

大規模マンガ画像判別手法の検討

山仲 一颯^{1,a)} 伊東 栄典²

概要：近年、海賊版サイトによるマンガ画像の公開が問題である。マンガは電子書籍と印刷媒体で発行されているため、マンガ画像のスキャンは簡単で、そのため DRM(デジタル権利管理) だけでは海賊版対策に有効とは言えない。我々は海賊版マンガを抑制するため、マンガ画像から、マンガのタイトル・巻数・話数・キャラクター名の判別器を研究している。判別精度が十分で、かつ処理速度が高速ならば、迷惑メールフィルタのような画像フィルタが実現でき、海賊版利用の抑止となる。以前より提案してきた判別手法と、その改善案を比較的大規模のマンガ画像に対して適用した。本発表では提案手法と判別結果について説明する。

キーワード：海賊版マンガ, フィルタリング, 画像認識, パターン認識, 機械学習

A study on large-scale manga image classification method

ISSA YAMANAKA^{1,a)} EISUKE ITO²

Abstract: Pirate edition of Manga is a serious problem. Since manga is published in digital e-book and paper book, it is easy to scan manga images. So DRM (digital rights management) is not so effective to against piracy. In order to suppress pirated manga, we are researching a discriminator for manga titles, volume numbers, story numbers, and character names from manga images. If classification accuracy is sufficient and the processing speed is high, it is possible to realize manga image filter. And the filter will deter the use of pirated copies. We applied the previously proposed classification method and its improvement plans to relatively large manga images. In this paper, we report the new classification method and the classification results.

Keywords: Japanese manga, pirate edition, bootleg, filtering, image recognition, machine learning

1. はじめに

近年、マンガ海賊版が問題となっている。海賊版とは著作物を、著作権者や発行者に無断で複製し提供するものを言う。2018 年には海賊版サイト「漫画村」が爆発的に利用された [1]。政府は 2019 年 4 月 13 日に知的財産戦略本部会合・犯罪対策閣僚会議を開き、ネット上の海賊版サイトに対する緊急対策を行い、日本国内の ISP で侵害コンテンツの通信をブロック可能にした。しかしながら 2023 年 1 月現在でも海賊版サイトは存在している。

正規版のマンガは印刷された紙媒体の本だけでなく、電

子書籍でも販売されている。紙のマンガはスキャンが簡単であり、電子書籍でもスクリーンショットが可能な場合がある。そのため DRM (デジタル権利管理) による利用制限だけでは有効ではない。

我々は、海賊版マンガを抑制を目指して、マンガの画像から、マンガのタイトル・巻数・話数を推定する判別手法を研究している [2]。判別精度が十分で、かつ処理速度が高速ならば、迷惑メールフィルタのような海賊版マンガフィルタが実現でき、海賊版利用の抑止となる。海賊版サイトはクラウド事業者のサーバ・ストレージにデータを保持しているため、クラウド事業者に対する海賊版マンガ画像検出も期待できる。

今回、収集プログラムを用いて、購入済の (利用権利が

¹ 九州大学工学部電気情報工学科

² 九州大学情報基盤研究開発センター

^{a)} yamanaka.issa.901@s.kyushu-u.ac.jp

ある)マンガの画像を収集した。購入済みマンガの、海賊版マンガページも収集した。本論文では、以前より提案してきた手法($k \times k$ 分割と螺旋順によるベクトル化)に加え新たに提案する手法(ひし形分割による行列化)を述べる。収集した比較的大規模なマンガ画像に対して2つの手法を適用する。

2. マンガ画像判別器の構成

本研究で考えるマンガ画像判別器は、1ページの画像、1話のページ画像リスト、1巻の画像リストを入力し、それらの画像または画像リストから特徴量を抽出することで、判別や分類を行う。マンガページ画像が1枚与えられた場合の判別を考える。最も粒度の細かい分類としては、そのページ画像が、どのマンガタイトル(シリーズ)の何巻の何話の何ページかを特定するものである。最も粒度の粗い分類は、マンガのタイトル推定である。人間は自分の知るマンガの画像を見れば、それがドラゴンボールなのかワンピースなのかのような、タイトル判別は容易にできる。画像に含まれるキャラクター部分を判別できれば、本研究の機械的判別機でもタイトル推定は可能であろう。中間的な分類は、タイトルおよび巻数までの推定や、タイトル・何巻・何話までの推定になる。

入力として1話の連続するページ画像リストからの判別も考えられる。判別器が推定するのは、タイトル、タイトルと巻数、タイトル・何巻・何話、の3通りが考えられる。連続するページ数が多ければ、確率的にで判別精度が向上するだろう。

マンガ画像判別器の入出力を表1にまとめる。また図1にマンガ画像判別器の構成を示す。

表1 マンガ画像判別の入出力

入力	1ページ 1話のページ画像リスト 1巻のページ画像リスト 1タイトルの全ページ
出力	タイトル, 巻, 話, ページ, キャラクタ, 作者

3. 用いたデータ

機械学習でマンガ判別器を作る場合、学習用のマンガ画像が必要である。マンガ画像の入手手段として、紙マンガのスキャン、電子書籍のキャプチャ、既存マンガデータセット、出版社からの提供が考えられる。今回は、正規版として我々が購入済みのマンガからページキャプチャした。

また、購入済みマンガと同じページを含む海賊版マンガページも入手した。海賊版サイトからの画像入手と分析は、著作権法 第三十条の四(著作物に表現された思想又は感情の享受を目的としない利用)に該当するため、購入していないマンガでも分析できるとの見解を法律家から受けてい

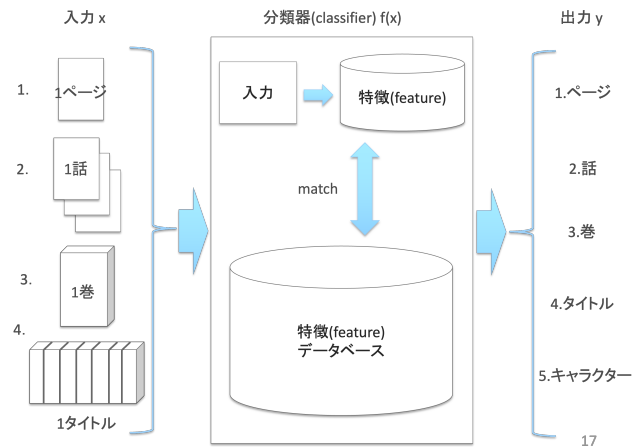


図1 マンガ画像分類器の構成

る。表2に、本研究で用いたマンガページ画像数を示す。

表2 正規版と海賊版・収集ページ数

版	収集元	ページ数
正規版	ジャンプ+ (2014年50号~2023年27号)	203,092
海賊版	Dr.STONE (1~232話)	4,694
	呪術廻戦 (1~224話)	4,360
	チェンソーマン (1~97話)	1,866
	鬼滅の刃 (1~205話)	4,250

4. 螺旋型の白比率の変化量列

1つのマンガページ画像が、どのマンガのページかを当てる推定手法について、各ページの画素の白比率(=白画素数/全画素数)を用いる方法を提案している[2]。提案済の手法を、表2で示した大規模データに適用した。

4.1 ベクトル化

白比率に基づいて1ページをベクトル化する手順を以下に示す。

- 1ページのマンガ画像を $k \times k$ に分割
- 各区域の白黒画素数を用いて数値化
- 中心から外側に時計回りに螺旋状に並べる
- 前区域との現区域の白画素率の差でベクトル化

図2に区画分割と螺旋順番を示す。なお k は $k=3, 5, \dots, 49$ の奇数とした。奇数にした理由は中心区画を決めるためである。

ページ判別はベクトル間のコサイン類似度で行う。ページ i と j の類似度を、ベクトル i と j との $\cos$ 類似度 $\cos(i, j)$ で算出すると、同一ページでは $\cos(i, i) = 1$ になり、2つのベクトルが大きく異なる場合に $\cos(i, j)$ は0に近くなる。入力ページ x に対し、 x と調査済ページとの $\cos$ 類似度が最も高いものを判別結果とする。

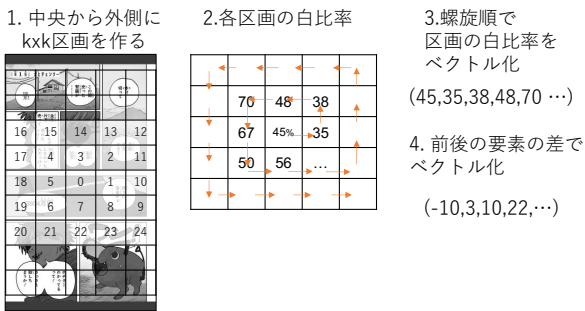


図 2 白比率によるベクトル化

4.2 正規版での比較

約 20 万ページの正規版マンガで、全ページを類似度で判別できるか調査した。約 20 万 $\times$ 20 万の類似度を表示すると見づらい。そこで一部分だけを図 3 に示した。図 3 は 500 $\times$ 500 画像の cos 類似度のヒートマップである。

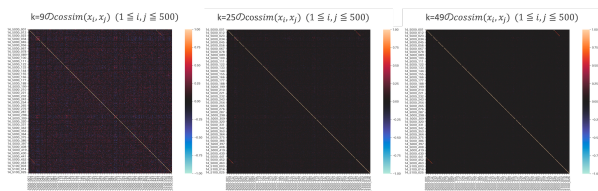


図 3 正規版画像の cos 類似度 (500 画像分, $k = 9, 25, 49$)

画像 i に対して最もコサイン類似度の高い j において、 $i = j$ ならば True Positive, $i \neq j$ ならば False Negative とし精度 (Accuracy) を計算した。表 3 に結果を示す。

表 3 正規版での判別精度

分割数 k	TruePositive	FalseNegative	平均精度
9	200,948	2,144	0.98944
25	200,951	2,141	0.98946
49	200,954	2,138	0.98947

判別が外れたページは、同じ画像が複数含まれていたためである。そのため類似度 1 であり、プログラムの判別ではハズレとされた。これらを正解とすると、平均 Accuracy は 100% になる。



図 4 異なるページが同一画像である例

4.3 海賊版ページの判定

表 2 に示す海賊版ページを、正規版ページから判別した。分割数 k を 3 から 49 に変化させ、海賊版と正規版ページの特徴量を作成し、海賊版と正規版との cos 類似度を計算した。

海賊版 4,694 画像 $\times$ 正規版 203,902 画像の cos 類似度の結果は見づらいため、「Dr.STONE」の第 1 話だけの結果を図 5 に表示する。縦軸が海賊版で横軸が正規版で、200 画像 $\times$ 200 画像の cos 類似度のヒートマップである。この図から、左上は海賊版 Dr.STONE 第 1 話と、正規版 Dr.STONE 第 1 話に該当する部分で、cos 類似度が高いことがわかる。

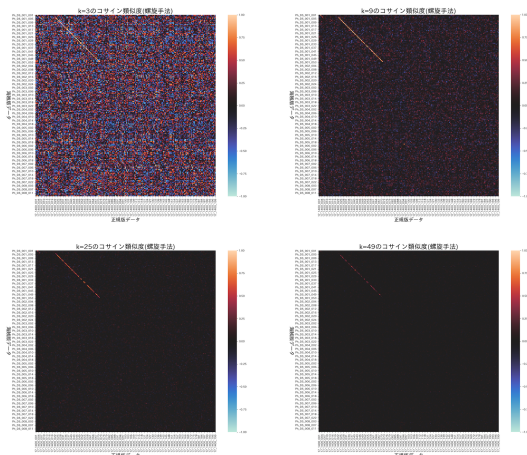


図 5 海賊版と正規版の類似度ヒートマップ (200 画像)

分割数 k が大きくなるにつれて判別精度が悪くなる。これは該当ページでも cos 類似度が低くなるためである。

海賊版ページ p_i の対し cos 類似度最大の正規版ページが r_j とする。 (p_i, r_j) が正しく同一のページを指す場合を True Positive, それ以外の場合を False Negative とし精度 (accuracy) を算出した。分割数 k の平均精度を図 6 に示す。分割数 k が 3 から 9 は精度は高くなり、分割数 $k = 9$ で最大値 0.972092 を得た。 $k = 11$ 以降は精度が低くなる。

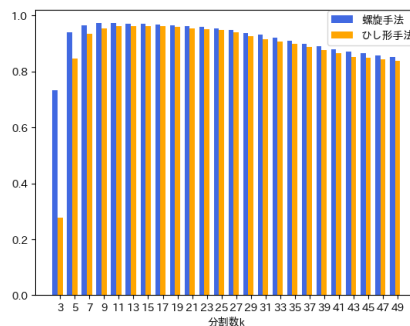


図 6 分割数 k による平均精度 (螺旋, ひし形)

5. ひし形分割による行列化

次にページ画像を数値の行列で表現する方法を提案する。図7に $k \times k$ 区画の分割、分割区画のひし形選定、行列の作成の手順を示す。この方法で各画像の特徴量（行列）を作成する。入力画像 x の行列と最も類似度が高い画像 y を判別結果とする。類似度は行列の $\cos$ 類似度とした。

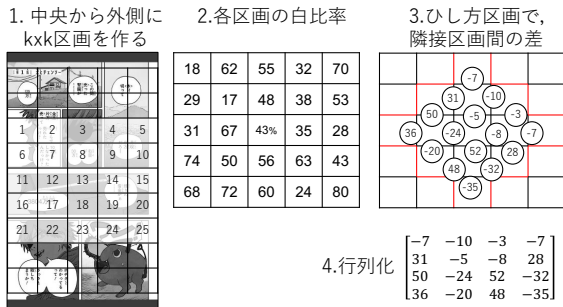


図7 ひし形区画の白比率による行列化

図5と同様に、ひし形分割による行列化で、海賊版ページと正規版ページの $\cos$ 類似度を算出した。200 画像の類似度ヒートマップを図8に示す。

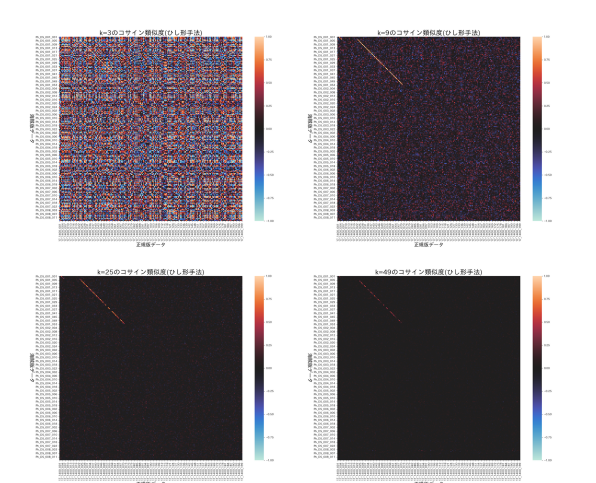


図8 海賊版と正規版間の類似度ヒートマップ (200 画像)

図6に分割数 k における平均精度を示している。この図によると、ひし形手法は分割数が小さいとき ($k = 3, 5, 7$) のとき螺旋手法に比べてに精度が低い。行列の要素が少ないため、画像の一意的な判別には不足するのだろう。

螺旋ベクトル化と、ひし形行列化の2つの手法で、海賊版ページの判別精度を比較する。分割数 k における平均判別精度を図6に示している。今回の場合、全ての分割数において螺旋手法の方が精度が高く、精度の最大値も螺旋手法が高い。本論文で示した2つの手法における海賊版ページ判別では、分割数 $k = 9$ の螺旋手法が精度最大である。

螺旋手法の方が精度が高い理由として、特徴量の要素数がある。螺旋手法のベクトル要素数は $k^2 - 1$ 個、ひし

形行列化による要素数は $(k - 1)^2$ 個である。そのため螺旋手法の方が特徴量が多い。また、ひし形手法は螺旋手法よりも画像中央部に着目した手法である。マンガ画像判別では、画像中央部に着目するより、ある程度大まかに全体を見るほうが良いのであろう。

データ量についても言及する。日本で出版されたマンガは膨大であり、ページ数も膨大になる。簡単な試算では、1年に出版されるマンガのページ数は10万ページ以上になる。そのため、マンガのページ画像判別では、各画像の特徴量は少ない方が良い。今回最大の精度である分割数 $k = 9$ の螺旋手法で前マンガページが判別できるとすると、1ページの特徴量は要素数80個 ($= 9^2 - 1$) のベクトルになる。ベクトルの要素がfloat型数値とすると、1画像の特徴量は1kB以下になり、大量の画像に対応できる。

6. おわりに

本論文では海賊版マンガを抑止するための、機械学習によるマンガ画像判別手法について述べた。画像分割と白比率によるページ推定手法を2つ提案した。ページ画像の中央部正方形を $k \times k$ に分割し、各区画で白比率 (= 白画素数/画素数) を用いる手法である。区画を螺旋状に並べるベクトル化と、ひし形の区画を行列化する手法を提案した。2つの手法における海賊版ページ判別では、どちらも単純な手法であるにもかかわらず高い精度で分類できた。分割数 $k = 9$ の螺旋手法が精度最大であった。

今回は説明していないが、登場キャラクターの顔画像判別も試している。キャラクター判別を用いたページ・話・巻の特定も試したい。他にも様々な類似度や距離を用いた手法も検討したい。提案した手法は、画像のズレ、傾き、ノイズに弱いと思われる。より耐性の高い判別手法を検討したい。更には自然画像やアニメ画像との比較も行いたい。

参考文献

- [1] 出版広報センター：深刻な海賊版の被害 | 出版広報センター, <https://shuppankoho.jp/damage/>. (Accessed on 2024/02/01).
- [2] 小城凱, 伊東栄典：マンガ画像の分類手法に関する研究, 情報処理学会 火の国情報シンポジウム 2023, pp. A2-3 (2023).