したグレースケールのヒートマップ画像 $I_{heatmap}$ を生成する。そして、新たに構築した畳み込みニューラルネットワーク (図 2 の “Landmark CNN”) と、ARM ブロックを使用して顔の特

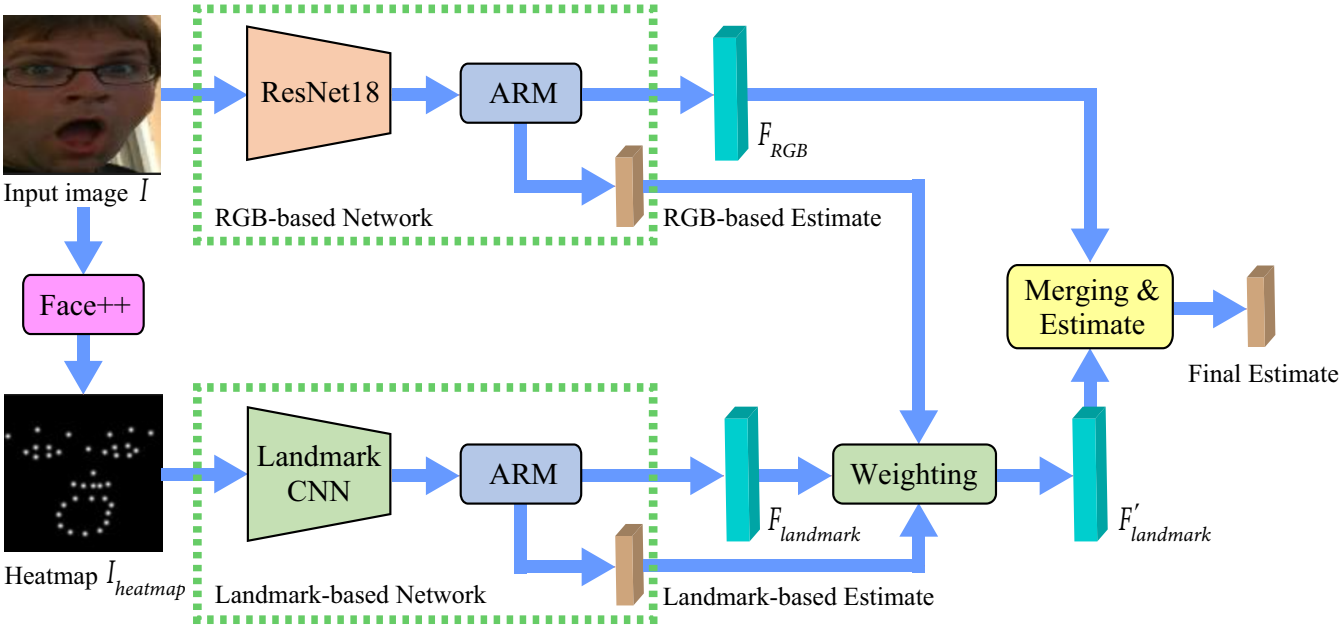


図 2 提案手法のネットワーク構造。

表 1 顔の特徴点ネットワークの構造。Output size の括弧内の数字は順に、縦方向の画素数、横方向の画素数、チャンネル数を表す。

Layer	Kernel size	Stride	Output size
input			(224,224,1)
conv	5	1	(220,220,32)
maxpool	2	2	(110,110,32)
conv	3	1	(108,108,64)
maxpool	2	2	(54,54,64)
conv	3	1	(52,52,128)
maxpool	2	2	(26,26,128)
conv	3	1	(24,24,256)
maxpool	2	2	(12,12,256)
conv	3	1	(10,10,512)
conv	4	1	(7,7,512)

徴点のヒートマップ画像の特徴マップ $F_{landmark}$ を得る。特徴マップ F_{RGB} および $F_{landmark}$ は、それぞれのネットワークの最終層の 1 つ手前から抽出している。

顔の特徴点のネットワーク (図 2 の “Landmark-based Network”) について、新たにネットワーク (図 2 の “Landmark CNN”) を導入する。具体的な内部構造を**表 1** に示す。その後、このネットワークの出力を ARM ブロック (3 節) に入力する。このネットワークをベースライン手法 [17] の ARM ブロックに接続するために、最後の畳み込み層で、サイズを調整している。その後、顔の特徴点のヒートマップ画像の特徴マップに関して、重み付け (図 2 の “Weighting”) を行う。具体的な重み付けの方法に関しては、4.2 項で詳しく説明する。

最終的に、特徴マップ F_{RGB} と重み付けされた特徴マップ $F'_{landmark}$ をチャンネル方向に結合し、全結合層を経て感情推定結果を得る。

4.2 顔の特徴点ネットワークの出力の重み調整

顔の特徴点ネットワークが出力する特徴マップに対する重み付け (図 2 の “Weighting”) の方法について説明する。重み付けには次の 2 つを考慮した:

(1) 顔の特徴点ネットワーク単独での推定結果の確信度
(2) 顔画像ネットワークの推定結果との一致・不一致

まず (1) について、Malinin ら [11] によれば、訓練データの不足、訓練データとテストデータの統計量の不一致など、様々な要因により、ネットワークの推定結果には不確実性が生じる。本研究では推定結果の確信度として、各感情カテゴリ c の確率 p_c のうち最大のものを $p_{landmark}$ とし、重み付けに用いる。

$$p_{landmark} = \max_c p_c \tag{1}$$

次に (2) について、顔画像ネットワークの推定精度に比べ、特徴点ネットワークの推定精度は劣る場合が多い。実際に実験したところ、前者の正解率は 0.92 程度であったのに対し、後者の正解率は 0.68 程度であった。しかし一方で、**表 2** に示す通り、前者では不正解だが後者は正解するデータも存在するため、単純に後者を無視するのではなく、後者の推定結果も考慮することは有益だと考えられる。そこで前者の推定結果と一致しない場合、後者の特徴マップに重み $w < 1$ を掛けて、後者の推定結果の影響を小さくする (一致する場合は $w = 1$)。本研究では実験により $w = 0.6$ とした (実験結果については 5.3 項参照)。

重み付けされた特徴点ネットワークの特徴マップ $F'_{landmark}$ は次式で与えられる。

$$F'_{landmark} = p_{landmark} \cdot w \cdot F_{landmark} \tag{2}$$

ここで $F_{landmark}$ は重み付け前の特徴点ネットワークの特

表 2 RAF-DB [10] のテストデータに対する顔画像と顔の特徴点のネットワークによる推定の正解と不正解の一致度合いの表。例えば右上は、顔画像のネットワークでは推定に成功しているが、顔の特徴点のネットワークでは推定に失敗しているデータの数を表す。

	顔の特徴点が入力で正解	顔の特徴点が入力で不正解
顔画像が入力で正解	2,018	806
顔画像が入力で不正解	62	182

徴マップである。

4.3 ネットワークモデルの学習

ネットワークモデルの学習について説明する。ベースライン手法の著者が公開しているコードを使用して訓練を行ったところ、ハイパーパラメータを調節するなど様々な工夫を行ってもベースライン手法の精度を再現できなかった。そこで本研究では、ベースラインのネットワークには事前に学習させたモデルを適用する。また、ベースラインの学習済みモデルを初期値として、顔の特徴点ネットワークを組み込んだパイプラインを学習すると、顔の特徴点ネットワークの学習が妨げとなって、全体のネットワークの学習がうまく進まなかった。そこで、人物の顔画像ネットワークの学習済みモデルのネットワークパラメータは固定した状態 (図 2 の赤い点線部分参照) で、顔の特徴点のネットワークを訓練させる。損失関数には、最終的な推定結果について正解とのクロスエントロピー損失を用いた。式 (3) に提案手法の損失関数を示す。

$$\mathcal{L} = \mathcal{L}_{Landmark} + \mathcal{L}_{overall} \tag{3}$$

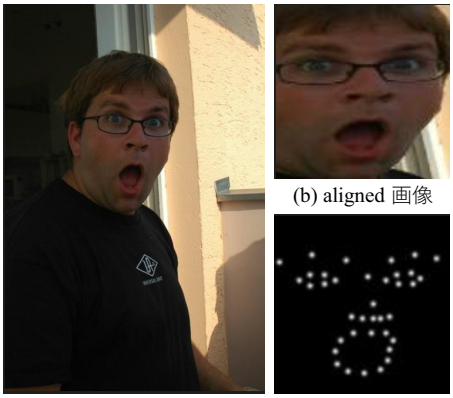
ここで $\mathcal{L}_{Landmark}$ は顔の特徴点のネットワークに対する損失、$\mathcal{L}_{overall}$ は顔画像と顔の特徴点のネットワークの出力をマージしたネットワークに対する損失である。それらを加算した損失を、全体の損失としている。損失の重み係数は全て 1 である。

4.4 データセット

本手法では、ベースライン手法 [17] でも使われている RAF-DB [10] という感情認識のための人物顔画像データセットを用いる。このデータセットには、実写の人物顔画像と、それに対応する 7 種類の感情カテゴリ (Surprise、Fear、Disgust、Happiness、Sadness、Anger、Neutral) が含まれている。**表 3** にそれぞれの感情カテゴリの画像の枚数を示す。感情カテゴリのラベル付けは主観的なタスクであるため、1 枚ずつ 40 人によってラベル付けされている。RAF-DB [10] では、顔画像 12,271 枚分が訓練データ、3,068 枚分がテストデータとなっており、画像の解像度は 224 × 224 画素となっている。さらに、人物の顔画像に対応する顔の特徴点もデータセットに加える。

表 3 RAF-DB の各感情カテゴリの画像枚数の内訳。

	Surprise	Fear	Disgust	Happiness	Sadness	Anger	Neutral
訓練データ	1,290	281	717	4,772	1,982	705	2,524
テストデータ	329	74	160	1,185	478	162	680



(a) original 画像　(b) aligned 画像　(c) ヒートマップ

図 3 RAF-DB の画像データの例。(c) ヒートマップは RAF-DB に含まれる特徴点データより独自に生成。画像の出典: 文献 [10]

顔の特徴点データの生成方法について説明する。RAF-DB [10] には背景も含まれた人物画像である original 画像 (**図 3**(a)) と、original 画像から顔の部分だけを切り取って整形された aligned 画像 (図 3(b)) が含まれており、ベースライン手法 [17]、本手法ともに感情推定には aligned 画像を入力として使用している。また、このデータセットには original 画像に対する、自動検出された 37 点の顔の特徴点の座標データ[*1]と手動でつけられた 5 点（両目、口の両端、鼻）の特徴点座標データが含まれているが、aligned 画像に対する特徴点の座標データは含まれていない。そこで original 画像から aligned 画像への変換方法を著者に確認し、original 画像の特徴点の座標データを aligned 画像に合わせた。具体的には、37 点の顔の特徴点以外に提供されている手動で付けられた 5 点（両目、口の両端、鼻）の顔の特徴点の座標データを使用し、左目と右目と口の中央（口の両端の位置の真ん中）の座標をそれぞれ (25,35)、(75,35)、(50,75) となるようにアフィン変換を行った。そして、CNN のプーリング層の役割を担っている ARM ブロックに接続できるように、画像を入力とする必要があるため、アフィン変換後の座標を用いて図 3(c) のようなヒートマップ画像を作成した。

5. 実験

5.1 実験環境

提案手法を Python および PyTorch を用いて実装し、NVIDIA RTX A5000 上で学習・推論を行った。ベースライン手法ではオプティマイザとして *Adam* [8] を使用していたが、提案手法では他の最先端手法でよく利用されてい

[*1] 顔の特徴点の検出器としては “Face++ Cognitive Services” `https://www.faceplusplus.com/` が用いられている。

表 4 本研究でのオプティマイザ AdaBelief [18] のパラメータ。

epslion	weight decouple	rectify	lr	betas	weight decay
1e-16	True	False	1e-3	(0.9,0.999)	1e-2

表 5 ベースライン手法と提案手法に対する各カテゴリの感情推定の精度比較。より精度が高い方を太字で示す。

	Surprise	Fear	Disgust	Happiness	Sadness	Anger	Neutral
ベースライン	0.8997	**0.8514**	0.6687	**0.9637**	**0.8808**	0.8025	0.9779
提案手法	**0.9058**	**0.8514**	**0.6875**	0.9629	0.8766	**0.8210**	**0.9824**

表 6 ベースライン手法と提案手法に対する感情推定の Overall Accuracy と Average Accuracy の比較。OA (Top k) は Top k の Overall Accuracy、AA は Average Accuracy を表す。より精度が高い方を太字で示す。

	OA (Top 1)	OA (Top 2)	OA (Top 3)	AA
ベースライン	0.9205	**0.9775**	**0.9886**	0.8635
提案手法	**0.9231**	0.9772	0.9870	**0.8696**

る *AdaBelief* [18] を採用した[*2]。具体的なパラメータの値を**表 4** に示す。値に関しては、実験により設定した。学習率スケジューラは PyTorch の API である ExponentialLR を用い、gamma を 0.9 とした。バッチサイズは訓練データに対しては 256、テストデータに対しては 64 とした。訓練にかかった時間は、1 つの GPU を用いて 3 チャンネルの 224 × 224 画素の顔画像データおよび、1 チャンネルの 224 × 224 画素の顔の特徴点のグレースケールのヒートマップ画像を入力した場合、1 エポックあたり約 15 分であった。推論にかかる時間は 224 × 224 画素の顔画像データおよび、224 × 224 画素の顔の特徴点のグレースケールのヒートマップ画像を入力とすると、テストデータ全てに対して約 1 分であった。

5.2 ベースライン手法と提案手法との比較

ベースライン手法と提案手法を使って、本研究で用意したテストデータ 3,068 枚の感情推定の結果の 7 つの感情カテゴリの正解率を**表 5** に示す。また、Overall Accuracy (Top 3 まで) と Average Accuracy の値を**表 6** に示す。正解の顔画像のサンプルに関しては、**表 9** や**表 10** に示す画像を参照されたい。定量的比較では、4 つのカテゴリに対する推定精度がベースライン手法よりも提案手法で改善している。また、Top 1 の Overall Accuracy と Average Accuracy でも推定精度がベースライン手法よりも提案手法で改善している。ここで Top k とは、モデルの予測した感情カテゴリの上位 k 位までに正解が含まれる割合を表す。

5.3 顔の特徴点ネットワークの重みの値の比較

式 (2) における顔の特徴点ネットワークの重み w を変化させると、ベースラインとの Overall Accuracy の差がどう変化するのかを検証した。**図 4** に検証結果を示す。結論として $w = 0.6$ のときにベースラインとの Overall Accuracy の差が最も大きいということがわかった。よって本研究の他の実験では $w = 0.6$ を採用した。

[*2] AdaBelief のバージョンは 0.2.0 を使用した。

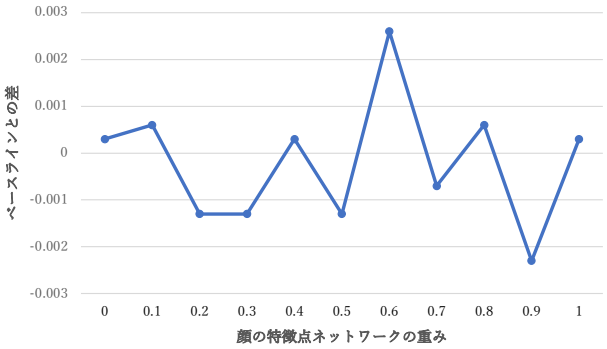


図 4 顔の特徴点ネットワークの特徴量の重み w を変化させた場合のベースラインとの Overall Accuracy の差。$w = 0.6$ のときに Overall Accuracy の差が最も大きい。

表 7 重み付けに関する各カテゴリに対するアブレーションスタディ。精度が最も高い値を赤字かつ太字で表し、精度が 2 番目に高い値を青字かつ太字で示す。

	Surprise	Fear	Disgust	Happiness	Sadness	Anger	Neutral
考慮なし	**0.9058**	**0.8514**	**0.7188**	0.9595	0.8661	0.7963	0.9794
不確実性のみ	**0.9149**	**0.8649**	0.6562	**0.9629**	0.8640	0.7901	**0.9853**
データ比較のみ	0.9027	**0.8514**	0.6813	**0.9646**	**0.8724**	**0.8025**	**0.9824**
両方考慮	**0.9058**	**0.8514**	**0.6875**	**0.9629**	**0.8766**	**0.8210**	**0.9824**

表 8 重み付けに関する全体のアブレーションスタディ。OA (Top k) は Top k の Overall Accuracy、AA は Average Accuracy を表す。精度が一番高い値を太字で示す。

	OA (Top 1)	OA (Top 2)	OA (Top 3)	AA
考慮なし	0.9198	0.9739	0.9876	0.8682
不確実性のみ	0.9198	0.9729	0.9857	0.8626
データ比較のみ	0.9214	0.9752	**0.9883**	0.8653
両方考慮	**0.9231**	**0.9772**	0.9870	**0.8696**

5.4 重み付けに関するアブレーションスタディ

重み付けに関してアブレーションスタディを行った。比較対象は、1) 重み付け考慮なし、2) 不確実性 ($p_{landmark}$ による重み付け) のみ考慮、3) データ比較 (w による重み付け) のみ考慮、4) 両方考慮、の 4 つである。各カテゴリの比較結果を**表 7** に示す。結果として、全てのカテゴリにおいて、両方考慮したモデルが上位 2 位以内の精度となった。また、全データに対する比較結果を**表 8** に示す。結論として、Top 1 と Top 2 の Overall Accuracy と、Average Accuracy に関して不確実性とデータ比較の両方を考慮したモデルが最も精度が高いことがわかった。

5.5 分類に成功している画像に対する定性評価

提案手法に対して、どのような画像の分類が成功しているのか分析を試みた。表 9 に、分類に成功している画像を示す。また、CNN が画像のどの部分を見て分類をしているのかを判断するために、CNN 可視化手法の最新手法

表 9 分類に成功した画像に対する Grad-CAM++による重要度可視化結果。図中の人物顔画像の出典: 文献 [10]

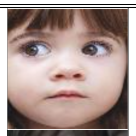
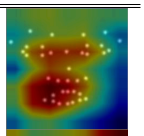
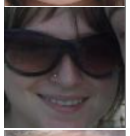
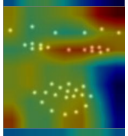
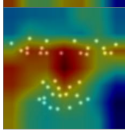
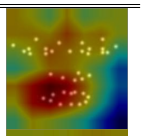
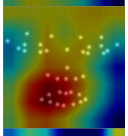
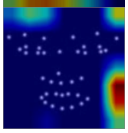
正解ラベル	予測ラベル	入力画像	顔画像可視化	特徴点可視化
Surprise	Surprise			
Happiness	Happiness			
Sadness	Sadness			

表 10 分類に失敗した画像に対する Grad-CAM++による重要度可視化結果。図中の人物顔画像の出典: 文献 [10]

正解ラベル	予測ラベル	入力画像	顔画像可視化	特徴点可視化
Disgust	Anger			
Surprise	Disgust			
Happiness	Neutral			

である *Grad-CAM++* [5] を、ベースラインネットワークと特徴点ネットワークの出力に適用した[*3]。表 9 の 4 列目と 5 列目の画像において、赤くなっている部分は CNN が着目している箇所、青くなっている部分は CNN が着目していない箇所である。表 9 を見ると、1 行目や 3 行目の Grad-CAM++の可視化画像では、CNN は RGB 画像と顔の特徴点画像のどちらも同じような場所に着目している。一方で、2 行目の画像の Grad-CAM++の可視化画像では、CNN は RGB 画像に対しては口元の辺りに着目しているが、顔の特徴点画像に対しては目元の辺りに着目していることが分かる。RGB 画像の CNN が口元の辺りに注目している理由として、入力画像の女性がサングラスをかけていて目元からでは感情推定の判定が難しいと判断したからであると考えられる。また、2 行目の画像の人物の口元が歯を見せていることから、CNN が Happiness と判断したと考えられる。このように、顔のいずれかの部位を見ることによって感情を推定することができるような画像に対して、モデルが分類に成功していると考えられる。

5.6 分類に失敗している画像に対する定性評価

次に、提案手法に対してどのような画像の分類が失敗しているのか分析を試みた。表 10 に、分類に失敗している画像と Grad-CAM++による可視化画像を示す。表 10 の 1 行目と 2 行目の画像については、そもそも入力画像の表情が曖昧であるため、分類に失敗したと考えられる。3 行目の画像に関しては、RGB 画像、顔の特徴点画像ともに CNN での信頼性が低いようなヒートマップ画像になっており、上手く表情を捉えられていないことが分かる。

5.7 ネットワーク構造に関するアブレーションスタディ

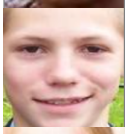
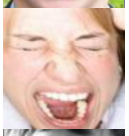
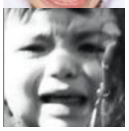
表 11 にネットワーク構造に関するアブレーションスタディの結果を示す。表の 1 列目は入力画像、2 列目は入力画像の正解ラベル、3 列目は顔の特徴点のみを入力としたネットワークの推定ラベル、4 列目は顔画像のみを入力としたネットワーク (ベースラインネットワーク) の推定ラベル、5 列目は顔の特徴点と顔画像を入力としたネットワーク (提案手法) の推定ラベルを表す。1 行目を見ると、顔画像を入力としたネットワークによる推定は失敗しているが、特徴点を入力としたネットワークによる推定は成功しており、それらをマージした結果として正解ラベルと同じラベルを推定していることがわかる。また、2 行目を見ると、全てのネットワークで正解ラベルと同じラベルを推定している。この結果より、顔画像のみでは推定できなかったが、顔の特徴点を用いることで推定に成功したデータが存在することがわかる。ここで 3 行目を見ると、顔の特徴点および顔画像のネットワークの両方で推定に失敗しているが、マージしたネットワークでは推定に成功していることがわかる。これは、顔の特徴点および顔画像のネットワークの推定尤度が最も高かったラベルはそれぞれ Neutral と Happiness であるが、2、3 番目に高かった推定尤度のラベルが Anger であり、それらをマージした結果、合算された Anger の推定尤度が Neutral や Happiness の推定尤度を上回ったからではないかと考えている。また、4 行目を見ると、マージしたネットワークでは推定に失敗しているが、特徴点を入力としたネットワークによる推定は成功していることがわかる。

6. まとめと今後の課題

本研究では、人物の顔画像を入力とした感情推定において、人物の顔画像だけでなく、別途抽出した顔の特徴点を利用することで、感情推定の精度を向上させる手法を提案した。当初の試みとして、人物の顔画像を入力とする既存のネットワーク (ベースライン) [17] に、顔の特徴点を入力とするネットワークを新たに追加し、それぞれのネットワークの最終層の 1 つ手前で得られる特徴マップをチャンネル方向に結合した後、感情カテゴリの分類を行った。しかし、ベースラインの学習済みネットワークは単独ですで

[*3] ベースラインネットワークについては ARM ブロックの手前の ResNet18 の第 4 層の出力に対して、特徴点ネットワークについては ARM ブロックの手前の 5 層目の畳み込み層の出力に対して、それぞれ Grad-CAM++を適用した。

表 11 ある入力画像に対して、顔の特徴点のみ、顔画像のみ、両方、を入力とした 3 通りのモデルの推定結果。1 列目は入力画像、2 列目は入力画像の正解ラベル、3 列目は顔の特徴点のみを入力としたネットワークの推定ラベル、4 列目は顔画像のみを入力としたネットワーク (ベースラインネットワーク) の推定ラベル、5 列目は顔の特徴点と顔画像を入力としたネットワーク (提案手法) の推定ラベルを表す。図中の人物顔画像の出典: 文献 [10]

入力画像	正解ラベル	特徴点のみ	顔画像のみ	両方
	Neutral	Neutral	Sadness	Neutral
	Happiness	Happiness	Happiness	Happiness
	Anger	Neutral	Happiness	Anger
	Sadness	Sadness	Happiness	Happiness

に一定の精度が得られている一方、顔の特徴点を入力とするネットワークの推定精度は低く、両者の出力を単純に合わせると、前者単独よりも精度が下がることがわかった。そこで、ベースラインネットワークはパラメータを固定し、顔の特徴点を入力とするネットワークの推定結果に基づいて、後者の出力する特徴マップを重み付けし、前者の特徴マップとマージした。特徴マップの重みは次の 2 つを考慮した:

ネットワーク単独での確信度: 顔の特徴点を入力とするネットワークが推定した、感情カテゴリごとの確率のうち最大のものを推定結果の確信度と見なし、その値を乗算した。

ベースラインネットワークの推定結果との一致・不一致: ベースラインネットワークの推定した感情カテゴリと顔の特徴点を入力とするネットワークの推定結果が一致しない場合、後者の信頼度が低いと見なし、前者の推定結果への影響を小さくするため、重みを小さくした。

その結果、感情推定の精度が向上し、顔画像だけでは感情推定に失敗していた画像 (表 11 の 3、4 行目のような画像) に対しても、顔の特徴点を利用することで正解することができた。

今後は、感情推定の精度をさらに向上させることが課題である。特徴マップの重み付けによって、顔の特徴点によるネットワークの影響が全体の中で弱まり、顔の特徴点のネットワークでは推定できているにも関わらず、全体のネットワークでは推定できていないというデータが存在する。このようなデータも正解できるように、さらに重み付けを工夫することによって感情推定の精度がさらに向上するのではないかと考えている。

参考文献

[1] Mouath Aouayeb, Wassim Hamidouche, Catherine Soladié, Kidiyo Kpalma, and Renaud Séguier. Learning vision transformer with squeeze and excitation for facial expression recognition. *CoRR*, Vol. abs/2107.03107, p. 13, 2021.

[2] L. F. Barrett. *How emotions are made: The secret life of the brain*. Houghton Mifflin Harcourt, 2017.

[3] Lisa Barrett, Batja Mesquita, and Maria Gendron. Context in Emotion Perception. *Current Directions in Psychological Science*, Vol. 20, pp. 286–290, 10 2011.

[4] P. Burkert, F. Trier, M. Afzal, A. Dengel, and Marcus Liwicki. DeXpression: Deep Convolutional Neural Network for Expression Recognition. *ArXiv*, Vol. abs/1509.05371, p. 8, 2015.

[5] Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth N. Balasubramanian. Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018, Lake Tahoe, NV, USA, March 12-15, 2018*, pp. 839–847. IEEE Computer Society, 2018.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[7] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Los Alamitos, CA, USA, jun 2016. IEEE Computer Society.

[8] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR (Poster)*, 2015.

[9] Ronak Kosti, Jose M. Alvarez, Adrià Recasens, and Àgata Lapedriza. Context based emotion recognition using EMOTIC dataset. *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 42, No. 11, pp. 2755–2766, 2020.

[10] Shan Li, Weihong Deng, and JunPing Du. Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2584–2593, 2017.

[11] Andrey Malinin and Mark Gales. Predictive Uncertainty Estimation via Prior Networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, p. 7047–7058, Red Hook, NY, USA, 2018. Curran Associates Inc.

[12] T. Mittal, P. Guhan, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha. EmotiCon: Context-Aware Multimodal Emotion Recognition Using Frege's Principle. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14222–14231. IEEE Computer Society, jun 2020.

[13] Maja Pantic, Michel François Valstar, Ron Rademaker,

and Ludo Maat. Web-based database for facial expression analysis. In *Proceedings of the 2005 IEEE International Conference on Multimedia and Expo, ICME 2005, July 6-9, 2005, Amsterdam, The Netherlands*, pp. 317–321. IEEE Computer Society, 2005.

[14] Tanapol Pumlumchiak and Sirion Vittayakorn. Facial expression recognition using local gabor filters and pca plus lda. In *2017 9th International Conference on Information Technology and Electrical Engineering (ICITEE)*, pp. 1–6, 2017.

[15] Michael David Resnik. The Context Principle in Frege's Philosophy. *Philosophy and Phenomenological Research*, Vol. 27, No. 3, pp. 356–365, 1967.

[16] Andrey V. Savchenko. Facial expression and attributes recognition based on multi-task learning of lightweight neural networks. In *19th IEEE International Symposium on Intelligent Systems and Informatics, SISY 2021, Subotica, Serbia, September 16-18, 2021*, pp. 119–124. IEEE, 2021.

[17] Jiawei Shi and Songhao Zhu. Learning to Amend Facial Expression Representation via De-albino and Affinity. *CoRR*, Vol. abs/2103.10189, p. 10, 2021.

[18] Juntang Zhuang, Tommy Tang, Yifan Ding, Sekhar C Tatikonda, Nicha Dvornek, Xenophon Papademetris, and James Duncan. Adabelief optimizer: Adapting stepsizes by the belief in observed gradients. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, Vol. 33, pp. 18795–18806. Curran Associates, Inc., 2020.