CH-007

漫画物体検出に向けた検出器の並列化
Parallel Detectors for Manga Objects

小川 徹* 山崎 俊彦* 相澤 清晴*
Toru Ogawa Toshihiko Yamasaki Kiyoharu Aizawa

1. まえがき

漫画は画像および文字を含む構造を持ったドキュメントの一種である．その構造や内容の認識・理解は検索や加工などの分野に応用が可能である．本研究では画像の認識・理解の基本技術である一般物体検出に着目する (図 1)．これは画像中に含まれる各要素の位置 (外接矩形) およびその種類を推定する課題である．一部の漫画はデジタル環境で製作されているため，各要素を検出することは容易であるが，従来の漫画製作は紙の上にインクで描くものであり，存在する漫画画像の大部分は紙の原稿をスキャンしたラスタ画像である．したがってこれらの画像を活用するためにはラスタ画像に対する検出手法が重要になる．

検出する要素としてはコマ，テキストおよびキャラクター (顔および全身) の 4 つを対象とする．コマとは漫画の各ページに含まれている領域であり，通常 1 つのコマが 1 つの物語上のシーンを記述している．このコマを辿っていくことで読者は物語を追っていく．コマの形状は通常矩形だが，漫画によっては台形などの形状をとることもある．テキストは通常キャラクターの発話内容や状況説明 (ナレーション) などを担う要素である．多くのテキストはフキダシと呼ばれる漫画特有の表現を用いて配置される．これは画像中に領域を設けて，その中にテキストを配置するという手法である．一部のテキストはフキダシを設けずに背景の上に直接置かれることもある．この場合は作者の手書きで書き込まれることが多い．キャラクターは漫画の主となる要素である．一般にデフォルメされた人間であるが漫画によってはイヌやネコなどを主要なキャラクターとするものもある．

漫画画像中の物体検出技術は様々な応用が考えられる．たとえば電子漫画の配信では，スマートフォンの画面に合わせて漫画のコマを並べかえる処理が行われることがあるが，コマを検出する技術によりコマを自動で切り抜くことができる．テキスト検出は OCR および機械翻訳と組み合せることで翻訳版の漫画製作やセリフ検索に利用できる．日本の漫画は海外でも人気を博しており，翻訳版の製作の需要は高い．また漫画研究の分野においても物体検出は重要なものとなっている．たとえば Chu ら [1, 2] は漫画の画風を表現するための特徴量を提案しているが，特徴量の一部としてコマやフキダシ，キャラクターの場所および形状を利用している．Rigaud ら [19] はフキダシと話者との対応を推定するための手法を提案している．この研究では前処理としてキャラクターとフキダシの検出を行っている．Le ら [12] は漫画のページを内容を効率よく検索するための特徴量を提案している．この特徴量はコマの位置を利用して計算されている．

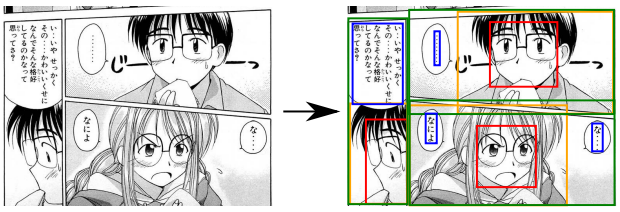


図 1: 本研究の目的 (ラブひな ©赤松 健)

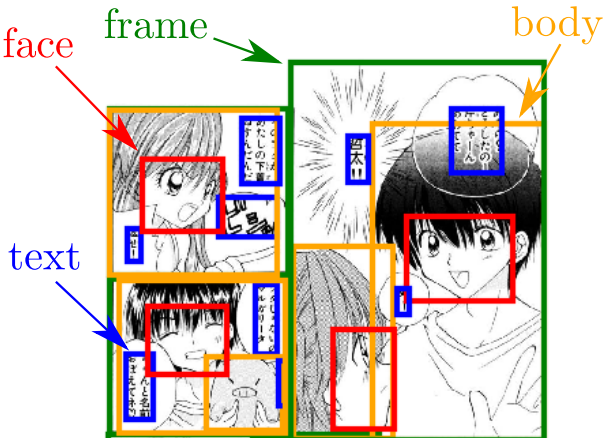


図 2: 藤本ら [6] による漫画アノテーション (爆烈！かんふー娘 ©うえだ 美貴)

2. 関連研究

2.1. 漫画データセット

松井らは Manga109 [15] という漫画画像データセットを公開している．このデータセットは様々なジャンル，年代の日本の商業漫画 109 冊分 21,142 ページの画像で構成されている．また藤本ら [6] により Manga109 へのアノテーションが行われている． この研究では Manga109 に含まれる画像に手動でアノテーションを行っており，主に “コマ (frame)”, “テキスト (text)”, “顔 (face)”, “全身 (body)” の 4 つの要素の外接矩形の座標がつけられている．(図 2)．さらに “テキスト” についてはその内容が，“顔” および “全身” についてはそのキャラクター名がメタ情報として付与されている．本研究ではこのデータセットを利用する．

Guérin らは eBDtheque [9] という漫画画像およびアノテーションのデータセットを公開している．このデータセットは 100 ページ分の漫画から構成されており，さらにコマやキャラクターなどに手動でアノテーションがつけられている．漫画研究において最も広く用いられているデータセットであるが，枚数が小規模であると

*東京大学大学院 情報理工学系研究科 電子情報学専攻

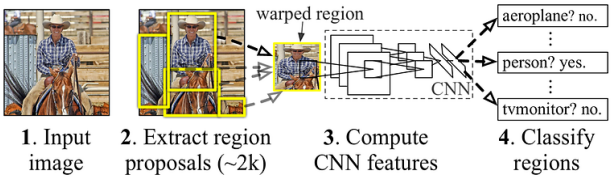


図 3: R-CNN [8] の概観

いう欠点がある．

Mohit らは COMICS [10] という漫画画像データセットを公開している．このデータセットはアメリカの漫画 3,948 冊から構成されており，漫画画像データセットとしては最大規模である．Mohit らは一部のページ (500 ページ) に手動でコマのアノテーションを付与し，それを用いて物体検出手法である Faster R-CNN [18] の学習を行っている．学習した Faster R-CNN をデータセット全体に適用することでコマの疑似アノテーションとして利用している．また OCR を利用してテキストの抽出を行っている．現時点では手動でのアノテーションが一部にしかついていないため，検出手法の評価には適していない．

2.2. 自然画像における物体検出手法

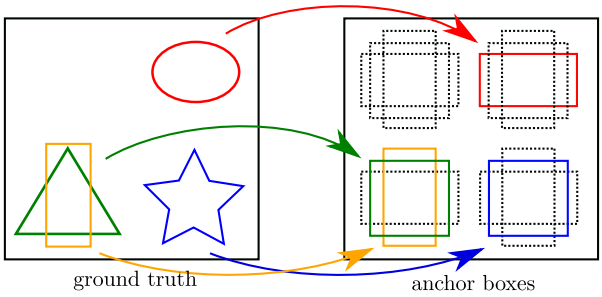
自然画像における物体検出は PASCAL VOC [4] や MS COCO [13] などのコンテストが開催されるなど，活発な研究がなされている分野である．自然画像での物体検出の基礎となる手法として R-CNN [8] がある．この手法ではまず Selective Search [22] などの手法を用いて，画像中から物体らしい候補領域を大量に選択する．そして各領域について特徴量を計算し，その領域がどの物体か (あるいは背景か) という分類問題を解くことで物体を検出している (図 3)．

R-CNN の後継として Fast R-CNN [7] や Faster R-CNN [18] がある．Fast R-CNN は画像全体の特徴を計算しておき，そこから候補領域を切り出すことで，候補領域ごとに特徴量を計算する必要があるという R-CNN の欠点を改善している．また外接矩形をより正確にするために候補領域に対する真の外接矩形のズレを推定するという問題も同時に解いている．Faster R-CNN は候補領域を提案する部分にも Region Proposal Network (RPN) と呼ばれるニューラルネットワークを用いることで，候補の提案と分類，矩形の推定という 3 つの問題を同時に解いている．

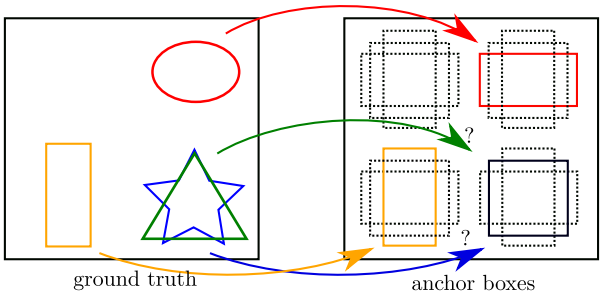
また近年では YOLO [16], YOLOv2 [17], SSD [14] や DSSD [5] などのより高速な手法も提案されている．これらの手法では RPN を用いず，あらかじめ決められた grid と呼ばれる領域ごとにクラスの分類問題および矩形の推定問題を解いている．また縦長や横長の物体に対応するために anchor (anchor box) と呼ばれる大きさや縦横比の異なる矩形を各 grid に複数個定めておき，物体と大きさや縦横比が近い anchor が反応するように学習を行っている．

上に挙げた手法は自然画像における物体検出で高い性能を示している．一方でその構造上，漫画のように物体が密に配置されている画像では上手く動作しない可

図 4: 漫画における密な物体の例 (極限サイクロン ⓒ 高波 伸)．“コマ” (緑)，“顔” (赤) および “全身” (橙) が大きく重なっている．



(a) 疎な画像



(b) 密な画像

図 5: “割り当て問題” の例．物体は anchor との重なりに応じて各 anchor に割り当てられる．疎な画像ではこの割り当てによって全ての物体が少なくとも一つの anchor に割り当てられる (a)．一方，密な画像では anchor 複数の物体間で競合することがあり，この場合割り当てをもたない物体が存在することになる．b では青い星と緑の三角形が競合をしている．

能性がある．たとえば漫画においては顔のアップのような構図 (図 4) がよく用いられる．このとき “コマ”，“顔” および “全身” の矩形は非常に近いものとなるが，領域のクラス分類を行う手法では領域ごとに 1 つのクラス (あるいは背景) を返すため，全てのクラスを正しく推定することはできない．特に YOLO や SSD はおおまかな anchor に正解領域を割り当てた上で学習を行っているため，密に配置された物体については一部が割り当てられない可能性がある (図 5)．本研究ではこの現象 (“割り当て問題” と呼称する) に着目し改善を図る．

3. 提案手法

2.2 章で述べたように, 本研究では密に物体が配置された画像における “割り当て問題” に着目する. 既存手法の問題点は, 図 5b のように大きさや縦横比が近い 2 つのクラスのオブジェクトがあった際に, 正しく割り当てを行うことができないことである. この問題を解決するために, あらかじめ決められている anchor box の集合 (anchor set と呼称する) を複製し, クラス数と同じだけの anchor set を用意する. そして各 anchor set は単一のクラスのみに割り当てを行う. これによって大きさや縦横比が近いオブジェクトであっても適切に割り当てが可能となる (図 6).

実装にあたり, ベースの手法としては SSD300 [14] を採用する. Liu らは Caffe [11] の実装を公開しているが, 本研究では拡張性の観点から Chainer [21] による再現実装を行い, 全ての実験で Chainer 実装を利用した. 図 7a に示すように SSD300 では出力の大きさは $(8732, 9)$ となっている. これは 8732 個の anchor についてそれぞれクラス分類 (5 次元 $(= \#class + 1)$, 1 次元は背景を示す) および矩形の推定 (4 次元) を出力するというものである. クラス分類は Softmax 関数により確率が計算されるようになっている. 提案手法では出力を $(4, 8732, 5)$ に変更する (図 7b). これは 4 つ $(= \#class)$ の anchor set についてそれぞれ 8732 個の anchor が含まれており, 各 anchor が物体らしさ (1 次元) および矩形の推定 (4 次元) を出力するというものである. 物体らしさは Sigmoid 関数により正規化を行う.

Loss 関数についても SSD を元に一部変更したものを用いる. まず各 anchor に対して ground truth の割り当てを求める. c ($c \in [1, 4]$) 番目の anchor set 中の a ($a \in [1, 8732]$) 番目の anchor に対する, ミニバッチ中の m ($m \in [1, M]$, M はミニバッチサイズ) 番目のサンプルの割り当て $s^m_{c,a}$ およびその重なり $j^m_{c,a}$ は次のように定義される.

$$s^m_{c,a} = \underset{g \in [1, G^m], c = t^m_g}{\operatorname{argmax}} \; Jaccard\left(B_a, B^m_g\right)$$

$$j^m_{c,a} = Jaccard\left(B_a, B^m_{s^m_{c,a}}\right)$$

ここで G^m は m 番目のサンプルに含まれる ground truth の個数であり, t^m_g および B^m_g は m 番目のサンプルの g 番目の ground truth のクラスおよび外接矩形である. また B_a は a 番目の anchor の外接矩形である.

Loss 関数 $L(z)$ は外接矩形の推定の項 $L^m_{loc}(z)$ と物体らしさの判定の項 $L^m_{conf}(z)$ の和として次のように定義される.

$$L(z) = \frac{1}{\sum_{m \in [1,M]} \left|A^m_{pos}\right|} \sum_{m \in [1,M]} \left(L^m_{loc}(z) + L^m_{conf}(z)\right)$$

z はネットワークの出力であり, $(4, 8732, 5)$ の形を持つ. ここで A^m_{pos} は m 番目のサンプルについてオブジェクトが割り当てられた anchor の添字の集合であり, 次式で定義される.

$$A^m_{pos} = \left\{(c, a) \middle| j^m_{c,a} \geq 0.5\right\}$$

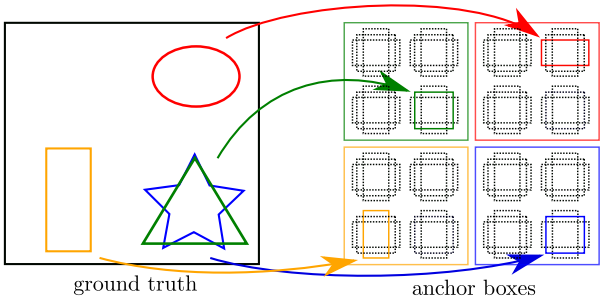


図 6: Fork された anchor を用いた場合の割り当て. それぞれの anchor set が各クラスに対応する.

$L^m_{loc}(z)$ は次のように定義される.

$$L^m_{loc}(z) = \sum_{(c,a) \in A^m_{pos}} huber\left(z(c, a, 2 \ldots 5), \log \Delta B\right)$$

$$\Delta B = B^m_{s^m_{c,a}} - B_a$$

また $L^m_{conf}(z)$ は次のように定義される.

$$L^m_{conf}(z) = \sum_{(c,a) \in A^m_{pos} \cup A^m_{neg}} l^m_{c,a}(z)$$

$$l^m_{c,a}(z) = sigmoid\_cross\_entropy\left(z(c, a, 1), j^m_{c,a} \geq 0.5\right)$$

ここで A^m_{neg} は hard negative の集合であり, オブジェクトに割り当てられていない anchor のうち $l^m_{c,a}(z)$ が大きい上位 $k|A^m_{pos}|$ 個を選ぶことで得られる (hard negative mining).

4. データセット

2.1 章で述べたように, 画像データには松井らが作成した学術漫画データセット Manga109 [15] を, 学習およびテスト用のアノテーションデータとして藤本らの研究 [6] を利用する. 松井らは漫画画像を片側 1 ページずつの状態で公開しているが, 本研究では左右のページを結合した状態を 1 ページとして扱う. これは藤本らが漫画中に含まれる見開きページ (左右のコマにシーンや物体がまたがって置かれているページの対, 図 8) に対応するために, アノテーションを付与する際に画像を 2 枚ごとに左右に結合して扱ったことに倣っている.

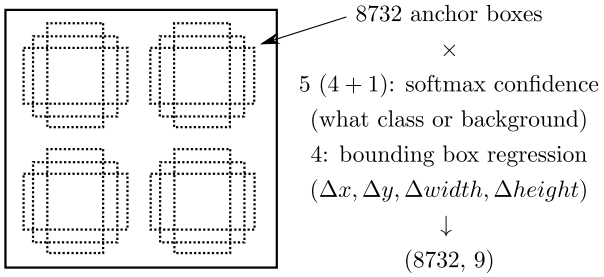
実験にあたって Manga109 のうちアルファベット順で先頭 99 冊を学習データ, 末尾 10 冊をテストデータとした (表 1). 実験にあたって, 1 つもアノテーションがついていないページを除外した (図 9). これは藤本らのアノテーションは表紙や扉絵などには付与されておらず, これらのページが含まれていると, 実際には物体が含まれているにも関らず物体がないと見なされてしまうからである. ここでページとは見開き 1 ページを指す.

5. 実験

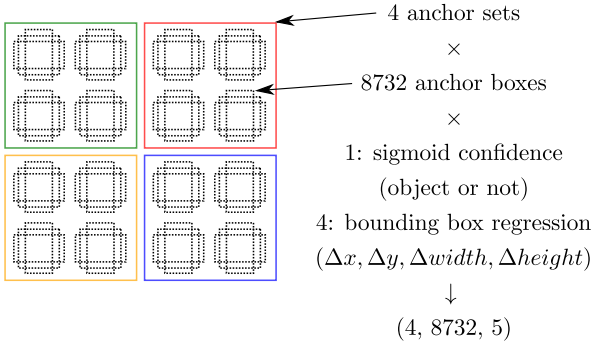
SSD300, SSD300-x4, SSD300-fork4 (提案手法) の 3 つの手法で物体検出を行った. SSD300-x4 は “割り当

表 1: Manga109 データセット

	冊数	ページ数	コマの数	テキストの数	顔の数	全身の数
学習	99	8,893	90,524	128,731	98,993	140,425
テスト	10	832	8,594	11,586	9,498	11,874
合計	109	9,725	99,118	140,317	108,491	152,299



(a) 元の手法



(b) Fork した場合

図 7: 出力されるテンソルの形. 提案手法 (b) では 4 つ (= *#class*) の anchor set を利用している.. 元の手法 (a) ではクラスを推定するために softmax 関数を利用しているが, 提案手法では sigmoid 関数を利用して各クラス独立に物体らしさを推定する.

て問題” を避けるためにクラスごとに独立したモデルを 4 つ学習するという手法である.

- SSD300: SSD300 [14] を 4 クラスの検出器として学習した. ネットワークの初期重みとしては Liu らと同様に ImageNet [3] で学習された VGG-16 [20] を利用した. 学習率などについても Liu らと同様のものを用いた.
- SSD300-x4: 1 クラスを検出する SSD300 を 4 本学習した. 学習やテストの時間, パラメータ数は通常の 4 倍となる. 基本的な設定は SSD300 と同様だが, 実験時間の短縮のために, 60,000 iteration まで学習した SSD300 の重みを初期解とした. 最終層については該当クラスのチャンネルおよび背景のチャンネルの 2 つのみを抽出した. 学習は 20000 iteration (全体での 80,000 iteration に相当) および 40,000 iteration (全体での 100,000 iteration に相当) で学習率を変化させ, 60,000 iteration (全体での 120,000 iteration に相当) 学習を行った.
- SSD300-fork4: 3 章で述べた手法. SSD300-x4 と同様に 60000 iteration まで学習した SSD300 を初期解として用いた. 最終層については該当クラスのチャンネルと背景のチャンネルの差を初期重みとして利用した. 出力層のチャンネル数が SSD300 と比べて増えているが, 学習時間の増加は 1%程度であった.

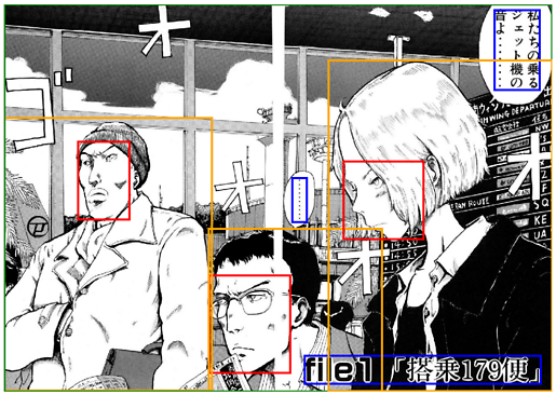


図 8: 見開きページの例 (犯罪交渉人 峰岸英太郎 ©記伊 孝). 左右のページを繋げることで一つのシーンを表現している. また中央のキャラクターは左右のページにまたがって描かれている.

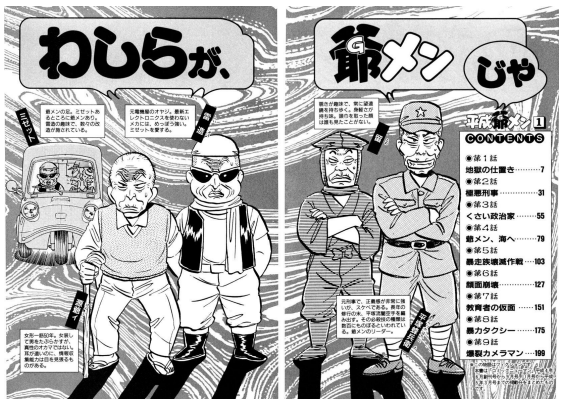


図 9: アノテーションがついていない目次ページの例 (平成爺メン ©やまだ 浩一). キャラクターやテキストが描かれているが, これらにアノテーションはついていない.

表 2: 結果

手法	mAP	コマ	テキスト	顔	全身
SSD300	79.4	96.9	81.6	61.6	77.3
SSD300-x4	84.8	97.8	87.2	74.1	80.1
SSD300-fork4	81.1	97.3	81.5	67.3	78.4

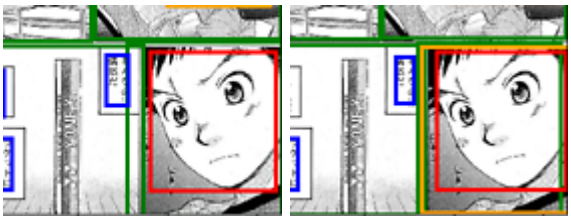
(a) IoU=0.5

手法	mAP	コマ	テキスト	顔	全身
SSD300	57.1	90.7	57.9	31.2	48.7
SSD300-x4	65.1	93.1	67.5	43.2	56.5
SSD300-fork4	59.9	91.9	57.9	36.8	52.9

(b) IoU=0.7

手法	mAP	コマ	テキスト	顔	全身
SSD300	16.9	57.1	7.2	0.4	2.7
SSD300-x4	24.2	72.4	15.4	1.1	8.0
SSD300-fork4	19.1	64.1	7.2	0.5	4.6

(c) IoU=0.9



やまとの羽根 ©咲 香里

雪の降る街 ©山田 雨月

図 10: SSD300 (左) と SSD300-fork4 (右) の比較. SSD300 では “コマ” (緑) と “全身” (橙) が大きく重なっているとき, “全身” が検出できていない. 一方, SSD300-fork4 では両方が正しく検出されている.

5.1. 手法ごとの比較

各クラスごとの Average Precision (AP) とその平均 (mAP) を表 2 に示す. しきい値としては物体検出において一般に用いられる IoU=0.5 (表 2a) に加え, IoU=0.7 (表 2b) および IoU=0.9 (表 2c) を採用した. これらは Zitnick ら [23] が高精度の領域検出の評価のために用いたものである.

表 2 からわかるように全てのしきい値および全てのクラスにおいて, SSD300-x4 が最も高い性能を示している. “テキスト” 以外の 3 クラスおよび mAP については SSD300-fork4 がそれに続いている. またしきい値が増加するに従って mAP の差は広がる傾向にある. クラスごとの傾向は手法には影響を受けず, しきい値が低いときには “コマ”, “全身”, “テキスト”, “顔” の順である. しきい値が増加するにつれて各クラスの AP は下がっていくが, 特に “顔” および “全身” のスコアの減少が顕著である. これは “コマ” や “テキスト” に比べて “顔” や “全身” は境界があいまいであることに起因すると考えられる.

本手法の目的である “割り当て問題” の解決については, 図 10 に示すようにいくつかの例で改善が確認できた. これらの例では “コマ” と “全身” が大きく重なっており, SSD300 では “コマ” しか検出できていないが SSD300-fork4 では両方とも検出できている.

5.2. 漫画ごとの比較

本研究で用いたテストデータセットにはさまざまな種類の漫画が含まれる. 漫画ごとの性能を比較するためにそれぞれの漫画における AP を表 3 に示す. 手法は SSD300-fork4 を利用し AP は IoU=0.5 で計算した.

表 3 で [1] となっている漫画は同じ作者の作品が学習データセットにも含まれているものである. したがってこれらの漫画については画風などを学習することが可能であり, それ以外の漫画はモデルにとって新規の作風ということになる. しかし表 3 からわかるように同じ作者の漫画で学習されたかどうかと検出性能との明確な関係は得られなかった. これは作者レベルでロバストなモデルが学習できたことを示している.

どの漫画でも “コマ” の性能は非常に高いものとなっている. 特に “幼稚園ぼうえい組” においては 100%の検出精度を達成している. この漫画は 4 コマ漫画と呼ばれるジャンルであり, 同じ大きさの矩形のコマが縦方向に 4 つ並んだ構造をしている (図 11). そのためテストデータセットの中でも特に “コマ” 検出が容易であったと考えられる.

10 冊の中で最も “全身” の検出精度が高い漫画は “やまとの羽根” である. この漫画はスポーツ漫画であり, 大きなコマに全身が含まれているような構図が多用されている (図 12a). このことが “全身” の検出精度に寄与していると考えられる. 一方, 最も “全身” の検出精度が低い “花影戦記 妖魔降臨” では多数の群集が小さく描かれたコマがいくつも存在する. たとえば図 12b の左上や左下のコマでは多くの人物が一コマに描かれているが, 1 つも検出できていない.

“顔” の検出精度は全体として高くないが, 特に “アンバランス・トーキョー” の検出精度は低いものとなっている. 一般的に漫画では顔を実際の比率より大きく描くが, この漫画では顔の領域をあまり大きくとっておらず多くの場合で検出に失敗している. たとえば図 13 のページでは “顔” が一つも検出できていない (アノテーションではそれぞれ 24 個および 13 個の “顔” が含まれている).

表 3: 漫画ごとの結果

漫画	ページ数	ジャンル	mAP	コマ	テキスト	顔	全身
うるとら☆イレブン	108	スポーツ	82.2	95.8	85.0	73.8	74.0
アンバランス・トーキョー	79	SF	76.4	98.8	82.5	51.2	72.9
ワレワレハ、オニデアル	89	ラブコメ	79.8	94.5	78.9	64.4	81.3
やまとの羽根	106	スポーツ	89.4	98.9	81.2	86.2	91.1
やさしい悪魔	85	ファンタジー	85.4	98.3	91.6	68.6	82.9
幼稚園ぼうえい組	14	4 コマ	75.5	100	66.6	58.4	76.9
花影戦記 妖魔降臨 [1]	99	ファンタジー	80.8	99.2	90.3	64.7	68.9
雪の降る街 [1]	83	恋愛	74.2	96.8	67.7	53.5	78.9
ゆめのかよいじ [1]	87	ファンタジー	82.1	95.0	72.9	77.4	83.2
ゆめ色クッキング	82	恋愛	85.2	97.5	84.2	71.5	87.4
合計	832	-	81.1	97.3	81.5	67.3	78.4

[1] 同じ作者の作品が学習データに含まれる

図 11: 4 コマ漫画におけるコマ検出の例 (幼稚園ぼうえい組 ©テンヤ). すべてのコマが正確に検出されている.

(a) やまとの羽根 ©咲 香里

(b) 花影戦記 妖魔降臨 ©島崎 譲, 鷹 司

図 12: “全身” の検出結果の例

6. まとめ

本稿では漫画画像中の物体検出の手法について提案した. 本手法は自然画像における物体検出手法を元に, 物体の重なりが大きいという漫画特有の問題 (“割り当て問題”) に対処するための工夫を加えたものである. 大規模な漫画画像データセットを用いた実験により, 既存手法を mAP で 1.5%, クラスによっては 6%改善したことを示した. また検出領域を正解とするしきい値が高くなるとより効果が大きくなることがわかった.

漫画ごとの解析では同じ作者の漫画を学習データに含んでいなくても性能が劣化しないことがわかった. また漫画のジャンルによって検出精度が影響を受けるこ

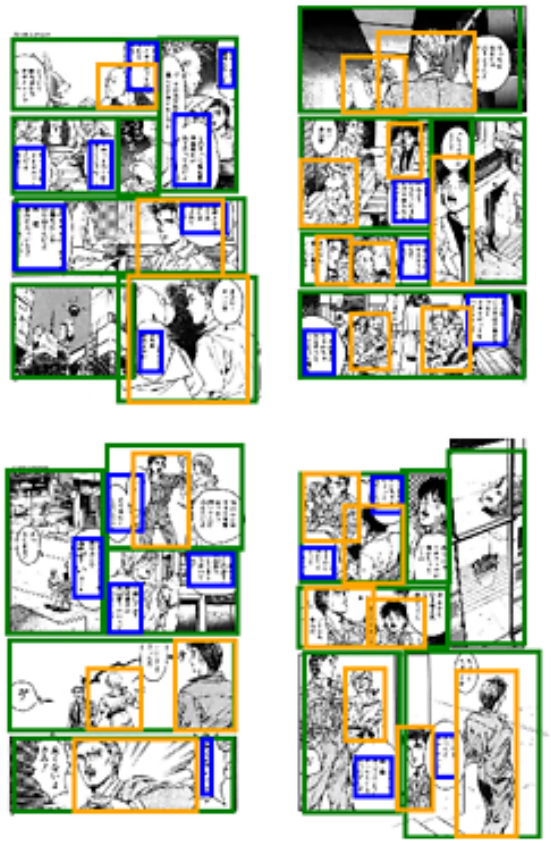


図 13: 顔検出に失敗した例 (アンバランス・トーキョー ©内田 美奈子). これらのページにはそれぞれ 24 個および 13 個の “顔” のアノテーションがついているが, 一つも検出できていない (顔の検出結果は赤い矩形で表示している).

と, およびそれらの多くは物体の大きさによるということがわかった.

より精度を高める工夫として, 小さい物体の認識精度を高めることが挙げられる. これは単純に高解像度画像を入力とする以外にもコマなどの小領域に分割した上で再度検出を行うなどの手法が考えられる.

7. 謝辞

本研究は, 戦略的情報通信研究開発推進事業 (SCOPE) の委託を受けた.

参考文献

[1] Wei-Ta Chu and Ying-Chieh Chao. Line-based drawing style description for manga classification. In *Proceedings of the 22nd ACM International Conference on Multimedia*, pp. 781–784. ACM, 2014.

[2] Wei-Ta Chu and Wei-Chung Cheng. Manga-specific features and latent style model for manga style analysis. In *International Conference on Acoustics, Speech and Signal Processing*, pp. 1332–1336. IEEE, 2016.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 248–255. IEEE, 2009.

[4] Mark. Everingham, S. M. Ali. Eslami, Luc. Van Gool, Christopher. K. I. Williams, John. Winn, and Andrew. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, Vol. 111, No. 1, pp. 98–136, January 2015.

[5] Cheng-Yang Fu, Wei Liu, Ananth Ranga, Ambrish Tyagi, and Alexander C Berg. Dssd: Deconvolutional single shot detector. *arXiv preprint arXiv:1701.06659*, 2017.

[6] Azuma Fujimoto, Toru Ogawa, Kazuyoshi Yamamoto, Yusuke Matsui, Toshihiko Yamasaki, and Kiyoharu Aizawa. Manga109 dataset and creation of metadata. In *Proceedings of the 1st International Workshop on coMics ANalysis, Processing and Understanding*, p. 2. ACM, 2016.

[7] Ross Girshick. Fast r-cnn. In *International Conference on Computer Vision*, pp. 1440–1448. IEEE, 2015.

[8] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Conference on Computer Vision and Pattern Recognition*, pp. 580–587. IEEE, 2014.

[9] Clément Guérin, Christophe Rigaud, Antoine Mercier, Farid Ammar-Boudjelal, Karell Bertet, Alain Bouju, Jean-Christophe Burie, Georges Louis, Jean-Marc Ogier, and Arnaud Revel. ebdtheque: a representative database of comics. In *Proceedings of the 12th International Conference on Document Analysis and Recognition*, pp. 1145–1149. IEEE, 2013.

[10] Mohit Iyyer, Varun Manjunatha, Anupam Guha, Yogarshi Vyas, Jordan Boyd-Graber, Hal Daumé III, and Larry Davis. The amazing mysteries of the gutter: Drawing inferences between panels in comic book narratives. In *Conference on Computer Vision and Pattern Recognition*. IEEE, 2017.

[11] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[12] Thanh-Nam Le, Muhammad Muzzamil Luqman, Jean-Christophe Burie, and Jean-Marc Ogier. Content-based comic retrieval using multilayer graph representation and frequent graph mining. In *the 13th International Conference on Document Analysis and Recognition*, pp. 761–765. IEEE, 2015.

[13] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755. Springer, 2014.

[14] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*, pp. 21–37. Springer, 2016.

[15] Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using manga109 dataset. *Multimedia Tools and Applications*, pp. 1–28, 2016.

[16] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Conference on Computer Vision and Pattern Recognition*, pp. 779–788. IEEE, 2016.

[17] Joseph Redmon and Ali Farhadi. Yolo9000: Better, faster, stronger. In *Conference on Computer Vision and Pattern Recognition*. IEEE, 2017.

[18] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pp. 91–99, 2015.

[19] Christophe Rigaud, Nam Le Thanh, J-C Burie, J-M Ogier, Motoi Iwata, Eiki Imazu, and Koichi Kise. Speech balloon and speaker association for comics and manga understanding. In *Proceedings of the 13th International Conference on Document Analysis and Recognition*, pp. 351–355. IEEE, 2015.

[20] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations*, 2015.

[21] Seiya Tokui, Kenta Oono, Shohei Hido, and Justin Clayton. Chainer: a next-generation open source framework for deep learning. In *Proceedings of Workshop on Machine Learning Systems in The Twenty-ninth Annual Conference on Neural Information Processing Systems*, 2015.

[22] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, Vol. 104, No. 2, pp. 154–171, 2013.

[23] C Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *European Conference on Computer Vision*, pp. 391–405. Springer, 2014.