CF-010

CNN を用いた服飾・風景画像に対する印象の推定
Estimation of Impressions for Clothing and Landscape Images Using CNN

神戸 瑞樹[1)]　横山 想一郎[2)]　山下 倫央[2)]　川村 秀憲[2)]
Mizuki Kambe　Soichiro Yokoyama　Tomohisa Yamashita　Hidenori Kawamura

1　まえがき

ファッション業界では，「ガーリー」「上品」などの印象や各ブランドのイメージをもとに商品の開発・販売が行われている．こうした印象やイメージは明確な定義がなく，受け手により変わり得るものの，ファッション業界では一定の共通認識が存在する．例えば，ファッション雑誌 [1] では「華やか見え！大人ガーリー」や「きちんと見え！上品ベーシック」などと商品が説明されている．しかし，定量的な分析が出来ないため，デザイナーなどの一部の人によってどのような商品を作るかが決定されている．このため，人の印象を定量的に評価できるシステムが求められる．

本研究では，印象を定量的に評価するということを 100 人中何人がある印象を感じるかと定義する．これを CNN を用いてタグ付けされた画像から推測するために服飾画像に対して印象語を付与したデータセットと，風景や屋内の画像に対応するブランドイメージを持つブランドを付与したデータセットを用いる．

2　関連研究

Fashion Dataset AI の研究においてファッションはメジャーな分野の一つであり，ファッション用のデータセットが数多く作られている．特定のファッションカテゴリのみを含むもの [2] や多様なファッションアイテムを含むものが存在する [3]．より細かい分析を行うために属性やランドマークといった補足情報を含んでいるデータセットもあるが，属性は視覚的なものが主であり，印象的な属性や感覚的な属性はほぼない．これは，データセットのほとんどが同一商品の画像検索を行うために設計されているためである．最大規模のデータセットとして DeepFashion があり，80 万枚の画像から構成されている [4]．また，人の全身画像から個別のファッションアイテムを解析することを目的としたデータセットも存在する [5, 6]．これらのデータセットはピクセルレベルでアノテーションを有している．

Attribute Learning 属性を学習する研究はファッションに限らず行われている [7, 8]．ファッションにおいては学習した属性を用いて，同一商品の画像検索 [4] やその属性を持っている服飾の検索 [9] などが行われている．また，WEB 上のノイズ混じりのデータから学習するものも存在する [10].

これらの研究で扱う属性は基本的に視覚的なものであるが，印象的な属性を学習するものも存在する．Xiong らはサポートベクター回帰を用いて携帯電話の形状から印象を学習させた [11]．Vaccaro らは服飾画像に与えられたテキストから属性を抽出し，多言語トピックモデルを用いて視覚的な属性と印象的な属性の関係性を学習した [12]．Talebi らは CNN を用いて画像の「きれいさ」について学習させた [13].

しかし，これらの印象的な属性を学習する研究では属性数・データ数といった面で規模が小さい．本研究では，属性数・画像枚数を大幅に増やしたデータセットを作成し，それを学習する．

3　データセット

服飾画像に印象語が付与された Fashion Impression Dataset と風景や屋内の画像に対して対応するブランドイメージを持つブランドが付与された Brand Image Dataset を用いる．

3.1　Fashion Impression Dataset

服飾画像に対して「カットソー」「スカート」といったカテゴリと「ガーリー」「甘い」といった印象語，「無地」「花柄」といった外観的な要素が付与されている．印象語，外観的な要素は 4 段階評価 (1. とても当てはまる 2. やや当てはまる 3. あまり当てはまらない 4. 全く当てはまらない) のものと，バイナリ評価 (そのキーワードが当てはまるか否か) のものがある．4 段階評価のものが 6 個，バイナリ評価のものが 142 個の計 148 個がある．「第一印象」「素材感」「服のイメージ」「色彩」「シルエット (フォルム)」「柄・模様」「特徴・効果 1」「特徴・効果 2」「用途・シーン」といった観点から選出され，ファッション雑誌でよく使われる単語や，デザイナーの意見を参考にして厳選したものが利用されている．4 段階評価には，通常であれば「どちらでもない」を加え，5 段階，7 段階といった奇数での評価段階を設定するが，今回は『とりあえず「どちらでもない」を選択する』という事象を避けるためあえて 4 段階評価とした．

ファッションの専門学校の学生 52 人でデータを作成している．印象の定量的な評価を 100 人中何人がある印象を感じるかと定義しているため，1 枚の服飾画像に対して複数人でタグ付けを行ったものの方が望ましい．しかし，タグ付けにかかるコストを勘案した結果，データ数を優先したためにこのデータセットは 1 枚の服飾画像に対して 1 人でタグ付けを行ったものになっている．これは，1 枚の服飾画像に対しての精度は下がるが，画像枚数が多ければ全体の傾向は掴めると考えた結果でもある．

表 1 から表 3 はカテゴリ，キーワードと付与した商品数を示している．バイナリ評価のタグは一部のみを抜粋している．また，各商品には色違いのものが複数存在しており，23852 個の商品，71658 枚の画像が存在している．ここで，色違いの商品に対しては同じタグが付くとしている．今回対象としている画像は EC サイトで公開されているものを利用しており，基本的には服飾単体のものであるが，人が着ているもの，マネキンが着ているものも多少混じっている．

1）北海道大学大学院情報科学研究科．Graduate School of Information Science and Technology, Hokkaido University, Sapporo, Hokkaido, Japan
2）北海道大学大学院情報科学研究院．Faculty of Information Science and Technology, Hokkaido University, Sapporo, Hokkaido, Japan

Copyright © 2019 by Information Processing Society of Japan and
The Institute of Electronics, Information and Communication Engineers
All rights reserved.

表 1　カテゴリーとその商品数

カテゴリー	商品数
カットソー	3291
ブラウス	3657
ニット	5172
コート	908
ジャケット	907
ブルゾン	449
スカート	3536
パンツ	2431
ワンピース	3424
スーツ	77

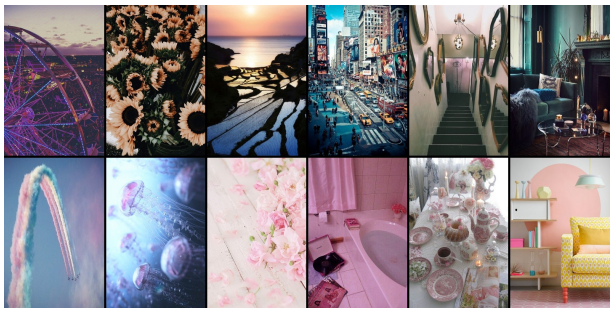

図 1　Brand Image Dataset の画像例　一行目がアヴァンギャルドなブランド A で二行目がキラキラかわいいブランド B に対応する画像

表 3　バイナリ評価のタグとその商品数 (一部)

タグ	観点	商品数
夏らしい，夏に着たい	素材感	4365
ガーリー	服のイメージ	2873
パステル	色彩	11622
フレア	シルエット (フォルム)	1467
花柄	柄・模様	1599
フリル	特徴・効果 1	970
着回ししやすそう	特徴・効果 2	13753
お祭り・花火大会	用途・シーン	1835

3.2　Brand Image Dataset

風景・屋内の画像に対して対応するブランドイメージを持つブランドと「ひまわり」「星」といった写っているもの，「引き込まれる」「優しい」といった印象などがタグとして付与されている．

ブランドは，アヴァンギャルドなブランド A とキラキラかわいいブランド B の 2 つである (図 1)．この 2 つのブランドは違いが分かりやすく，イメージの差も分かりやすく出ると考えられる．ファッションの専門学校生 90 人に対して Pinterest[14] 上で，ブランドイメージに合う画像の収集及び，その収集された画像に対してタグ付けを依頼し，画像に即したタグを自由に付けたものとなっている．画像は建物内部と風景に絞って収集してもらい，タグは後で整理・統合を行った．最終的に 5562 枚の画像が集まり，ブランド A に対応する画像が 2906 枚，ブランド B に対応する画像が 2656 枚である．タグの種類は 1051 種類となった．表 4 は出現回数が上位 20 個のタグである．

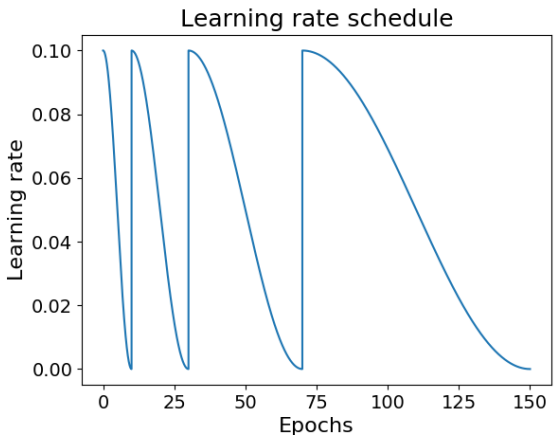

図 2　SGDR における学習率の変化

表 4　Brand Image Dataset における出現回数上位 20 個のタグ

タグ	出現回数	タグ	出現回数
優しい	1497	インパクトがある	1047
素敵	1446	優雅なひと時	1029
真っ白	1370	緑	1027
女の子が好きそう	1341	雑誌に出てきそう	998
ピンク	1287	清潔感がある	972
スタイリッシュ	1213	素敵な	963
美しい	1171	魅了される	940
きれい	1166	写真を撮りたい	934
清楚	1115	可愛い	931
植物	1111	映画のワンシーン	926

4　データの学習

学習には ResNet-50 を用い，ImegeNet で事前学習されたものをファインチューニングしている．今回用いるデータは不均衡データであるため，損失関数は重み付きのクロスエントロピーを基本的に用い，4 段階評価を行ったタグに関しては重み付き MSE を用いた．最適化手法としては SGDR[15] を用いた．これは，SGD+momentum[16] の学習率をコサインアニーリングで変化させるという手法である (図 2)．$momentum = 0.9, batch\_size = 50, weight\_decay = 0.0001, epoch = 150$, 初期学習率 = 0.1 とした．

4.1　Fashion Impression Dataset の学習

23852 個の商品の内，9 割分の 21466 個を学習データに，1 割分の 2386 個をテストデータとした．商品 1 個につき，色違いなど含めて対応する画像は複数枚あるので，画像枚数としては，学習データが 64462 枚，テストデータが 7196 枚となっている．

バイナリ評価のタグに対しては，通常の正解率で性能を評価すると，全て 0 と出力するだけで正解率が 8 割，9 割を超えてしまい，適切な評価が出来ない．このため正解ラベルが 1 であるものの正解率と，正解ラベルが 0 であるものの正解率の平均を性能評価に用いる．

表 5 に学習結果を示す．但し，4 段階評価のタグ，バイナリ評価のタグの評価値は各タグの評価値の平均を

Copyright © 2019 by Information Processing Society of Japan and
The Institute of Electronics, Information and Communication Engineers
All rights reserved.

表２　4 段階評価のタグ (第一印象) の商品数

	とても当てはまる	やや当てはまる	あまり当てはまらない	全く当てはまらない
かわいい	7285	9138	5793	1636
きれい	8247	9908	4526	1171
かっこいい, クール	2922	6436	8098	6396
モテ服，モテる	2005	9145	9581	3121
セクシー	1422	5143	9822	7465
おしゃれ	3072	11073	8456	1251

表５　Fashion Impression Dataset の学習結果

MAE (4 段階)	精度 (バイナリ)	精度 (カテゴリ)
0.236	70.0%	91.0%

表６　カテゴリーとその正解率

カテゴリー	精度	商品数
カットソー	82.0%	3291
ブラウス	91.7%	3657
ニット	93.3%	5172
コート	88.3%	908
ジャケット	75.4%	907
ブルゾン	62.1%	449
スカート	98.4%	3536
パンツ	94.5%	2431
ワンピース	95.7%	3424
スーツ	3.7%	77
カテゴリ正解率	91.0%	23852

とったものになっている．

バイナリ評価のタグの精度は 7 割となっているが，これは学習がうまくいかなかったというわけではなく，人によって意見が分かれるものに対して一人でしかタグの付与を行っていないというデータの性質の問題であると考えられる．

表 6 から表 8 は個別のキーワードにおける学習結果である．表 6 を見るとスーツだけかなり正解率が低いが，これはスーツのデータ数が他と比べてかなり少ないためだと考えられる．また，表 8 を見ると「シルエット (フォルム)」「柄・模様」「特徴・効果 1」といった，外観的な要素に対する正解率は高めで，他の印象語のタグは低めという傾向がある．これは，外観的な要素のほうが人の意見も一致しやすく，学習させやすいということが考えられる．また，「かわいい」系のブランドと「大人っぽい」系のブランドの間で「かわいい」や「大人っぽい」といったタグの出力値にはっきりと分かる差が出たことで学習が出来ていることを確認した．

図 3，図 4 は入力画像と各キーワードに対する CNN の出力値である．画像に書かれている数字は CNN の出力値であり，0.2 刻みの区間の出力値であった画像が左から並べてある．

図３　入力画像と「フェミニン」というキーワードに対する出力値

図４　入力画像と「デート」というキーワードに対する出力値

図５　入力画像と出力値
上の行はブランド A に，下の行はブランド B に対応する画像．出力値が 1 に近いほどブランド A に，0 に近いほどブランド B に対応したと判断している

4.2　Brand Image Dataset の学習

5562 枚の内，4449 枚を学習データに，1113 枚をテストデータとした．今回の学習ではタグ情報のみでの学習，画像のみでの学習，タグと画像の両方を用いた学習，予想したタグと画像を用いた学習を行う．タグ情報のみの学習では多層パーセプトロンを用いる．タグと画像の両方を用いた学習では，ImageNet で事前学習した ResNet-50 で特徴量抽出を行い，その特徴量とタグを多層パーセプトロンに入力して学習する．予想したタグと画像を用いた学習では，タグを学習させた ResNet-50 でタグの予測と特徴量抽出を行い，これらを多層パーセプトロンに入力して学習する．

表 9 に学習結果を載せる．これらを見るとタグから得られる情報量が非常に大きく，CNN では十分な特徴抽出が出来ていないと考えられる．また，予想したタグを入力に追加しても精度が上がっていないことが分かる．これは，タグ予想の精度が 65%程度と低かったことや，タグの予想には CNN を用いていたため事前学習した CNN の特徴量から得られる程度の情報しか得られなかったといったことが原因として考えられる．

図 5 は入力画像とその出力値である．上の行はブランド A に，下の行はブランド B に対応する画像である．出力値が 1 に近いほどブランド A に，0 に近いほどブランド B のイメージに沿ったものと判断されたものである．

Copyright © 2019 by Information Processing Society of Japan and
The Institute of Electronics, Information and Communication Engineers
All rights reserved.

表 7　4 段階評価のタグ (第一印象) の評価値 (MAE)

	1(とても当てはまる)	0.66(やや当てはまる)	0.33(あまり当てはまらない)	0(全く当てはまらない)	ALL
かわいい	0.290	0.143	0.276	0.550	0.247
きれい	0.278	0.127	0.308	0.590	0.236
かっこいい、クール	0.464	0.224	0.162	0.338	0.258
モテ服、モテる	0.445	0.186	0.166	0.396	0.224
セクシー	0.547	0.291	0.131	0.281	0.235
おしゃれ	0.389	0.126	0.223	0.512	0.217

表 8　バイナリ評価のタグとその正解率 (一部)

タグ	観点	正解率
夏らしい，夏に着たい	素材感	74.2%
ガーリー	服のイメージ	67.8%
パステル	色彩	60.6%
フレア	シルエット（フォルム）	81.3%
花柄	柄・模様	90.9%
フリル	特徴・効果 1	76.6%
着回ししやすそう	特徴・効果 2	63.0%
お祭り・花火大会	用途・シーン	69.87%

表 9　Brand Image Dataset の学習結果

tag only	image only	image and tag	image and predicted_tag
89.1%	72.3%	90.1%	72.1%

5　まとめと今後の展望

今回の研究では，人の印象を定量的に評価するシステムの作成を目指し，データの作成と CNN を用いた学習を行った．しかし，今回の精度は 1 人でタグ付けしたものとの比較であるため，十分な評価が出来ていない．このため，複数人でタグを付けたものの平均と CNN の出力値を比較してどれだけ近づいているかを検証する必要がある．また，今回は教師データを通常の物体認識と同様の枠組みで使用した．しかし，印象という主観が交じるデータであるため全く同様に扱っていいかには疑問が残る．このため，ノイズ混じりのデータを学習するアルゴリズムなどを参考にして教師データの扱いや損失関数の設計などを行う必要がある．

参考文献

[1] steady. 2017 年 8 月号. 宝島社, 2017. pp. 20-21.

[2] Aron Yu and Kristen Grauman. Fine-Grained Visual Comparisons with Local Learning. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 192–199. IEEE, jun 2014.

[3] M. Hadi Kiapour, Xufeng Han, Svetlana Lazebnik, Alexander C. Berg, and Tamara L. Berg. Where to Buy It: Matching Street Clothing Photos in Online Shops. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 3343–3351. IEEE, dec 2015.

[4] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1096–1104. IEEE, jun 2016.

[5] Xiaodan Liang, Si Liu, Xiaohui Shen, Jianchao Yang, Luoqi Liu, Jian Dong, Liang Lin, and Shuicheng Yan. Deep Human Parsing with Active Template Regression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 37, No. 12, pp. 2402–2414, dec 2015.

[6] Shuai Zheng, Fan Yang, M. Hadi Kiapour, and Robinson Piramuthu. ModaNet. In *2018 ACM Multimedia Conference on Multimedia Conference - MM '18*, pp. 1670–1678, New York, New York, USA, 2018. ACM Press.

[7] Tetsu Matsukawa and Einoshin Suzuki. Person re-identification using CNN features learned from combination of attributes. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pp. 2428–2433. IEEE, dec 2016.

[8] Kota Yamaguchi, Takayuki Okatani, Kyoko Sudo, Kazuhiko Murasaki, and Yukinobu Taniguchi. Mix and Match: Joint Model for Clothing and Attribute Recognition. In *Procedings of the British Machine Vision Conference 2015*, pp. 51.1–51.12. British Machine Vision Association, 2015.

[9] Bo Zhao, Jiashi Feng, Xiao Wu, and Shuicheng Yan. Memory-Augmented Attribute Manipulation Networks for Interactive Fashion Search. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6156–6164. IEEE, jul 2017.

[10] Sirion Vittayakorn, Takayuki Umeda, Kazuhiko Murasaki, Kyoko Sudo, Takayuki Okatani, and Kota Yamaguchi. Automatic attribute discovery with neural activations. In *ECCV*, 2016.

[11] Yan Xiong, Yan Li, Peiyuan Pan, and Yu Chen. A regression-based Kansei engineering system based on form feature lines for product form design. *Advances in Mechanical Engineering*, Vol. 8, No. 7, p. 168781401665610, jun 2016.

[12] Kristen Vaccaro, Sunaya Shivakumar, Ziqiao Ding, Karrie Karahalios, and Ranjitha Kumar. The Elements of Fashion Style. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology - UIST '16*, pp. 777–785, New York, New York, USA, 2016. ACM Press.

[13] Hossein Talebi and Peyman Milanfar. NIMA: Neural Image Assessment. *IEEE Transactions on Image Processing*, Vol. 27, No. 8, pp. 3998–4011, aug 2018.

[14] ピンタレスト: Pinterest. `https://www.pinterest.jp/`.

[15] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic Gradient Descent with Warm Restarts. aug 2016.

[16] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, Vol. 323, No. 6088, pp. 533–536, oct 1986.

Copyright © 2019 by Information Processing Society of Japan and
The Institute of Electronics, Information and Communication Engineers
All rights reserved.