Vol.2021-MPS-134 No.6
2021/7/27

Deep Convolutional Neural Network を用いた 美容カウンセリングにおける顔印象の画像分類

黒沢正治 [1,2]　庄野逸 [1]

概要：美容カウンセリングにおいて美容専門家による感性や経験に基づく印象評価は重要な役割を果たす．本研究では顔画像から美容専門家の印象評価を予測するモデルの構築を目的として，深層畳み込みニューラルネットワーク（DCNN）モデルを適用した．また，少数データセットという課題の解決と顔印象の表現に適した顔特徴量抽出を目指し，大規模顔画像データセットを用いた事前学習による転移学習手法の適用を試みた．その結果，さらなる精度の改善の余地はあるが，いわば専門家の暗黙知と言える顔画像からの印象評価を DCNN モデルで表現できる可能性を見出した．

キーワード：顔印象分類，美容カウンセリング，専門評価者のスキル表現，深層畳み込みニューラルネットワーク

Face Impression Classification in Cosmetic Counseling Using Deep Convolutional Neural Network

MASAHARU KUROSAWA[1,2]　HAYARU SHOUNO[1]

Abstract: In cosmetic counseling, the evaluation of facial impression by beauty experts plays an important role. In this study, we used a deep convolutional neural network (DCNN) model to predict the impression evaluated by beauty expert from facial images. We also used a transfer learning method with a large-scale face image dataset to achieve effective features of beauty experts' evaluation with a small amount of data. As a result, although the model's accuracy needs to be improved for practical use, we conclude that the DCNN model can represent the unique evaluation of face impressions in cosmetic counseling, as provided by beauty experts.

Keywords: Face impression, cosmetic counseling, expert skill representation, deep convolutional neural network

1. 背景と目的

スキンケアやメイクの目的の一つは，自分のなりたい印象を形成することである[1]．印象に影響を与える顔の特徴は個人間で大きく異なるため[2]，経験豊富な美容専門家による対面での美容カウンセリングが顧客の商品購買やサービス選択において重要な役割を果たす．一方で，対面でカウンセリングを受ける機会は場所的および時間的に制約があり，美容専門家の育成にもコストがかかる．そこで熟練の美容専門家の感性と経験に依存していた顔の印象評価技術のモデリングが求められている．

顔の印象評価やその予測モデリングに関する研究は古くからなされている[3, 4]．モーフィング画像や平均画像を刺激画像として用いた印象評価では対象印象への顔特徴の関与がわかりやすい一方で，提示画像と実際の人の顔とに乖離があるなどの課題がある．また，近年は顔認識や印象分類，感情識別などの顔関連タスクのための様々な大規模画像データセットがオープンソースとして活用できるが，これらに付与されている評価ラベルの多くは非専門家によるもので，美容カウンセリングシーンへ直接活用することはその精度から不適であった．

そこで，本研究では顧客の顔画像から美容専門家の印象評価を予測するモデルの構築を目的とした．まず新たに日本人女性の顔画像を取得し，個々の顔画像に美容の専門家が Skin-Power（SP） と呼称する感性と経験に基づく印象評価値ラベルを付与したデータセット（SP dataset）を構築した．SP スコア予測モデルの作成においては，二つの課題があった．一つ目は SP スコアがいわば暗黙知であり，明確な画像特徴の言語化や数値化が非常に難しいという課題である．これに対しては画像の特徴量抽出から分類器までのパラメータの学習を end-to-end で行う深層畳み込みニューラルネットワーク（Deep Convolutional Neural Network : DCNN）モデルの適用により解決を試みた．DCNN モデルは，福島らによって提案されたネオコグニトロンと呼ばれる視覚認知における階層型の神経回路モデルを起源にもち[5]，畳み込み層とプーリング層を多層に組み合わせた構造を有するモデルである．二つ目の課題は DCNN モデルが膨大な数のパラメータを持つため一般的に SP dataset のような少数データセットでは十分にパラメータの学習が進まないという点である．この課題に対して，任意の大規

1 電気通信大学大学院
Graduate School of Informatics and Engineering, University of Electro-Communications.
2 株式会社コーセー
KOSÉ Corporation.

模画像データセットを用いて事前学習させた DCNN パラメータを初期値として，目的のデータセットとタスクで再学習させる転移学習と呼ばれる手法を用いた．また，画像分類における転移学習には ImageNet と呼ばれる大規模自然画像データセット[6]を用いることが一般的であるが，ImageNet には顔に関するクラスは存在しない．また，先行研究にて画像ドメインが特殊且つ少数データセットである肺疾患 CT 画像の分類において，ImageNet の事前学習に加えてテクスチャ画像を用いて 2 段階の転移学習をすることで分類精度が向上した報告がある[7, 8]．そこで，より SP スコア分類に対して有効な特徴抽出と予測精度向上を可能とするために，大規模顔画像データセット（FFHQ-aging）を用いた事前学習を試みた．

2. 実験方法

2.1 使用データセットと美容専門家による SP スコア付け

10 歳代から 70 歳代の日本人女性 377 名の正面・真顔の画像を取得し，美容専門家により SP スコアを 1 から 7 の 7 段階で付与した（1 が最良）．SP スコアは「活き活きとした，エネルギッシュな，健康的な」などの印象語で近く形容される評価指標であり[9]，本研究ではこの複合的な印象値を Skin-Power（SP）スコアと呼称する．DCNN モデルへの入力時は 256×256 サイズにリサイズ後，224×224 サイズで任意のクロップとフリップ処理を加えて画像の水増しを行った．前処理を行う前の各クラスの画像枚数と画像例を表 1 および図 1 に示す．

表 1　データセットの各クラスの画像枚数

SP Class	Images
SP_1	23
SP_2	66
SP_3	63
SP_4	44
SP_5	67
SP_6	84
SP_7	30
Total	377

図 1　SP 画像例（入力時は目元隠蔽は無い状態で使用）

2.2 深層ニューラルネットワークモデル

これまで SP dataset と同質・同規模の画像データセット及びタスクに DCNN モデルを適用した先行知見はほとんどないため，本研究ではまず基本的なアーキテクチャである VGG16 を用いた．VGG16 は 2014 年に提案されたモデルで，ネオコグニトロンの自然派生形であり，図 2 に示すように畳み込み層とプーリング層のみからなるシンプルな構造の DCNN モデルである[10].

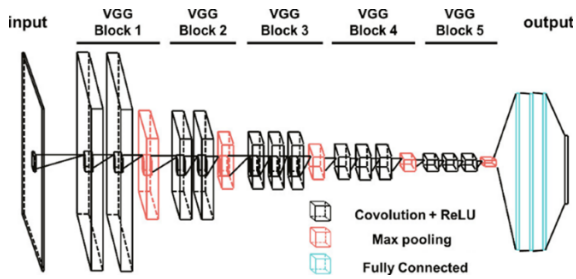

図 2　VGG16 アーキテクチャの模式図（[11]から引用編集）

2.3 モデルの訓練と検証条件

本研究のタスクは前処理をした RGB 画像を入力とし，SP スコアの 7 クラス分類タスクとした．データは訓練用と検証用に 8:2 の割合で分割し使用した．損失関数は交差エントロピー誤差を用い，最適化手法は Stochastic Gradient Descent （SGD）を用いた．学習率は 0.01, 0.001, 0.0001 を検討し，後述する結果では 0.001 の結果を示す．バッチサイズは 16，エポック数は 60 とした．検証データにおける正解率が最も高いエポックのモデルを最高精度モデルとして選択した．尚，検証用データにより算出される損失値は学習（パラメータ更新）には用いていない．また，各精度指標の比較では 5 分割交差検証法を行い，5 セットにおける各最高精度モデルの精度の平均値を代表値として示した．

2.4 ImageNet および FFHQ-Aging を用いた転移学習

DCNN の事前学習には ImageNet と FFHQ-aging を用いた．ImageNet は 1000 種類の物体に対して平均 1000 枚程度の画像が含まれる自然画像データセットである[6]．FFHQ-aging は高画質の顔画像にクラウドソーシングで見た目の年齢を付与した 7000 枚からなる顔画像データセットである[12]．VGG16 を ImageNet で事前学習した後に畳み込み層のパラメータを固定し全結合層のみを SP dataset で再学習させたモデルを VGG16_fixed，畳み込み層および全結合層を SP dataset でファインチューニングしたモデルを VGG16_ft と記す．また，VGG16 を FFHQ-aging を用いて年代の 10 クラス分類タスクをスクラッチで学習，または ImageNet 事前学習モデルを FFHQ-aging でファインチューニングした後に SP dataset で再度ファインチューニングしたモデルをそれぞれ，VGG16_ft-on-scr, VGG16_ft-on-ft と記す．

FFHQ-aging を用いたモデルの学習では訓練データに 63000 枚，検証用に 7000 枚を用いて SP dataset を用いた学習と同条件で訓練した．

2.5 SP 評価におけるモデルの注視領域の可視化

SP 分類タスクにおいて各クラス出力に対する寄与度の高い画像エリアを可視化するために，勾配情報を用いた gradient-weighted class activation mapping （Grad-CAM）[13] という手法を適用した．各 SP クラスの予測で正解した画像のみを用いて SP クラス毎にヒートマップの平均画像を作成した．尚，正解データ数で除して平均化した後，正規化はしない状態で可視化した．結果の画像としては各クラスの入力画像（前処理した顔画像）の平均画像にヒートマップを合成した画像を示した．

3. 結果

3.1 VGG16 を用いた SP スコアの分類

VGG16 モデルを複数の 4 つの学習方法で訓練したモデルの分類精度と混同行列をそれぞれ表 2 と図 3 に示す．VGG16_fixed と比較して VGG16_ft では正解率 59.4%，F1-スコア 0.516 と共に精度向上が見られた．また，混同行列をみると VGG16_ft においては SP スコア 4 の予測精度が向上し，±1 クラス正解率も同様に向上していることから SP スコアの連続性を捉えていることが示唆された．さらに，興味深いことに VGG16_ft-on-ft において，正解率は 62.1%，VGG16_fixed と比較して 4.3%，VGG16_ft と比較して 2.7% の増加がみられた．F1 score は 0.533 となり，VGG16_fixed と比較して 0.053，VGG16_ft と比較して 0.017 の増加がみられた．一方で，VGG16_ft-on-scr においては VGG16_ft-on-ft のような精度改善は見られなかった．VGG16_ft-on-ft における精度改善については，混同行列では VGG16_ft と同様の傾向で SP スコア 4 の予測精度が向上した．また，±1 クラス正解率も同様に向上しており，SP スコアの連続性をより捉えていることが示唆された．

表 2 SP スコア分類の精度比較

モデル	正解率 (%)	±1 クラス正解率 (%)	F1-スコア
VGG16_fixed	57.8	93.1	0.480
VGG16_ft	59.4	94.4	0.516
VGG16_ft-on-scr	59.9	93.6	0.459
VGG16_ft-on-ft	62.1	96.3	0.533

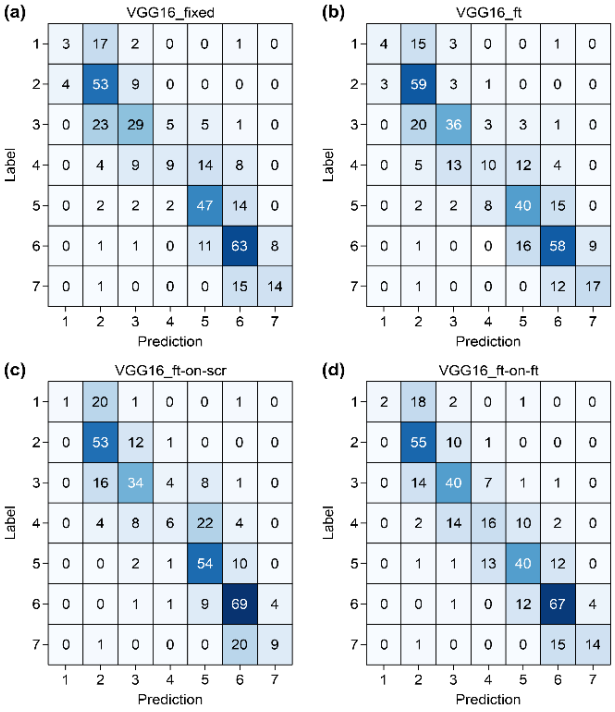

図 3 各 VGG16 モデルの混同行列(A)VGG16_fixed (B)VGG16_ft (C)VGG16_ft-on-scr (D)VGG16_ft-on-ft. 結果は 5 分割交差検証における検証結果を全て統合して表示.

3.2 VGG16 予測モデルの注視領域の可視化と比較

3.1 の結果から SP dataset で訓練した DCNN モデルでは SP スコア予測に効果的な特徴量を獲得できていることが予測される．そこで，どのような画像特徴量が SP スコア予測に寄与しているのかを知るために，Grad-CAM を用いた．Grad-CAM は DCNN の説明性や解釈性の研究領域において提案された手法で，最終畳み込み層特徴マップの予測クラスの出力に対する勾配を用いて特徴マップをヒートマップ化する手法であり[13]，勾配が大きい特徴量は予測したクラスへの寄与が大きい特徴量であると解釈できる．訓練後の各モデルで検証データを用いて Grad-CAM 画像を作成し，図 4 に示す．VGG16_fixed では SP スコア 1 から 3 にて顔の中心部，SP スコア 5 から 7 で頭・額部分と顎・首部分など比較的顔以外の部分がハイライトされた．一方で，VGG16_fr-on-ft では目元や鼻，口，額など顔の各パーツがハイライトされ，さらに各クラスによって寄与度の高いエリアが異なることが明らかとなった．VGG16_ft や VGG16_ft-on-scr でも同様に目元や口元の寄与度が高い傾向にあるが，勾配の大きい領域（赤色で示される領域）が比較的少ないことから同クラス内での入力画像間で注視領域がばらついていることがわかった．

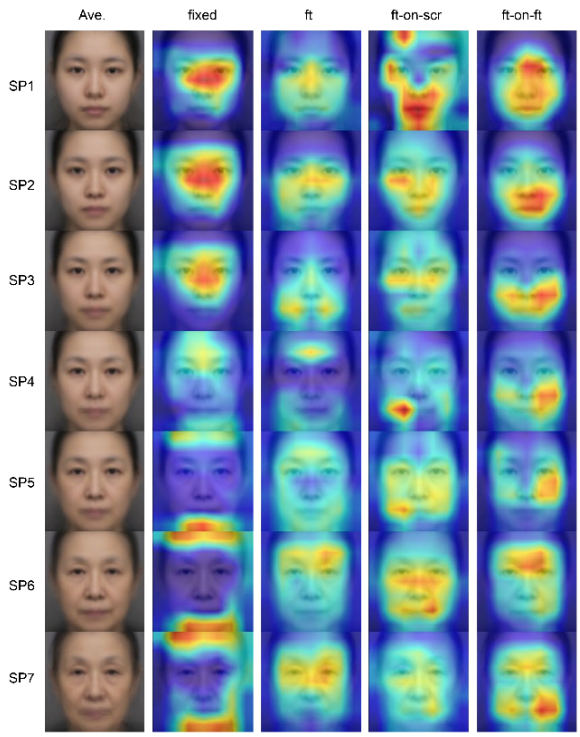

図 4　各モデルの SP スコア推論時の Grad-CAM 画像. 各クラスにおける平均画像を表示.

4. 議論

本研究では，美容専門家の顔印象評価（SP スコア）を予測する DCNN モデルを検討した. 分類精度は 57.8%～62.1%の精度，±1 クラス正解率においては 93%を超える精度を得ることができ，少数データセットかつ難しいタスクにおいて一定の精度が得られたと考える．また，VGG16_fixed 精度が最も低く，Grad-CAM 画像では顔の一部のみや末端に注視領域があることから ImageNet のみの学習では SP スコア表現のための十分な特徴量を得られていないことが示唆された．本研究では ImageNet 事前学習では「顔特徴の理解」が不十分という事前の仮説から，大規模顔画像データセットの FFHQ-aging を用いた事前学習を検討した．その結果，ImageNet 事前学習モデルにさらに FFHQ-aging を学習させた二段階転移学習により，検討したモデルの中では最も高い精度を得た．このことから SP スコア分類には顔画像データから得られる特徴のみならず，多様な画像分類に必要な特徴量獲得が寄与することが示された．これは先行研究[7, 8]と同様の結果となった.

5. 結論と今後の研究展望

本研究結果は分類精度のみからでは実用上まだ検討の余地が残されているが，これまで美容専門家の感性や経験による評価技術を DCNN で表現したことはなく，新しい知見の獲得に繋がった．本研究では FFHQ-aging と年代分類タスクによる事前学習のみを検討したが，他の顔画像データセットやタスクの違いによる転移学習効果について今後検討する予定である．また，美容専門家との議論において SP スコアに寄与する可能性のある部分として目元と口元と頬が挙げられていた. 今回用いた DCNN のモデル構造や学習手法に専門家の知見を明示的に組み込むことはしていないが，Grad-CAM で可視化した注視領域と専門家が指摘する顔部位情報との関連性については，DCNN モデルが人の視覚認知や神経回路との関連がある背景を考えると大変興味深い．また，顔画像から人種を予測するタスクにおいて DCNN の注視領域と実際の人の視線測定データとの関連や, 人の視線情報を DCNN モデルの注視情報として与えることで画像キャプショニングタスクでの精度向上に寄与する報告もあり[14, 15], 本研究における美容専門家の視線情報と DCNN モデル注視領域の関連性評価やモデルへの組み込みによる精度向上の可能性も期待され，今後検討予定である.

参考文献

[1] Workman JE, and Johnson KKP. The Role of Cosmetics in Impression Formation. Cloth Text Res J. 1991;10: (1).
[2] Jaeger B, Wagemans FMA, Evans AM, and Van Beest I. Effects of Facial Skin Smoothness and Blemishes on Trait Impressions. Percept. 2018;47(6):608-625.
[3] Kortli Y, Jridi M, Al Falou A, and Atri M. Face Recognition Systems: A Survey. Sens. 2020;20(2):342.
[4] Wang M, and Deng W. Deep face recognition: A survey. Neurocomputing. 2021; 429:215-244.
[5] Fukushima K. Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position. Biol Cybern. 1980; 36:193-202.
[6] Deng J, Li K, Do M, Su H, and Fei-Fei L. Construction and Analysis of a Large Scale Image Ontology. Vis Sci Soc. 2009.
[7] Shouno H, Suzuki A, Suzuki S, and Kido S. Deep Convolution Neural Network with 2-Stage Transfer Learning for Medical Image Classification. Brain Neural Netw. 2017;24(1): 3-12.
[8] Shouno H, Suzuki S, and Kido S. A transfer learning method with deep convolutional neural network for diffuse lung disease classification. In: Arik S, Huang T, Lai W, Liu Q, editors. Neural Information Processing. ICONIP 2015. Lecture Notes in Computer Science, vol 9489. Springer, Cham. https://doi.org/10.1007/978-3-319-26532-2_22.
[9] Jones AL. The influence of shape and colour cue classes on facial health perception. Evol Hum Behav;2018. ;39(1): 19-29.
[10] Simonyan K, and Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv pre-print server. 2015.
[11] Hamano Y, and Shouno H. Analysis of Texture Representation in Convolution Neural Network Using Wavelet Based Joint Statistics. Springer International Publishing; 2020. p. 126-136.
[12] Or-El R, Sengupta S, Fried O, Shechtman E, Kemelmacher-Shlizerman I. Lifespan Age Transformation Synthesis. In: Vedaldi A, Bischof H, Brox T, Frahm JM, editors. In Eur Conf Comput Vis – ECCV 2020. Lecture Notes in Computer Science. 2020, vol 12351. Springer:Cham
[13] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, and Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. Int J Comput Vis. 2020;128(2): 336-359.
[14] Jack RE, Blais C, Scheepers C, Schyns PG, and Caldara R. Cultural confusions show that facial expressions are not universal. Current Biol. 2009; 19:1543-1548.
[15] Sugano Y, and Bulling A. Seeing with Humans: Gaze-Assisted Neural Image Captioning. arXiv pre-print server. 2016.