スタイル特徴を利用したDNNによる印象推定に寄与する画像領域の可視化

○大萩 優哉† 飛谷 謙介† 谷 伊織‡ 橋本 翔† 長田 典子†
†: 関西学院大学理工学部人間システム工学科／感性価値創造研究センター
Yuya-Ohagi@kwansei.ac.jp

概要： 本研究では，画像の印象評価において生じる「個人のばらつき」を包含した高度な印象推定モデルを構築することを目的とする．まず，画像と印象との関係性をモデル化するため，画像から抽出したスタイル特徴と，離散確率分布で表現された印象値を学習し，印象推定を行うCNNを構築する．その後，学習モデルへの寄与の度合いを可視化する手法（Grad-CAM）を利用して，構築したCNNにおいて印象に寄与する画像領域を可視化した．その結果，評価傾向の異なるサンプル間では，印象に寄与する画像領域が異なることが確認された．これにより印象評価における個人のばらつきを空間的に説明できる可能性を示した．
＜キーワード＞ DNN，スタイル特徴，印象推定，可視化，感性情報，主観評価

1. 背景・目的

近年，深層学習によりAIの可能性が急速に広がり，社会の注目を集めるほど高性能なものとなっている．これに伴って，取り扱うデータも多様になってきており，AVAデータセットのような主観評価が付与された画像データセットも整備されてきた[1]．こうした主観的データでは個人によって評価がばらつくといった感性情報固有の特徴[2]を考慮することが必要である．

一方で，AIの解釈性，すなわち深層学習の判断した根拠を明らかにする研究が盛んに行われてきている．特に，CNNにおいては学習した概念を画像として可視化する技術[3]が提案されている．

そこで本研究では，画像の印象評価において生じる「個人のばらつき」を包含した高度な印象推定モデルを構築することを目的とし，構築したモデルにおいて印象に寄与する画像領域の可視化を行うことでその妥当性を確認する．

2. 先行研究

印象推定の研究において，Sunda et al.はスタイル特徴を用いて衣服の柄に対する印象と物理特性との関係性をモデル化し，高精度な印象推定モデルを構築した．スタイル特徴はGatys et al.がスタイル転写[4]の際に定めたもので，CNNの特徴マップをグラム行列化することで得られる．

一方で，CNNの可視化の研究において，Selvaraju et al.はモデルが学習した概念を可視化するGrad-CAMという手法を提案した．Grad-CAMはCNNにおいて可視化したい畳み込み層の各領域が出力に寄与する度合いを疑似的にヒートマップとして表現する手法である．

本研究では，上記2つの手法を用い，スタイル特徴を学習し印象推定を行うCNNを構築し，構築したCNNに対し，Grad-CAMにより評価傾向の違うサンプル間での印象に寄与する画像領域の違いを可視化し，個人のばらつきによる差があることを確認する．

3. スタイル特徴を学習するCNNの構築

3.1. スタイル特徴の抽出

スタイル特徴は，視覚的印象との関わりが強いとされており[5]，これを利用することで高精度な印象推定が実現できると考えられる．

本研究では，入力データである画像から学習済みVGG19[6]を通して抽出したスタイル特徴を利用する．Sunda et al.は各Pooling層から抽出されたスタイル特徴を使用している．その際，深いPooling層から抽出されたスタイル特徴ほど高次元である．そのため，データセットのサンプル数が少ない場合，学習の際に過学習することが予想される．そこで本研究では，最初のpooling層から抽出される低次元のスタイル特徴を用いる．

3.2. 印象推定を行うCNNの構築

本研究で扱うCNNは，画像を入力データとし，スタイル特徴を抽出，全結合層を挟んで出力層でクラス分類を行う．その際，クラス分類は印象の強弱と対応し，クラス番号と印象値の段階が一致している．構築したCNNの構造および詳細をそれぞれ図1，表1に

示す. 表1において, Ave pool の層でグラム行列の抽出を行う. そのうえで, 10-fold 交差検証による 10 個の学習済みの CNN のうち推定誤差が最小のモデルを 4 章で述べる画像領域の可視化に用いる.

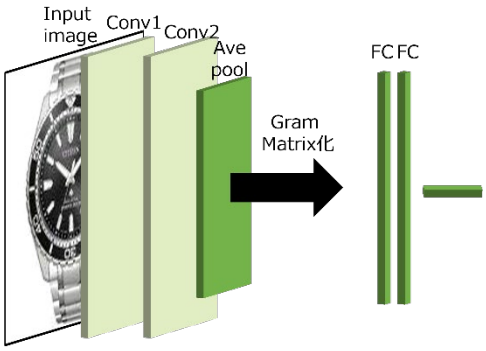

図 1 CNN の構造

表 1 CNN の構造

Layer	Input size	Output size	Kernel	Stride
Conv1	224 × 224 × 3	224 × 224 × 64	3×3	1
Conv2	224 × 224 × 64	224 × 224 × 64	3×3	1
Ave pool	224 × 224 × 64	112 × 112 × 64	2×2	2
FC1	1 × 1 × 4096	1 × 1 × 4096		
FC2	1 × 1 × 4096	1 × 1 × 4096		
FC3	1 × 1 × 4096	1 × 1 × 7		

CNN の学習における最適化手法として adam[7]を使用し, バッチサイズは 128 とする. 全結合層での重みの初期値には He の初期値[8]を適用し, 学習率の初期値は 0.000001 とする. 活性化関数には Rectified Linear Unit (ReLU) 関数[9]を使用し, 全結合層では Batch Normalization[10]を適用する. また, 出力層では softmax 関数を使用する. softmax 関数を式(1)に示す.

$$p_i = \frac{\exp(y_i)}{\Sigma_{k=1}^{\mathrm{n}} \exp(y_k)} \qquad i = 1, \ldots, \mathrm{n}, \quad \mathrm{n} = \mathrm{Z} \qquad (1)$$

ここで, y_i は, クラス番号 i に対応する前層の各ユニットとその結合重みを乗じた値であり, y_k は, 前層の値に対して, 結合重みを乗じて総和した値である. この活性化関数を出力層で使用することによって, クラス番号 i に対する尤度 p_i を算出する. また, Z はラベル数である. この算出した尤度 p_i と画像の教師ラベルを用いて損失関数より誤差 E を算出し, これを最小化するように重みを更新することによってネットワークの重みを学習する. さらに, 印象推定の精度指標として MSE (Mean Squared Error)を用いる. 使用する MSE を式(2)に示す. これより誤差 L を算出する.

$$L = \frac{1}{\mathrm{N}}\left[\sum_{i=1}^{\mathrm{z}} (\hat{y}_i - \mathrm{y}_i)^2\right] \qquad (2)$$

ここで N はテストデータの総数であり, $\hat{y}_i$ および y_i はそれぞれ推定された分布, 印象値の分布のクラス番号 i に対応した値である.

3.3. 損失関数の選定

深層学習においては, パラメータを最適化する際のアルゴリズムに加えて, 損失関数の選定も推定精度に大きく関わってくる. 本研究では印象値を離散確率分布として扱うため, 当該形式に適した損失関数を選ぶ必要がある.

そこで, クラス分類問題でよく用いられるクロスエントロピー誤差(CE), および Jin らにより提案された, 確率分布の差を表す Jensen-Shannon 情報量に基づく損失関数(CJS)[11]とで精度比較を行う. 誤差の算出方法は印象値の分布と推定された分布との MSE とする.

比較のために用いる DNN の詳細を表 2 に示す. 構築した DNN はスタイル特徴を入力データとし, 隠れ層は全結合層, 出力層はクラス分類とする. また, DNN の学習方法は 3.2.で述べたものと同様とする.

表 2 DNN の構造

Layer	Input size	Output size
FC1	1 × 1 × 4096	1 × 1 × 4096
FC2	1 × 1 × 4096	1 × 1 × 4096
FC3	1 × 1 × 4096	1 × 1 × 7

DNN の学習で得られた推定誤差を表 3 に示す. クロスエントロピー誤差を用いたモデルの推定誤差の方が低いことがわかる. したがって本研究では, 損失関数としてクロスエントロピー誤差を用いる.

表 3 推定誤差

	CE	CJS
MSE	0.117	0.136

3.4. 対象データセット

本研究では, 定量化された印象 (印象値) が付与された腕時計の画像データセット[12]を用いる. その理由としては, 腕時計はデザインを構成する要

素が少ないため，デザイン要素と印象との関係が解釈しやすく，結果の有用性・妥当性を示しやすいためである．

また，本研究では視覚的な印象を扱うため，データセットに付与されている印象のうち,「上品だ, 可愛らしい，小さい」を対象とした[12].

本データセットは, 画像 1 枚あたり 20 人分の印象値が付与された計 2000 枚の画像で構成されている．全画像において背景が白色であり，腕時計が中央に位置している．また, 印象値は, 主観評価実験を通して 1~7 の 7 段階で付与されている．実験は，クラウドソーシングサービスである「クラウドワークス」を利用して実施された. 実験では, 各実験参加者は画像を 1 枚ずつ観察し，評価項目に対して「非常に当てはまらない」,「当てはまらない」,「やや当てはまらない」,「どちらでもない」,「やや当てはまる」,「当てはまる」,「非常に当てはまる」の 7 段階で評価した. その際, 画像の呈示順序は実験参加者ごとにランダムとした．本研究では，20 人分の評価結果から，7 段階の印象値を 1 から 7 の値に変換し，縦軸と横軸にそれぞれ人数（0~1 に正規化）と印象値（1~7）をとった離散確率分布の形式で扱う．

3.5. 印象推定結果

提案手法における印象推定精度と既存手法との比較を行った．比較には，10-fold 交差検証における推定誤差の平均を用いる．比較対象は VGG19（ImageNet[13]にて学習済み）の fine tuning（VGG19）により全結合層の重みを再学習したものとした．また，fine tuning における学習方法は 3.2.で述べたものと同様とする．比較の結果，表 4 のように本手法の推定誤差が低いことが確認された．これにより，本手法が画像の印象を推定するうえで有効であることが示唆された.

表 4 CNN による推定誤差

	本手法	VGG19
MSE	0.094	0.192

4. 印象に寄与する画像領域の可視化

4.1. 可視化フロー

本研究では，印象に寄与する画像領域の可視化を行うために 2 つの処理を行う．まず，(i) 各印象値に寄与する画像領域の可視化を行う．その後，(ii) 寄与度の高い画像領域の統合を行う．

(i) ではクラス分類における各クラス（印象の強弱）に寄与する画像領域を Grad-CAM[3]により可視化を行い，ラベル数分のヒートマップを取得する．(ii) では (i) で得られた全てのヒートマップを用いて画素毎に比較を行い，最も寄与度の高いクラスに対応する色（既定）をその画素に付加する．以上の処理を行い得られた画像を「ヒートマップ統合画像」とする. また, 一定の閾値以上の寄与度を持つ画素以外は黒色とする.

4.2. 可視化結果

選定された CNN モデルに対し，(i) の処理を適用し，7 クラス分類であることから 7 枚のヒートマップが得られた. その際, 可視化する層は図 1 中の Conv2 とした．得られた 7 枚のヒートマップに対して (ii) の処理を適用し，ヒートマップ統合画像を取得した.

次に，得られた結果の傾向を把握するため，使用したデータセットにおける代表的なサンプルを選定した．その際，印象値の分布（gt）に対して Affinity Propagation によるクラスタリングを行うことで選定した．クラスタリングの結果，本データセットにおける評価傾向は 16 クラスタに分類でき，クラスタ内での代表的なサンプルを選定した．図 2 (a)は，比較的高い印象値，(b) は低い印象値を付与されたクラスタにおける代表サンプルで，それぞれ，印象値の分布（gt），推定された分布（pred），入力画像，および Grad-CAM によって得られたヒートマップを統合した画像を列挙したものである．ヒートマップ統合画像の下にクラス番号に対応した色を示す.

5. 考察

図 2 から，腕時計の形状を保持した状態で可視化されていることが分かる．これは入力層に近い畳み込み層を用いたためと考えられる.

また 2 つのヒートマップ統合画像を比較すると，(a) では，高評価（印象値 5～7）を示す箇所は腕時計の輪郭付近であり，低評価（同 1～3）を示す箇所が腕時計のバンド付近であることがわかる．このことは，高評価の推定には輪郭が強く効いており，低評価の推定にはバンドが効いていると言える．すなわち，全体的な評価が高い腕時計画像では, 高い評価を付ける人は輪郭や光沢を重視し, 低い評価を付ける人はバンドを重視している可能性が考えられる．同様に (b) では，高評価（同 5～7）を示す箇所は腕時計の輪郭付近であり，低評価（同 1～3）を示す箇所は文字盤の文字付近であ

ることがわかる．このことから，全体的な評価が低い腕時計画像では，高い評価を付ける人は輪郭を重視しており，低い評価を付ける人は文字盤の文字を重視している可能性が考えられる．

以上のように，Grad-CAM を用いることで，評価のばらつきが生じる理由について，重視している箇所や寄与度の違いによって説明できる可能性を示すことができた．

(a)高い印象値を付与された代表サンプル

(b)低い印象値を付与された代表サンプル

図 2　Grad-CAM による印象に寄与する画像領域

6.　まとめと今後の課題

本研究では，印象評価における個人のばらつきを対象とし，CNN が学習した概念を可視化する技術と組み合わせることで，個人のばらつきを包含したうえで印象推定モデル（CNN）を構築し，印象推定に寄与する画像領域の可視化を行い，CNN が捉えた個人のばらつきによる差があることを確認した．まずスタイル特徴を利用して，高精度に推定可能なモデルを作成する CNN を構築した．その上で，印象値が付与されたデータセットを用いて学習を行い，印象推定に寄与する画像領域の可視化を行った．その結果，評価毎に重視している箇所や寄与度が異なることを確認した．

今後の課題としては，人が評価の際に重視している画像領域の確認，CNN の推定精度向上に加え，対象データによる結果の差異にも着目した研究を行う．

参考文献

[1] N. Murray et al.: AVA: A large-scale database for aesthetic visual analysis, 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp.2408-2415, 2012.

[2] 井口征士他: 感性情報処理, 電子情報通信学会編ヒューマンコミュニケーション工学シリーズ, オーム社, 1994.

[3] R. R. Selvaraju et al.: Grad-cam: Visual explanations from deep networks via gradient-based localization, Proceedings of the IEEE International Conference on Computer Vision, pp.618-626, 2017.

[4] L. A. Gatys et al.: Image style transfer using convolutional neural networks, Proceedings of the IEEE conference on computer vision and pattern recognition, pp.2414-2423, 2016.

[5] N. Sunda et al.: Impression estimation model and pattern search system based on style features and Kansei metric, Proceedings of the 24th ACM Symposium on Virtual Reality Software and Technology, p.98, 2018.

[6] K. Simonyan et al.: Very Deep Convolutional Networks for Large-scale Image Recognition, arXiv preprint arXiv:1409.1556, 2014.

[7] J. Kingma et al.: Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980, 2014.

[8] K. He et al.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, Proceedings of the IEEE international conference on computer vision, pp.1026-1034, 2015.

[9] X. Glorot et al.: Deep sparse rectifier neural networks, Proceedings of the fourteenth international conference on artificial intelligence and statistics, pp.315-323, 2011.

[10] S. Ioffe et al: Batch normalization: Accelerating deep network training by reducing internal covariate shift, arXiv preprint arXiv:1502.03167, 2015.

[11] X. Jin et al.: Predicting aesthetic score distribution through cumulative jensen-shannon divergence, Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[12] 鈴木秀通他: レビューテキストと画像を用いた機械学習によるプロダクトの感性指標構築, 精密工学会誌 in press.

[13] J. Deng et al.: Imagenet: A large-scale hierarchical image database, 2009 IEEE conference on computer vision and pattern recognition, pp.248-255, 2009.

大萩優哉:現在関西学院大学理工学部人間システム工学科在学中. 感性情報学および質感に関する研究に従事.

飛谷謙介:2002 年早稲田大学理工学部応用物理学科卒業. 2004 年岐阜県立情報科学芸術大学院大学(IAMAS) 修士課程修了. JST 地域結集型共同研究事業特別研究員を経て, 2010 年岐阜大学大学院工学研究科博士後期課程修了. 同

年岐阜大学産官学融合本部研究員. 2014 年より関西学院大学理工学部／感性価値創造研究センター特任講師. 博士(工学). 主に感性工学, コンピュータビジョンに関する研究に従事. 電気学会, 精密工学会, 日本顔学会, ACM など各会員.

谷伊織:2014 年神戸大学理学研究科地球惑星科学専攻博士課程後期課程修了. 早稲田大学客員次席研究員, 総合研究大学院大学学術情報基盤センター助教を経て, 2018 年より関西学院大学理工学部感性価値創造研究センター研究特別任期制助教. 博士(理学). 人工知能による感性情報処理に関する研究に従事. 自然計算手法を用いた感性情報処理に興味を持つ. 計測自動制御学会, 共創学会正会員.

橋本翔:関西学院大学理工学部感性価値創造研究センター研究特別任期制助教. 博士(人間科学). 専門は心理統計学および多変量解析論. 現在は感性工学における感性の指標化研究に従事.

長田典子:1983 年京都大学理学部数学系卒業. 同年三菱電機(株)入社. 1996 年大阪大学大学院基礎工学研究科博士後期課程修了. 2003 年より関西学院大学理工学部情報科学科助教授, 2007 年教授. 2009 年米国パデュー大学客員研究員. 2013 年感性価値創造研究センター長. 2015 年革新的イノベーション創出プログラム「感性とデジタル製造を直結し, 生活者の創造性を拡張するファブ地球社会創造拠点」サテライトリーダー. 博士(工学). 専門は感性工学, メディア工学等.