The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

キャラクター顔特徴量の個別漫画への適応手法

Character Representation Adaption in Individual Manga

坪田亘記 Koki Tsubota　小川徹 Toru Ogawa　山﨑俊彦 Toshihiko Yamasaki　相澤清晴 Kiyoharu Aizawa

東京大学
The University of Tokyo

Our goal is character clustering of individual manga with bounding boxes of characters' faces and frames. We propose a method to adapt character representations to individual manga for clustering. We use deep metric learning by making positive and negative pairs considering manga specific nature that the same character tends to appear in the near page and the characters in the same frame almost always differ. We showed that adapting to individual manga improves the accuracy of clustering.

1. はじめに

漫画のキャラクターの顔分類は重要な課題である．もしこれが出来れば，顔画像を名前と対応させることでキャラクター名での検索，吹き出しと対応させることで話者同定に応用できる．また，キャラクターを区別することで着色等一貫した処理に応用できる．

本研究では個別の漫画に対して，クラスタリングをすることでキャラクターの顔分類を行う．コマとキャラクターの顔のバウンディングボックスは得られているものとする．

クラスタリングにおいては，教師データを用いて汎用的な特徴抽出器を作成し，その特徴抽出器をテストデータに適用するのが一般的である. しかしこの手法は教師データとテストデータの分布が十分に同一であることを仮定しており, 漫画のように画風の幅が広い画像では性能が劣化しうる. そのため本研究では個別の漫画に対して特徴量を適応させることで性能の向上を図る．

今回前提としているバウンディングボックスを得る研究として，漫画における顔とコマの検出の研究がある．例えば Chu ら [Chu 17] は顔検出を行っており，小川ら [小川 17] は漫画における物体（顔，体，テキスト，コマ) の検出を行っている．また小川らはキャラクターの顔よりもコマが検出しやすいことを示している．そのため，顔分類に必要なキャラクターの顔のバウンディングボックスだけではなく，コマのバウンディングボックスも与えられることは大きな制約ではない．

2. 関連研究

2.1 漫画における画像特徴量

漫画における画像特徴量の作成・検討は，クラスタリングと検索の分野で行われている．柳澤ら [柳澤 17] は SURF 特徴量 [Bay 06] を用いた漫画の顔画像のクラスタリングを行っている．また松井ら [Matsui 16] は EOH 特徴量 [Levi 04] という hand-crafted な特徴量を用いた検索を行っている．その後成田ら [Narita 17] によって深層学習を用いた特徴量を用いたことで検索が改善されたことが報告されている．しかしながら，これらの手法は汎用性のある漫画の特徴量を目的としており，個別の漫画に適応した特徴量を作成をしているわけではない．

連絡先: 東京大学　情報理工学系研究科
〒 113-8656 東京都文京区本郷 7-3-1
TEL: 03-5841-6761
Email: tsubota@hal.t.u-tokyo.ac.jp

2.2 深層距離学習

画像に対して同じクラスが近く，異なるクラスが遠くなるような特徴量を作成する手法の一つに，深層距離学習がある．深層距離学習は fine-grained な画像認識・検索・顔認識といった分野で使用されている．

深層距離学習は深層学習において同じクラス同士は近づけ，異なるクラス同士は遠ざけるような損失関数を定めて学習を行う手法である．損失関数のうちの一つとしては constrastive loss [Hadsell 06] がある．contrastive loss の式は画像 I_i, I_j に対応する特徴量 x_i, x_j に対して t_{ij} を正ペアのときに 1，負ペアのときに 0 としたとき，以下の式で表される．

$$L_{contrastive} = \frac{1}{2} t_{ij} ||x_i - x_j||_2^2 + \frac{1}{2}(1 - t_{ij}) \left(\tau - ||x_i - x_j||_2\right)^2 \quad (1)$$

なお，τ はマージンである．

個別のドメインに適応した深層距離学習としては，Zhang ら [Zhang 16] の手法がある．この手法では，ビデオにおける顔のトラッキング情報とフレーム情報をもとに正ペア/負ペアを構築し深層距離学習を行い，最終的に顔画像のクラスタリングを行っている．本研究ではこの手法に倣い漫画特有の情報をもとに正ペア/負ペアを定義し，深層距離学習を行う．

3. 手法

3.1 概要

漫画に対して汎用性のある深層学習ベースの特徴抽出器から，個別の漫画に適応した特徴抽出器を作成する手法を提案する．手法の概要を図 1 に示す．最初に教師あり学習によって漫画に汎用性のある特徴抽出器（漫画用 CNN）を得る．その後，キャラクターの名前は与えられず，顔とコマのバウンディングボックスが与えられているという条件で，個別の漫画に対して適応した特徴抽出器を距離学習により作成する．漫画に汎用性のある特徴抽出器の学習に用いられる漫画と，その後適応する個別の漫画は異なることに注意が必要である．最後に，得られた特徴抽出器を用いて個別の漫画の顔画像に対する特徴量を抽出し，抽出した特徴量に対して K-means でクラスタリングを行う．

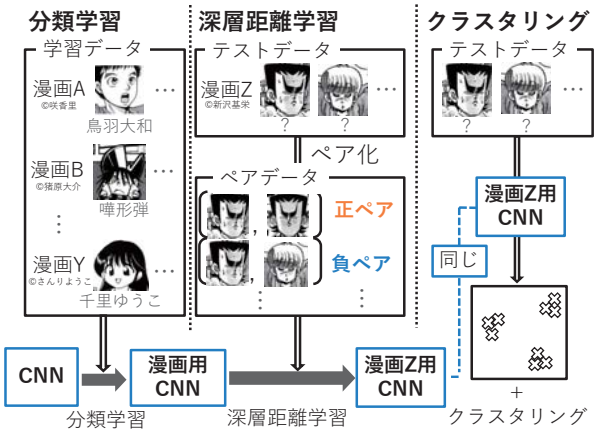


図 1: 手法の概要．分類学習によって漫画に対して汎用性のある特徴抽出器（漫画用 CNN）を作成し，その後深層距離学習により個別の漫画（漫画 Z）に適応した特徴抽出器（漫画 Z 用 CNN）を作成する．最後に K-means により分類を行う．

汎用性のある特徴抽出器の学習は成田ら [Narita 17] の手法に従って行い，個別の漫画に適応した特徴抽出器の学習は提案手法により行う．順に説明する．

3.2 分類学習による特徴量作成

成田ら [Narita 17] の手法に従い漫画に汎用性のある特徴抽出器（漫画用 CNN）を作成する．まず ImageNet [Russakovsky 15] で事前学習されたニューラルネットワークを用意する．その重みを初期の重みとして使用し，キャラクターの顔画像と名前のペアを用いて分類学習を行う．最後にネットワークの後方を除去することにより汎用性のある特徴抽出器を得る．

本研究では，ネットワークとして ResNet [He 16] の 50 層のネットワークを用いた．分類学習後には最終層の fc 層を除去し，代わりに L_2 正規化を行う層を追加したものを特徴抽出器とした．またこの特徴抽出器をベースラインの特徴抽出器とした．

3.3 深層距離学習による適応化

深層距離学習を行うためには正ペア/負ペアを作成する必要がある．しかしながら個別の漫画に対しては顔画像と名前のペアが与えられないので，明示的に正ペア/負ペアは与えられていない．一方で顔とコマのバウンディングボックスは得られているため，顔画像とその顔が属するコマとページが分かる．この情報に基づき正ペア/負ペアを推測する．

ここで漫画特有の性質として，同じキャラクターは近くのコマやページに登場する傾向があるということと，同一のコマに現れるキャラクター同士はほぼ異なるということが挙げられる．前者については，漫画では文脈としてキャラクターが一度登場するとしばらく登場するということから分かる．例を図 2 に示す．後者については，コマは基本的には情景を切り取ったものであるため，同じ情景には同じ人がほぼ登場しないということから分かる．例を図 3 に示す．

よって全ての顔画像について，漫画に汎用性のある特徴抽出器（漫画用 CNN）から得られた特徴量が近くかつ登場ページが近いものを正ペア，同じコマ内のキャラクターは負ペアと定めた．前者について具体的に説明すると，ある顔画像が与えられたときに，その顔画像が登場するページと前後一ページの顔

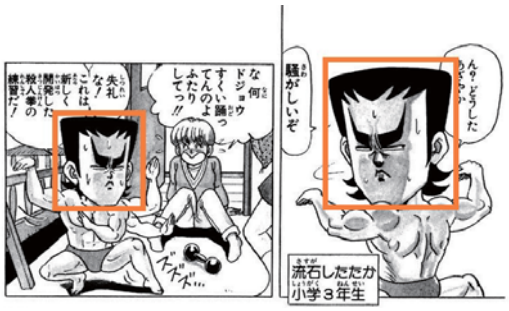


図 2: 同じキャラクターが近くに登場する例． ©新沢基栄

図 3: 同じコマに現れるキャラクター同士が異なる例． ©新沢基栄

画像の中で最も特徴量が近い顔画像を正ペアとした．また後者については，ある顔画像が与えられたときに，同一のコマに登場する全ての顔画像を負ペアとした．

さらに，教師データを増やすことによる性能改善を期待して，ペアの関係を伝搬させ，ペア数を増やした．すなわち I_i と I_j が正ペアかつ I_i と I_k が正ペアならば I_j と I_k は正ペア，I_i と I_j が正ペアかつ I_i と I_k が負ペアならば I_j と I_k は負ペアとした．なお，$I_\cdot$ は画像を意味する．

こうして定義した正ペア/負ペアを用いて深層距離学習で contrastive loss [Hadsell 06] を用いて fine-tune を行うことで個別の漫画に適応した特徴抽出器を作成する．最後に得られた特徴抽出器を用いて全ての顔画像から特徴量を抽出し，特徴量に対して K-means でクラスタリングを行う．

4. 実験

4.1 データセット

データセットとして Manga109 [Matsui 16, Fujimoto 16] を使用した．Manga109 とは学術利用可能な漫画データセットで，様々なジャンルの漫画が 109 冊含まれている．本研究では 109 冊のうち，1 タイトル 2 巻以上あるものに関しては 1 巻を除外し残った 104 冊を使用する．ページにまたがった顔が存在する可能性もあるため，ページは見開きとして扱った．

藤本ら [Fujimoto 16] による顔へのアノテーションは髪や耳を内包する形式ではないので，バウンディングボックスを上下左右に均等に拡大し 2 倍にした．バウンディングボックスは拡大しても漫画の全体ページの画像からはみ出ないようにクリッピングした．また元画像のサイズが小さい場合は顔認識が困難であるため，大きさが 30x30 pixel 以上である顔画像を対象として実験を行った．

104 冊のデータセットのうち，83 冊を漫画に汎用性のある特徴抽出器を作成するための分類学習用データ，10 冊を深層

表 1: テストデータにおける “その他” のクラスを除いたときのクラス数と画像数．括弧内は “その他” のクラスの有無（1/0）と画像数である．

タイトル	クラス数	画像数
ARMS	19 (1)	312 (6)
愛さずにはいられない	16 (1)	860 (60)
あっけら貫刃帖	14 (1)	596 (106)
あくはむ	17 (1)	1012 (9)
青すぎる春	23 (1)	508 (38)
天晴れ！ カッポーレ	14 (1)	831 (22)
ありさ[2]	15 (1)	863 (77)
BEMADER・P	17 (1)	1105 (82)
爆裂！ かんふー娘	61 (1)	988 (22)
ベルモンド	22 (0)	819 (0)
ラブひな 14 巻	20 (1)	1164 (44)

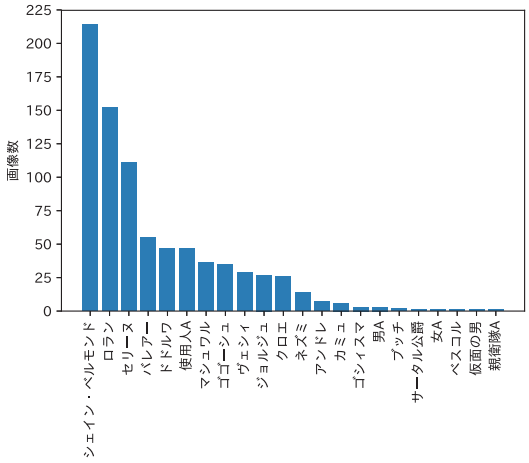


図 4: “ベルモンド” におけるクラスの分布をヒストグラムで表した．クラスのバランスが悪いデータであることが分かる．

距離学習のパラメータ決定用データ，11 冊をテストデータとして分割した．分類学習には，83 冊の漫画で 10 回以上出てくるキャラクターを用いた．クラス数は 1,031 であり，画像数は 67,328 であった，

テストデータにおけるクラス数と画像数を表 1 に示す．漫画ごとにクラス数，画像数が異なるため，クラスタリングの難易度が異なることに注意する必要がある．さらに図 4 に “ベルモンド” の顔画像のクラスの分布をヒストグラムで表した．クラスのバランスが悪いことが分かる．なおこの性質はどの漫画についても言える性質である．

また漫画においては “その他” に相当するクラスがある．通常のクラスはクラス内は同一人物である一方，“その他” のクラスは群衆の一人であったりクラス内でも異なる人物である場合がある．そのためクラスタリング時には除外して扱った．

4.2 評価方法

個別の漫画ごとのクラスタリングを評価の対象とする．クラスター数をキャラクター数としてクラスタリングを行ったときの正規化相互情報量（NMI）と精度で評価を行った．NMI, 精度はクラスタリングの評価に一般に用いられる指標であり順に説明する．説明のために文字の表記を定める．N をデータ数, データ i $(1 \leq i \leq N)$ に対して y_i をクラスター，c_i を正解のラベルとする．NMI は $Y = \{y_1, y_2, \cdots, y_N\}, C = \{c_1, c_2, \cdots, c_N\}$ と表すと，式 2 で定義される．

表 2: クラスタリングの評価

	NMI	精度
ベースライン	0.63	0.48
提案手法	**0.71**	**0.64**
提案手法（ページ条件なし）	0.67	0.58

$$\mathrm{NMI} = \frac{I(Y, C)}{[H(Y) + H(C)]/2} \tag{2}$$

$I(\cdot, \cdot)$ は相互情報量，$H(\cdot)$ は情報量である．また精度は式 3 で定義される．

$$(\text{精度}) = \frac{1}{N} \max_{mapping} \sum_{i=1}^{N} \delta_{c_i, mapping(y_i)} \tag{3}$$

mapping は排他的にクラスターにラベルを割り当てる関数であり精度が最大になるような *mapping* を使用する．また δ はクロネッカーのデルタである．

4.3 実装の詳細

分類学習時においては，200 epoch 行い，学習率は最初の 100 epoch は 10^{-3}，残りは 10^{-4} とした．ミニバッチのサイズは 64，weight decay は 10^{-4} とした．入力画像は 256x256 の画像にリサイズし，224x224 にランダムクロップし，ミラーリング，ガンマ補正を使用してデータ拡張を行った．

深層距離学習においては，100 epoch 行い，学習率は 10^{-3} を使用した．ミニバッチのサイズは 64 とし，ミニバッチ内での全てのペアについて学習を行った．データ拡張としては先述したランダムクロップとミラーリングを使用した．式 1 におけるマージン τ はパラメータ決定用データを用いてパラメータの決定を行い，1.4 とした．

4.4 結果

まず，テストデータの各漫画でクラスタリングを行ったときの，精度と NMI の平均値を比較することで提案手法の有効性を示す．表 2 はベースラインと提案手法，さらに提案手法の条件を変えた場合のクラスタリング結果を比較したものである．条件の変え方として，ページを考慮せず特徴量が近いものを正ペアとした場合との比較を行っている．ベースラインと提案手法を比較すると NMI は 0.08，精度は 0.16 向上している．特徴量が改善されていることが分かる．またページ条件なしと比較して提案手法が NMI，精度において上回っていることから漫画特有の情報であるページという情報を入れた効用があることが分かる．

次に各漫画に対して評価したときの NMI を図 5 に示す．いずれの漫画についても，ベースラインよりも提案手法の方が上回っていることが分かる．

なお，テストデータの各漫画において作成した正ペア/負ペアを，適合率を用いて評価したところ，それぞれ 0.84, 0.93 であった．提案した正ペア/負ペアは十分に正しいことが分かった．

最後に “天晴れ！ カッポーレ” に対して得られた特徴量を t-SNE を用いて可視化した．得られた可視化結果を図 6 に示す．特徴量がよりまとまり，クラスタリングしやすい特徴量になっていることが分かる．

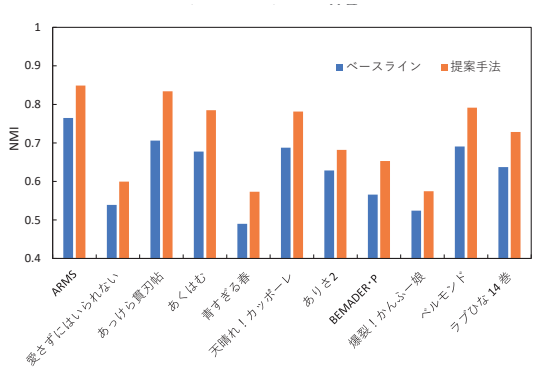
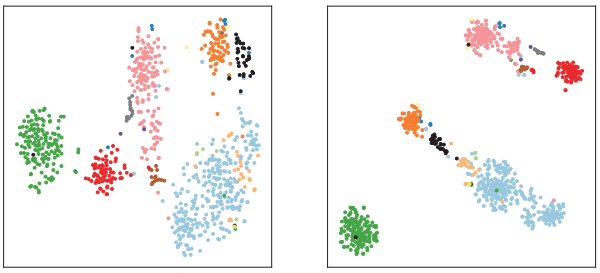


図 5: 各漫画におけるクラスタリングの結果の NMI での評価．いずれの漫画においても提案手法がベースラインよりも上回っている．



ベースライン　　提案手法

図 6: ”天晴れ! カッポーレ”における特徴量の可視化．各点の色がキャラクターに対応する．提案手法の方が特徴量がより集まっていることが分かる．例えば薄い水色のクラスと薄い橙色のクラスは，ベースラインでは区別することが難しいものの，提案手法により区別しやすくなっている．

5. まとめ

本研究では，コマとキャラクターの顔のバウンディングボックスが得られているという条件で，個別の漫画に対して漫画特有の正ペア/負ペアを導入して深層距離学習を行うことで特徴量を作成し，クラスタリングを行った．深層距離学習を行わない場合と比較して NMI が平均 0.08，精度が平均 0.16 向上した．

本手法を新しく与えられた漫画に適用するときには，実際にはコマと顔の検出を行う必要がある．そのため，顔分類の性能は検出の精度に依存する．そこでバウンディングボックスと顔の分類を同時に行うということも発展として考えている．

参考文献

[Bay 06] Bay, H., Tuytelaars, T., and Van Gool, L.: Surf: Speeded up robust features, *European Conference on Computer Vision*, pp. 404–417 (2006)

[Chu 17] Chu, W.-T. and Li, W.-W.: Manga FaceNet: Face Detection in Manga Based on Deep Neural Network, in *ACM International Conference on Multimedia Retrieval*, pp. 412–415 (2017)

[Fujimoto 16] Fujimoto, A., Ogawa, T., Yamamoto, K., Matsui, Y., Yamasaki, T., and Aizawa, K.: Manga109 Dataset and Creation of Metadata, in *International Workshop on coMics ANalysis, Processing and Understanding* (2016)

[Hadsell 06] Hadsell, R., Chopra, S., and LeCun, Y.: Dimensionality reduction by learning an invariant mapping, in *IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 2, pp. 1735–1742 (2006)

[He 16] He, K., Zhang, X., Ren, S., and Sun, J.: Deep Residual Learning for Image Recognition, in *IEEE Conference on Computer Vision and Pattern Recognition* (2016)

[Levi 04] Levi, K. and Weiss, Y.: Learning object detection from a small number of examples: the importance of good features, in *IEEE Conference on Computer Vision and Pattern Recognition* (2004)

[Matsui 16] Matsui, Y., Ito, K., Aramaki, Y., Fujimoto, A., Ogawa, T., Yamasaki, T., and Aizawa, K.: Sketch-based manga retrieval using manga109 dataset, *Multimedia Tools and Applications* (2016)

[Narita 17] Narita, R., Tsubota, K., Yamasaki, T., and Aizawa, K.: Sketch-based Manga Retrieval using Deep Features, in *International Workshop on coMics ANalysis, Processing and Understanding* (2017)

[Russakovsky 15] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge, *International Journal of Computer Vision*, Vol. 115, No. 3, pp. 211–252 (2015)

[Zhang 16] Zhang, Z., Luo, P., Loy, C. C., and Tang, X.: Joint face representation adaptation and clustering in videos, in *European Conference on Computer Vision*, pp. 236–251 (2016)

[小川 17] 小川 徹, 山﨑 俊彦, 相澤 清晴 : 漫画物体検出に向けた検出器の並列化, 情報科学技術フォーラム (2017)

[柳澤 17] 柳澤 秀彰, 渡辺 裕 : Deep Learning 特徴量を用いたマンガキャラクター顔画像の分類, 情報科学技術フォーラム (2017)