修 士 学 位 論 文

漫画画像における
吹き出しの話者推定

平成 30 年 1 月 30 日

東京大学大学院　情報理工学系研究科
電子情報学専攻
48-166452　山本和慶

指導教員　相澤 清晴 教授

概要

漫画の意味理解の基礎的な要素として, 漫画画像中の吹き出しとその吹き出しの話者キャラクターの紐付けがある. 本研究は, 漫画のページ画像中の吹き出しに対して, その話者となるキャラクターを推定することを目的とした. 本研究では, 深層学習を用いて学習ベースの手法で話者を推定する手法を提案した. 学習ベースの手法を用いるためには, 学習用のデータセットが必要となる. 本研究では吹き出しのデータセットとして, 漫画画像データセットに吹き出し領域の位置, 大きさ, 話者キャラクターの情報をそれぞれ付与したデータセットを構築した.

話者推定の手法として, 吹き出しとキャラクターのペアに対して漫画画像とメタデータから抽出した特徴量を入力として深層学習によりスコアを計算し, 最もスコアが高いキャラクターを話者として選択する手法を提案した. 評価実験により, 提案手法が従来手法と比べて高い精度での吹き出しの話者推定が可能であることを明らかにした. また, 提案手法に用いた特徴量について, 各特徴量の有効性についても検証を行った.

目次

表目次

図目次

第 1 章

序論

1.1 背景

漫画は日本を代表する文化の 1 つである．近年, スマートデバイスの普及に伴い電子コミックの市場が急激に拡大している．電子コミックの普及に伴い, 漫画の検索システム [17, 24], 漫画の翻訳, 漫画のモバイル端末向け表示システム [4] などの重要性が高まっている．これらのタスクにおいては, 漫画の内容理解に基づく高次の情報が有用となるようなケースもある．

漫画の内容理解タスクの 1 つとして漫画中のセリフとそのセリフの話者のキャラクターの紐付けがある．漫画中のセリフとキャラクターの紐付けにおいては, 漫画画像中のキャラクター・吹き出し・フレーム等の情報が非常に重要な手がかりとなる．R-CNN を用いた吹き出し検出 [34], 木構造条件付確率場モデルによるフレーム検出 [19], 特徴抽出による漫画の顔画像検出と認識 [32] など, 漫画画像中の物体検出に関する研究はさかんに行われている．そして, 吹き出しとキャラクターの紐付け [8] など, メタデータに基づいた吹き出しの話者推定の研究もおこなわれている．

1.2 目的

本研究では, メタデータを用いて学習ベースの手法で吹き出しの話者推定の手法を提案する．吹き出しの話者推定は, 漫画画像からのキャラクター, 吹き出しなどの物体検出と, それらのメタデータに基づいて話者推定を行う二段階の処理からなるが, 本研究は後者の部分を中心的に取り扱う．

図 1.1: 日本の漫画の一例．表紙などの一部のページを除いて白黒のものが主流である.ⓒ赤松健

1.3　課題

漫画画像中の吹き出しの話者推定における課題について述べる．日本の一般的な漫画の一例を図 1.1 に示す．漫画画像は自然画像とは異なる特徴を持つため, 漫画画像特有の課題が存在する.

漫画において, キャラクターの発言は図 1.2 のように吹き出しの内側にテキストを書き込むことによって表現される. このような形でイラストと文字を共存させるのは漫画に固有の表現である. そのため, 吹き出しの話者推定というタスクも漫画に固有のものであり, 一般画像において同様のタスクが存在しないため, 一般画像において同様のタスクを取り扱った先行研究も存在しない.

吹き出しの話者推定の先行研究としては, 距離情報に基づく吹き出しの話者推定 [8] がある. この研究では吹き出しとの距離が最も近いキャラクターを話者として選択するという手法が提案されている. しかし, この手法では距離情報だけを利用しているため, 吹き出し

から最も近いキャラクターが話者である場合にしか正しく話者を推定することができず, 対応できる範囲が極めて限定的であるという問題がある.

1.4　提案

先行研究には, 距離情報しか利用しておらず対応できる範囲が限定的であるという課題がある. この課題に対し、本研究ではディープニューラルネットワークを用いて学習ベースの手法を用いて話者推定を行う手法を提案する.

学習ベースの手法を用いるにあたり, まず吹き出し領域とその話者キャラクター情報のデータセットを構築した. Manga109 データセット [9] に対して, 漫画画像中の吹き出しそれぞれに位置と大きさを示す外接矩形および話者キャラクターのアノテーション ID を付与したデータセットを構築した. このデータセットを用いることにより, 学習ベースの手法によるアプローチが可能となった.

続いて, 構築したデータセットを用いて学習ベースの手法を用いて話者推定を行う手法を提案する. 漫画画像から切り出した吹き出しの画像と, 吹き出しとキャラクター領域に付与されているメタデータから抽出したいくつかの特徴量を入力として用いる. これらの画像とデータを入力としてディープニューラルネットワークを用いて吹き出しの話者推定を行う. これにより, 既存手法では話者推定が困難なものに対しても対応が可能となる.

1.5　本論文の構成

本論文の校正を以下に示す.

1. 第 1 章 序論
2. 第 2 章 関連研究
3. 第 3 章 提案手法
4. 第 4 章 実験
5. 第 5 章 結論

第 1 章では漫画画像に対する画像処理の現状と, 本研究の目的および課題について述べた. 第 2 章では, 漫画画像中の吹き出しの話者推定に関する研究と, 一般的な物体認識の手法について述べる. 第 3 章では, 漫画画像とそれに付属するメタデータを基に吹き出しの話者推定を行う学習ベースの手法を提案する. 第 4 章では, 提案手法の性能評価および考察を行う. 第 5 章では, 本研究のまとめと今後の課題について述べる.

図 1.2: 漫画に含まれる吹き出しの例．なめらかな曲線を用いたものやギザギザの直線を用いたものなど、多様なデザインが存在する．©八神健，赤松健

第 2 章

関連研究

本章では「一般画像認識の研究」,「漫画における画像処理の研究」について述べる

2.1 一般画像認識

画像認識の主要なタスクは入力画像をいずれかのクラスに分類する Classification, 入力画像から物体候補領域を抽出し, 多くの場合は分類も同時に行う Detection, ピクセル単位での領域分割を行う Semantic Segmantation などが存在する．このうち, 本研究の提案手法との関連が深い Classification に関する研究および, 前処理の漫画画像物体検出に深く関わる Detection に関する研究について詳しく述べる．

Classification や Detection の分野においては, 領域ベースマッチングと特徴ベースマッチングがかつて主流であった．領域ベースマッチングはテンプレート画像を少しずつずらしながら, テンプレート画像と入力画像の類似度を計算し, しきい値以上の類似度となるものを抽出することで認識および検出を行う手法である．類似度の計算にはテンプレート画像と入力画像の各画素の差の絶対値の和や, テンプレート画像と入力画像の各画素の差の二乗和などが用いられた．この手法は, テンプレート画像と入力画像を画素単位でマッチングする手法であるため, ノイズに弱く, 既知の画像にしか対応できない, 画像のスケールの変化にも弱いという問題がある．特徴ベースマッチングは, テンプレート画像と入力画像からそれぞれ特徴点の抽出を行い, 特徴点の周囲の局所領域から特徴量を計算し, 特徴量に基づいて類似度の計算を行う手法である．特徴点の抽出手法としては, Maximally stable extremal regions (MSER) [23], Difference of Gaussian region (DoG) [14] などの手法があり, 局所領域からの特徴量抽出手法としては, Scale Invariant Feature Transform (SIFT) [22] や SIFT を改良した Principal Component Analysis-SIFT [7],

Affine-SIFT [25] などの手法が用いられる．これらの手法は領域ベースマッチングと比べてノイズに強く, 画像のスケールの変化に対しても頑強であるという利点がある．一方, 特徴点抽出がうまくいかなかった部分は対応付けが行われないため, マッチングの結果が疎なものになってしまう可能性があるという問題がある．

領域ベースマッチングと特徴ベースマッチングのどちらも未知の画像には対応できないという問題があるが, 近年ではより汎用性の高い手法として学習ベースの手法が数多く提案されている．特にディープニューラルネットワークを用いた手法は高い精度が報告されており, ディープニューラルネットワークを用いた ImageNet の分類 [2] を皮切りにディープニューラルネットワークを用いた手法が多く提案されている．ニューラルネットワークは, 人間の脳の神経回路の機構を模したネットワーク構造である．このニューラルネットを何層も重ねたディープニューラルネットワークを用いた手法が画像認識の分野においては主流となっている．画像認識においては全結合層以外に畳み込み層とプーリング層を用いた Convolutional Neural NetworkNetwork (CNN) [18] が有効であることが知られている．CNN は, 畳み込み層によって広い範囲の情報を畳み込むことで, 空間的な情報を保持することが出来ることが特長である．畳み込み後にプーリング層でダウンサンプリングを行うことにより, 特徴を検出する働きをする．Network in network [20]，VGG [30]，GoogLeNet [31] など，多くの CNN のモデルが提案されている．本項では, 画像分類のタスクにおいて特に高い精度が報告されており, 本研究の提案手法にも用いた Deep Residual Network (ResNet) [15] について 2.1.1 で詳しく述べる．

Detection の分野にもディープニューラルネットワークを用いる試みは行われており, Regions with CNN features (R-CNN) [11] をはじめとしたディープニューラルネットワークを用いた手法が多数提案されている．ディープニューラルネットワークを用いた物体検出では, まず何らかの手法で物体候補領域を生成し, 各物体領域から特徴抽出を行い物体かそうでないかの判定およびクラス分類を行うという手順で処理を行う手法が主流となっている．R-CNN による物体検出の手順を図 2.1 に示す．まず Objectness [3], SelectiveSearch [33] などの手法を用いて候補領域を生成し, それぞれの候補領域から CNN を用いて特徴抽出を行う．そして抽出した特徴量を用いてカテゴリ分類を行うという手順で物体検出を行う．全ての候補領域に対してそれぞれ CNN で特徴抽出を行うという手順を踏むため, 処理に非常に時間がかかるという欠点がある．この欠点を改善した手法として Fast R-CNN [10] や Faster R-CNN [28] がある．Faster R-CNN は, Region Proposal Network という, 物体候補領域を生成するネットワークを提案している．物体候補領域の生成の部分にもディープニューラルネットワークを用いることにより, 高精度で少数の候補を生成することができるようになり, R-CNN や Fast R-CNN と比べて高速に

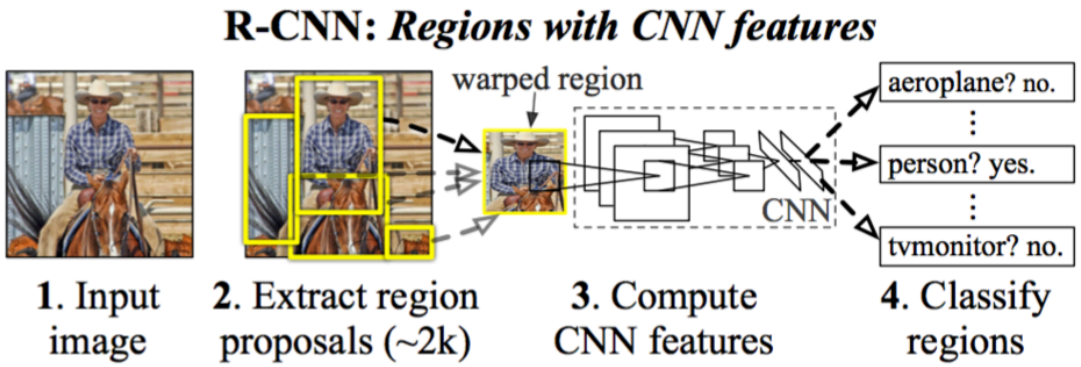


図 2.1: R-CNN を用いた物体検出の手順. R-CNN を用いた物体検出より [11].

なり精度も向上している. Faster R-CNN と同程度の精度でさらに高速化した手法として You only look once (YOLO) [27] や, Single shot multibox detector (SSD) [21] がある. 特に高速で, 話者推定の前段処理の吹き出し, キャラクター, フレームの検出に用いる手法として有力な SSD について 2.1.2 で詳しく述べる.

2.1.1　Deep residual network

ディープニューラルネットワークにおいては, 層を深くすることによってネットワークの表現力が高まると共により高度な特徴抽出が可能になると考えられる. しかし, 実際には単純に層を深くするだけでは学習がうまく進まず, 精度が向上しないことが知られている. ResNet では, 学習に残差関数を用いることでこの問題を解決することを試みている. ResNet のネットワーク構造を図 2.2 に示す. 図 2.3 のように残差ブロックと呼ばれる, 通常の畳み込み層と短絡結合を組み合わせた構造を導入していることがこのネットワークの特徴である. 残差ブロックを用いることにより, その層における変換が不要な場合に畳み込み層の重みを 0 にすることで短絡結合を用いて入力をそのまま出力とし次の層に流すことができる. これによって層が深くなっても効率的に学習を進めることができる.

2.1.2　Single shot multibox detector

SSD [21] は Faster R-CNN [28] と同程度の精度で Faster R-CNN や YOLO [27] よりも高速な検出手法である. SSD のネットワーク構造を図 2.4 に示す. Faster R-CNN と同様, 候補領域の生成をニューラルネットワークを用いて行うため, 高精度で少数の候補領域の生成が可能となり, さらに End-to-End での学習が可能となるという利点がある. Faster

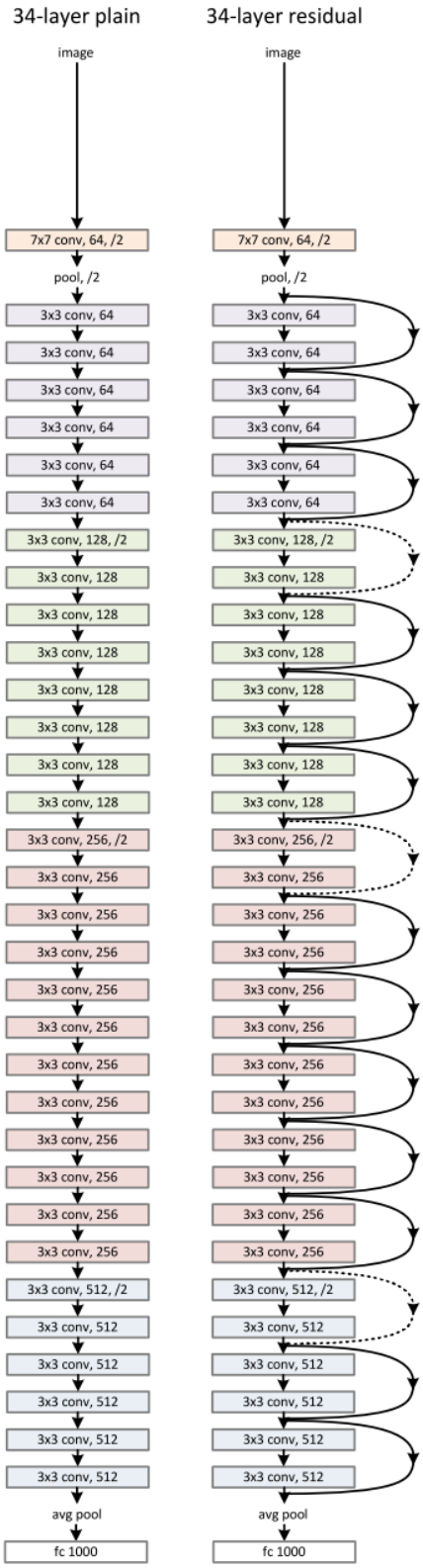


図 2.2: ResNet のネットワーク構造．左は一般的な 34 層の CNN の図, 右は 34 層の ResNet の図である．Deep residual network を用いた物体認識より [15].

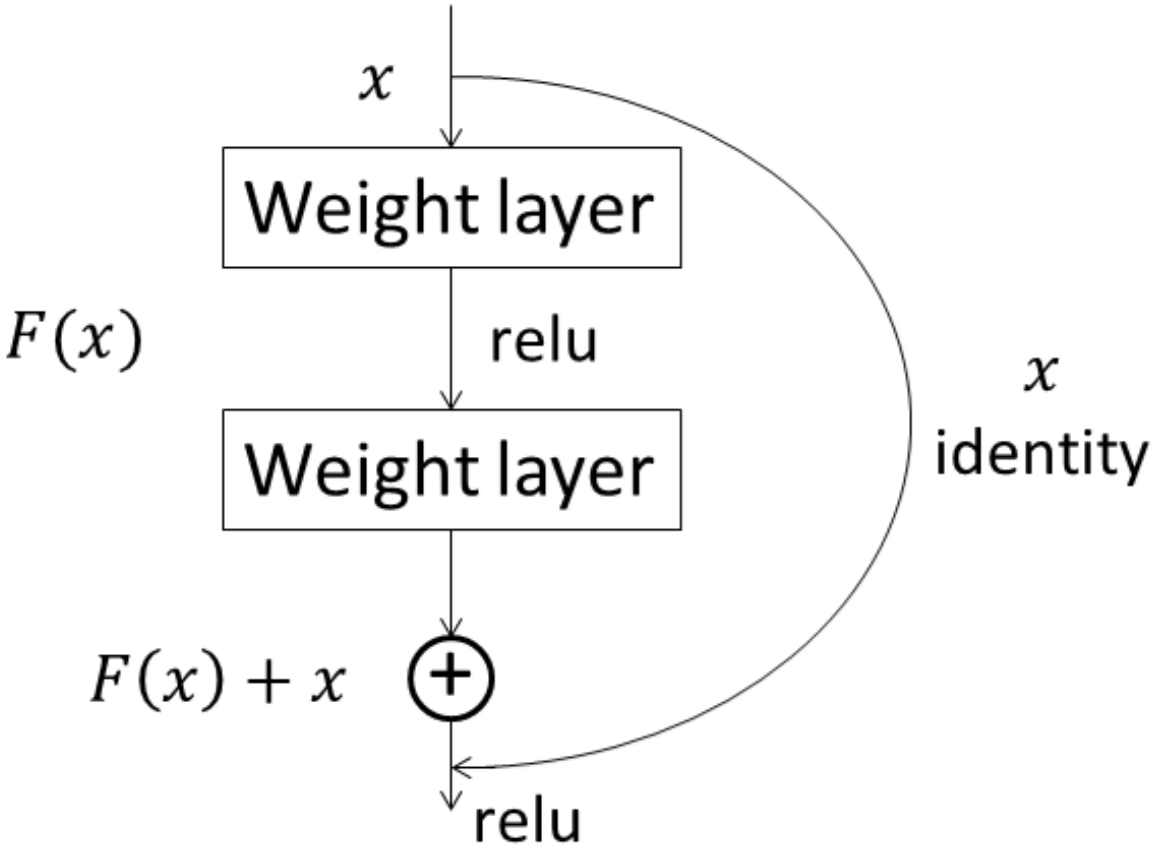


図 2.3: ResNet に用いられる残差ブロック Deep residual network を用いた物体認識より [15].

R-CNN では候補領域から再度特徴抽出を行うという処理が必要であったが, SSD では先に複数のスケールでの特徴分布を抽出しておき, その特徴分布を使ってクラス分類を行うという手順を取っている．そのため, Faster R-CNN と比べて高速での処理が可能となっている．また, 複数スケールでの特徴分布の抽出を行い, さらに様々なアスペクト比での出力を行うため高精度での物体検出が可能となっている．

2.2　漫画における画像処理

漫画における画像処理の研究で特に本研究と関連の深いものとして以下のようなものがある．

1. 漫画における物体検出
   漫画においても一般画像認識同様, Classification や Detection に関する研究が行わ

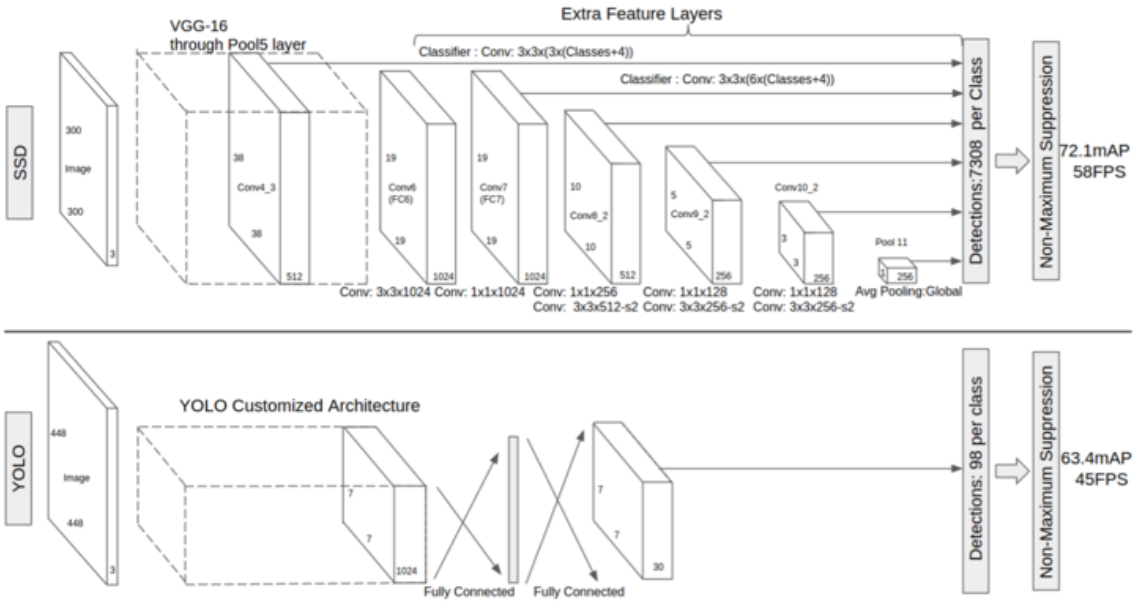


図 2.4: SSD のネットワーク構造．下の図は YOLO のネットワーク構造．Single Shot MultiBox Detector より [21].

れてきた．Detection については本研究で扱うタスクである話者推定の前処理として非常に関連の深いタスクであるため,2.2.1 で詳しく述べる．

2. 意味理解と検索

   漫画の意味理解やメタ情報を用いる試みとして, セリフの読み順推定 [16] やスケッチによる漫画検索 [24] などがある．また, 本研究のテーマとしている吹き出しの話者推定も漫画の意味理解の一種とみなすことができる．吹き出しの話者推定の研究としては, 吹き出しとキャラクターの距離の情報を利用した手法 [8] などがある．本研究と同じタスクを取り扱う先行研究であるため, 2.2.2 で詳しく述べる．

3. 漫画画像データセット

   画像処理の研究において, 画像データセットはかかせない存在である．漫画そのものは数多く存在するものの, 漫画画像には著作権が存在するため学術利用には著作者の許諾を得る必要がある．かつては作者から個別に許諾を得た上で漫画画像を研究に用いる必要があったが, 近年ではこの問題をクリアしたデータセットが公開されている．いくつかの条件の下, 作者から個別的に許可を得る必要なく利用できるデータセットとして, 図 2.5 に示すようなフランス, ベルギー, アメリカの漫画画像を集めた eBDtheque データセット [12] や, 日本の漫画画像を集めた Manga109 データ

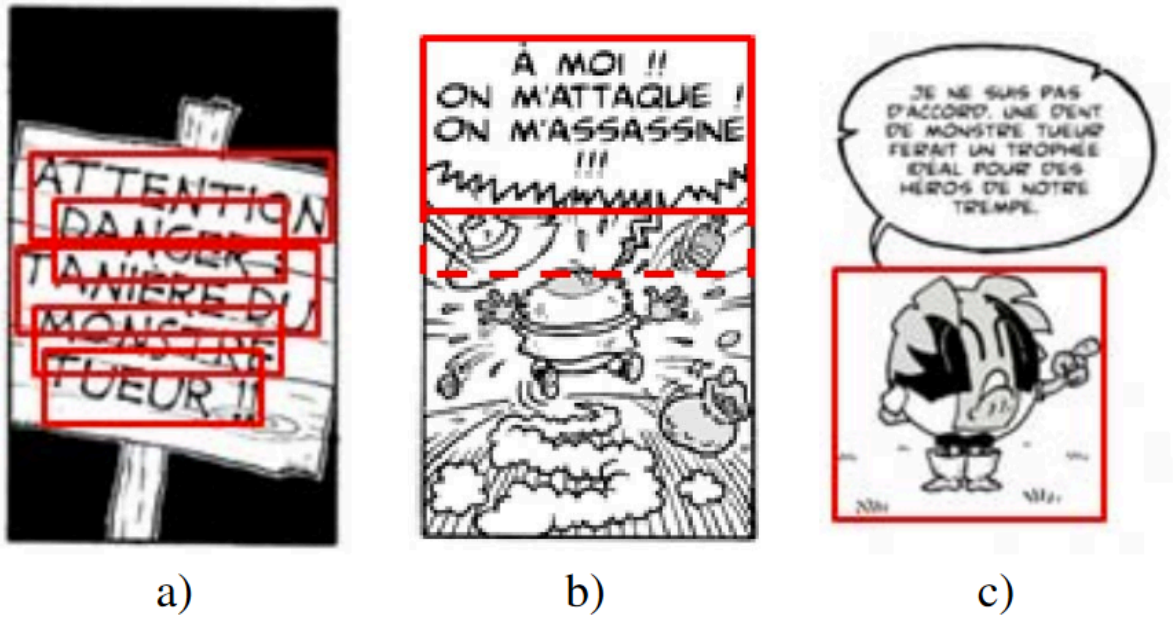


図 2.5: eBDtheque に含まれる漫画とアノテーションの一例. (a) 文字領域のアノテーション (b) 吹き出し領域のアノテーション (c) キャラクター領域のアノテーション eBDtheque より [12].

セット [9] がある. 特に大規模なデータセットである Manga109 データセットについて 2.2.3 で詳しく述べる.

2.2.1　漫画画像の物体検出

漫画画像を対象とした Detection においては, 特徴抽出による漫画の顔画像検出と認識 [32], Deformable Part Model を用いた漫画画像中の顔の検出 [36], 直線成分の組み合わせによるフレーム検出 [35], 木構造条件付確率場モデルによるフレーム検出 [19], リアルタイムでの電子漫画からの文字抽出 [5], 連結成分を用いた漫画の文字検出 [6], 動的輪郭モデルを用いた輪郭検出 [29] などの研究がなされている. これらの研究では, 顔, 吹き出し, フレームなどの特定のオブジェクトに着目し, そのオブジェクトに固有の特性を利用して Detection を行う手法となっている. 一方で, 近年では, R-CNN を用いた吹き出し検出 [34], Faster R-CNN を用いたマンガ画像からのメタデータ抽出 [13] や, 漫画物体検出に向けた検出器の並列化— [26] のように, Faster R-CNN や SSD などのディープニューラルネットによる物体検出手法を用いた手法も提案されている. Faster R-CNN を用いた Detection の結果の一例を図 2.6 に示す. これらのディープニューラルネットワークを用

いた手法は各オブジェクトの固有の特性を考慮する必要がなく, どのオブジェクトにおいても高い精度での物体検出が可能となっている.

2.2.2　吹き出しの話者推定

漫画の吹き出しとキャラクターの対応付けと意味理解 [8] では, 吹き出しとキャラクターそれぞれにアンカーポイントを設定し, そのアンカーポイント間の距離が最も近い吹き出しとキャラクターを対応付けるという手法が提案されている. また, アンカーポイントとしては, 吹き出しの外接矩形の中心点, 吹き出しの重心, 吹き出しのテールの位置が提案されている. この 3 つのうち, 後に述べたものほど精度が高くなるが, 吹き出しの重心を求めるにはピクセル単位でのアノテーションが必要となり, 吹き出しのテールの位置情報を用いるためにはテールの位置情報のアノテーションが必要となる. この手法では, 吹き出しとキャラクターの距離だけに着目しているため, 話者が吹き出しから最も近いキャラクターである場合しか話者を正しく推定することができないという問題がある.

2.2.3　漫画のデータセット

eBDtheque データセット [12] には, 漫画の作者と直接交渉して許諾を得る必要がないという利点があるが, 漫画画像が全部で 100 ページ分しかなく, 学習ベースの手法の学習データなどに用いるにはデータ数が十分でないという問題がある. この問題を解決したデータセットとして Manga109 データセット [9] がある. Manga109 データセットは 109 冊の漫画で構成される漫画画像のデータセットである. 20000 ページ以上のデータからなり, 学習データに用いることも十分可能な大規模データセットとなっている. また 109 冊の全ての漫画にフレーム, キャラクター (全身) , キャラクター (顔) , 文字の領域の大きさと位置の情報と, 文字領域に含まれる文字列がアノテーションとして付与されている.

このデータセットに含まれる漫画は 109 冊全てが日本のプロの漫画家によって描かれたものであるため, 品高品質なものとなっている. 1970 年代から 2010 年代という幅広い時代に公開されたものを用いており, ジャンルも幅広い. このデータセットの基本情報を表 2.1 に示す. 収録されている漫画のほとんどはマンガ図書館 Z（旧絶版漫画図書館）[1] にて公開されている. 漫画およびアノテーションデータの一例を図 2.7 および図 2.8 に示す. Manga109 に含まれる漫画は，そのコマの加工も含めてあらかじめ作者からの利用許諾を受けており, コピーライトを明記する等のいくつかの条件の下, 以下の目的での使用が認められている.

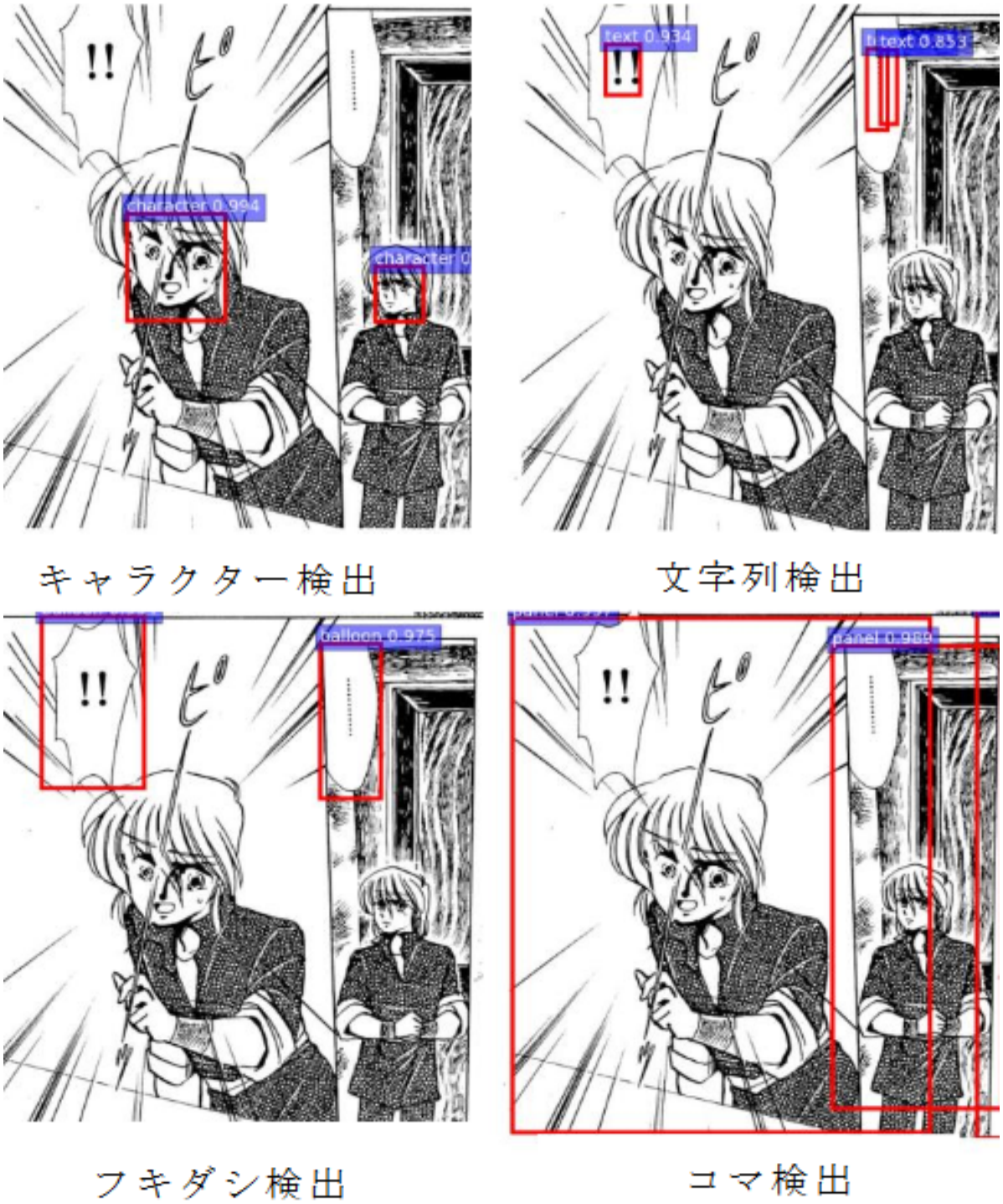


図 2.6: Faster R-CNN を用いた Detection 結果の一例 Faster R-CNN を用いたマンガ画像からのメタデータ抽出より [13]. ©島田ひろかず

表 2.1: Manga109 の基本情報．Manga109 データセットとそのメタデータの構築より [9]

巻	109 巻
ページ	21,142 ページ
作品	104 作品
著者	94 人
年代	1970 年代–2010 年代
出版社	エニックス，学習研究社，小学館，少年画報社，徳間書店，朝日ソノラマ，東京三世社，白泉社，秋田書店，竹書房，芳文社，角川書店，講談社，集英社
ジャンル	4 コマ，SF, ギャグ，サスペンス，スポーツ，バトル，ファンタジー，ホラー，ラブコメ，恋愛，時代物

- データセットを利用した実験.
- データセットの漫画の学術論文への掲載利用.
- 学会等のデジタルライブラリーへの学術論文の収録.
- 学術成果を示すデモビデオ等のデジタル媒体での利用.

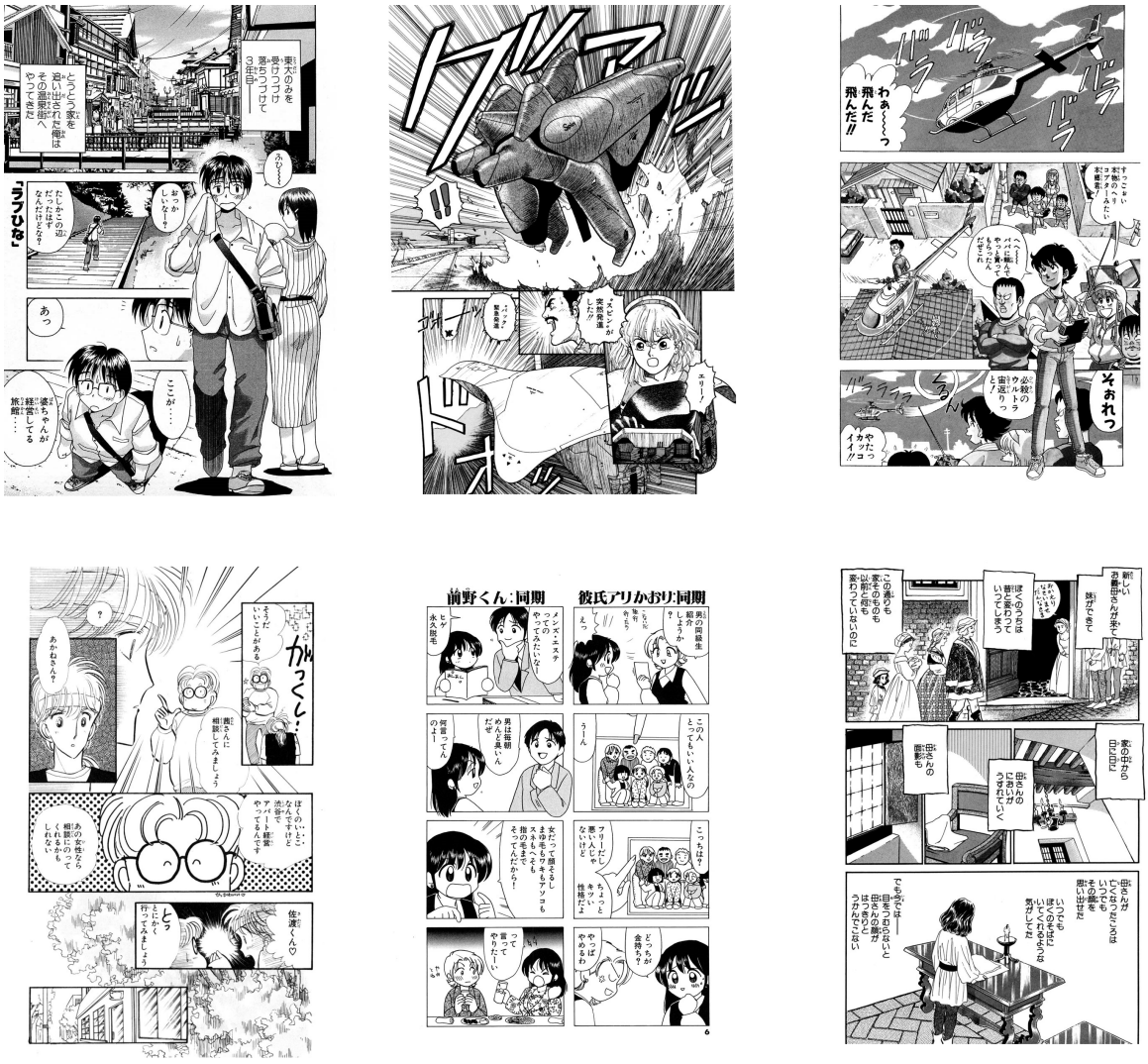


図 2.7: Manga109 に含まれる漫画の例．幅広い年代, ジャンルの作品が網羅されている．
ⓒ赤松健, 加藤雅基, 新沢基栄, 愛田真夕美, さんりようこ, 里中満智子

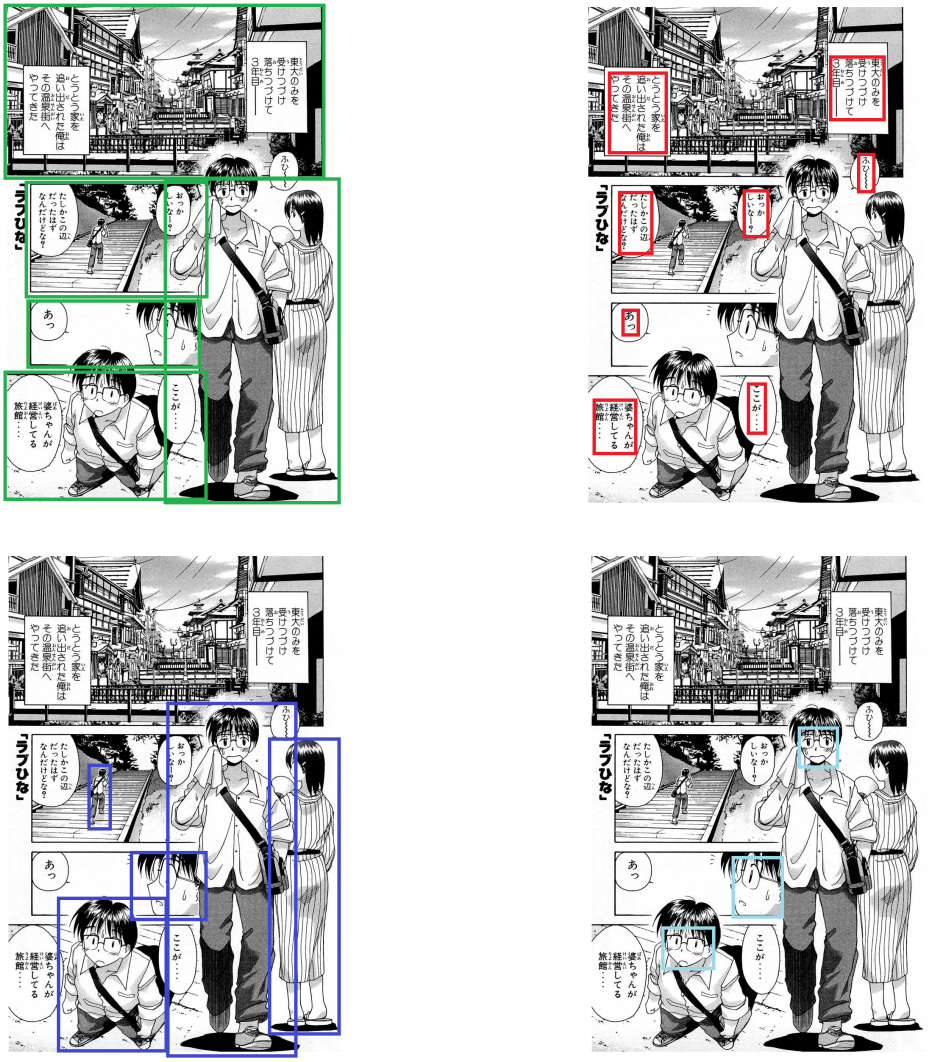


図 2.8: アノテーションデータ付きの例. 左上:フレーム, 右上:文字 , 左下:キャラクター (全身) , 右下:キャラクター (顔) . ©赤松健

第 3 章

提案手法

3.1 概要

吹き出しの話者推定のためにはまずキャラクターや吹き出しなどの物体検出が必要となる．2.2.1 で述べたディープラーニングを用いた手法は, 高い精度での物体検出が可能となることが報告されている. そのため, 本研究では吹き出し, キャラクター, フレームなどの物体が適切に検出されているという仮定の下, 吹き出しの話者推定を行った.

吹き出しとキャラクターの距離に基づいた話者推定 [8] では, 吹き出しとキャラクターの距離だけに着目しているため, 話者が吹き出しから最も近いキャラクターである場合しか話者を正しく推定することができないという問題がある. そこで, 本研究では吹き出しやキャラクターの大きさ, 位置関係などの情報も統合的に利用して話者推定を行う手法を提案する.

学習ベースの手法を用いるにあたっては, 学習用のデータセットが必要となる. そのため, 本研究ではまず話者推定のための吹き出しアノテーション付きデータセットを構築した. Manga109 データセット [9] に対して吹き出しの位置, 大きさとそれぞれの吹き出しに対応する話者の情報を持つアノテーションデータを付与した. 次に，吹き出しの話者推定の手法として, 話者推定は, 吹き出しとキャラクターのペア毎に, 吹き出しの画像とメタデータを入力としてスコア計算を行い, 最もスコアが高かったキャラクターをその吹き出しの話者とするという手法を提案する.

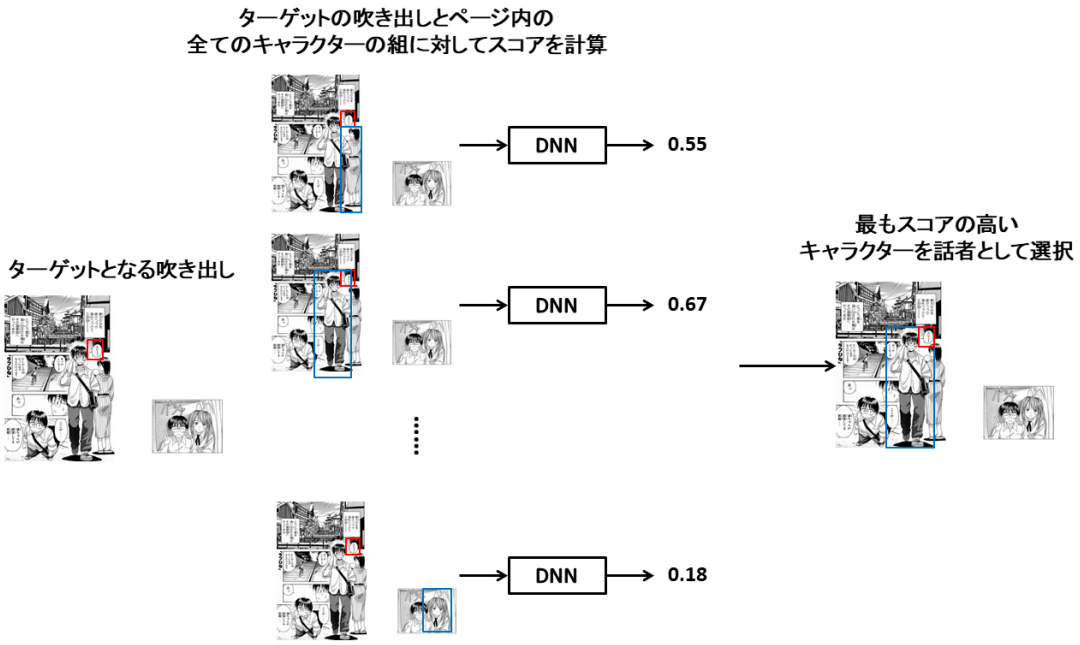


図 3.1: 提案手法の概要．見開きページ内の吹き出しとキャラクターの全ての組に対してスコアの計算を行い, 各吹き出しに対して最もスコアの高かったキャラクターと紐づける. ©赤松健

3.2　漫画のデータセットへの吹き出し領域のアノテーションの付与

本研究では学習ベースの手法を用いるため, 学習データとして漫画画像に加えてフレーム, キャラクター, 文字の位置情報などが付与されたアノテーションデータが必要となる. また, 漫画は時代や作風, ジャンルによりデザインに非常に大きな幅があるため, 様々な漫画を網羅した多様で大規模なデータセットが必要となる. これらの条件を満たすものとして，本研究では Manga109 データセット [9] を実験に用いる.

吹き出しの話者推定のタスクにおいては, フレーム, キャラクターのアノテーション以外に吹き出し領域の位置情報と吹き出しの話者の情報が必要となる. そこで, Manga109 の 109 冊の漫画のうち 15 冊を抽出し, これらの漫画に対してアノテーションを付与した. この際, 学習用データに用いた 5 冊の漫画は全てのページにアノテーションを付与し, 検証用データに用いた 5 冊およびテストデータに用いた 5 冊の漫画はそれぞれ最初の 40 ページにアノテーションを付与した. 付与した情報は, 吹き出しの位置および大きさを示す外接矩

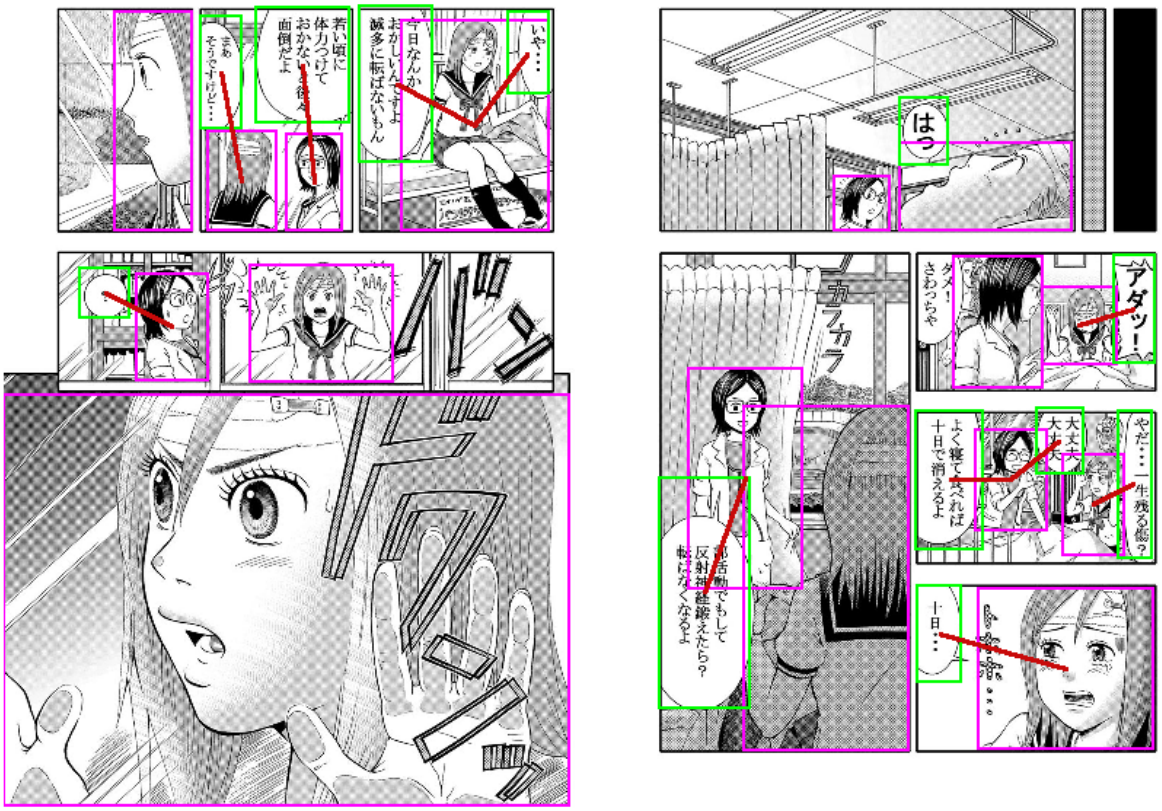


図 3.2: 吹き出しアノテーションの一例. 緑色の矩形が新たに追加した吹き出しのアノテーション. 吹き出し領域の位置と大きさの情報を持つ矩形と吹き出しの話者の情報からなる.ⓒ高波伸

形, 吹き出しの話者キャラクター (0 人の場合や 2 人以上の場合もある) , 吹き出しのテールの向き (右上, 右, 右下, 下, 左下, 左, 左上, 上, 向き無しの 9 通り) の 3 つである. 吹き出しのテールの向きの情報は,3.4.1 にて述べる Pre-training において用いるために付与した. 実験に用いた漫画の詳細な内訳は 4.1 にて述べる.

吹き出しアノテーションデータの一例を図 3.2 に示す. 話者は 1 人であるとは限らず, 0 人 (話者が描かれていない) である場合や 2 人以上の話者が存在する場合もある. また, ナレーションや心の声の表現などはアノテーションの対象とせず, 発話のみをアノテーションの対象とした. アノテーションを付けるかどうかの判定例を図 3.3 に示す. 右の図の例は吹き出しの無いもの, ナレーション, 心の声なのでいずれもアノテーション対象ではないと判定した.

(a) 吹き出しとみなしてアノテーションを付けるもの

(b) 吹き出しとみなさずアノテーションを付けないもの

図 3.3: 吹き出しにアノテーションを付けるかの判定例. ©赤松 健

3.3　メタデータの抽出

アノテーションデータはそのまま学習に用いるのではなく, 吹き出しとキャラクターのペア毎にいくつかの特徴量に抽出, 変換した上で入力データとする. 用いる特徴量は以下の 6 つである.

1. 吹き出しとキャラクターの距離
2. 吹き出しに対して話者候補の全てのキャラクターとの距離を順位付けした時, その吹き出しとキャラクターのペアは何番目に近いか
3. 吹き出しとキャラクターのペアが同じフレーム内に存在しているかどうか
4. キャラクター領域の面積
5. 吹き出し領域の面積
6. 吹き出しとキャラクターの位置関係

つまり, 最終的に入力として用いられるメタデータは 6 次元のベクトルとなる. それぞれの特徴量について順に解説する.

3.3.1　距離情報

吹き出しとキャラクターの距離は, 吹き出しの話者推定において重要な手がかりとなる. キャラクターと吹き出しの距離に基づく話者推定 [8] においてもその有効性が示されてい

表 3.1: 吹き出しの話者が吹き出しから最も近いキャラクターであるものの割合

吹き出しの話者が最も近いキャラクターであるもの	70.35%
吹き出しの話者が最も近いキャラクターであるもの	29.65%
吹き出しの話者をページ内からランダムに選択した場合の正答率	6.26%

る．学習データに用いたアノテーションデータにおいて, 吹き出しの話者がその吹き出しから最も近い位置にいるキャラクターであるものの割合を調査した結果を表 3.1 に示す．ページ内の全てのキャラクターの中からランダムに選択したキャラクターを話者と判定した場合の正答率が 6% 程度であるのに対して最も近いキャラクターを話者として選択するだけで 70% 以上の正答率が得られることから, 吹き出しと話者の距離は非常に重要な手がかりであることが分かる．また, 吹き出しとキャラクターの絶対的な距離だけでなく, 他の話者候補のキャラクターと比較した相対的な距離も重要な手がかりとなると考えられるため, 吹き出しに対して話者候補の全てのキャラクターとの距離を順位付けした時, その吹き出しとキャラクターのペアは何番目に近いかという情報も付与した．

吹き出しとキャラクターの距離を x_c, 見開きページ全体の左上の端から右下の端までの距離を X として $\frac{x_c}{X}$ で正規化したものを特徴量として用いた．距離の順位情報については, 1 位である (話者候補キャラクターの中で吹き出しに最も近い) 場合は 1, 最下位である (話者候補キャラクターの中で吹き出しに最も遠い) 場合は 0 とし, 間の順位は 0〜1 の範囲で等間隔となるように正規化する．例えば, ある吹き出しの話者候補となるキャラクターが 5 人いてその中で 2 番目に吹き出しに近い場合は 0.75, 3 番目に近い場合は 0.5 となる．

3.3.2　フレーム情報

4 の実験で学習データに用いたアノテーションデータにおいて, 吹き出しとその話者が同じフレーム内にいるものの割合を調査した結果を表 3.2 に示す．ここでは, 吹き出しおよびキャラクターのバウンディングボックスの中心点があるフレームの内側にある時, その吹き出しあるいはキャラクターはそのフレーム内にあるとみなしている．そのため, 1 つの吹き出しおよびキャラクターが複数のフレームに所属する可能性もある.90% 以上の吹き出しの話者は同じフレーム内に存在していることから, 吹き出しとキャラクターが同じフレーム内に存在しているかどうかは重要な手がかりとなると考えられる．特徴ベクトル化する際には吹き出しとキャラクターのペアが同じフレーム内に存在する場合は 1, 異なる場合は 0 としている．

表 3.2: 吹き出しとその話者が同じフレーム内にいるものの割合

吹き出しと話者が同じフレーム内に存在する	94.62%
吹き出しと話者が違うフレーム内に存在する	5.38%

3.3.3　面積と位置関係

極端に小さいキャラクターは吹き出しの話者になりにくいなどの特徴が考えられるため, 吹き出しとキャラクターのバウンディングボックスの面積も特徴量として用いた. 吹き出し領域のバウンディングボックスの面積を s_b, キャラクター領域のバウンディングボックスの面積を s_c, 見開きページ全体の面積を S としてそれぞれ $\frac{s_b}{S}$, $\frac{s_c}{S}$ で正規化したものを特徴量として用いた.

吹き出しとキャラクターの位置関係の情報は, 吹き出しとキャラクターの中心間を直線で結んだ時のその直線の向きを特徴量として用いた. 吹き出しから右方向に進む向きを基準として, 吹き出しとキャラクターの中心間を結んだ直線の角度を 0〜1 の範囲で正規化している. この特徴量は, 3.4 で述べる吹き出し画像から抽出した特徴量と組み合わせて利用することを想定した.

3.4　ニューラルネットワークを用いたスコア計算

吹き出し領域の画像を抽出したものと, 3.3 で述べたアノテーションデータから抽出した特徴量を入力として, 吹き出しとキャラクターの組ごとにスコア計算を行う. 計算に用いたニューラルネットワークの構造を図 3.4 に示す. 画像分類のタスクにおいて高い精度を誇る ResNet50 [15] をベースとして用いる. ResNet50 の最終層を取り除いたものに吹き出し画像を入力し, 抽出された特徴量にメタデータ特徴量を結合し, スコア計算を行う. ResNet50 の部分については吹き出し画像を用いて Pre-training を行う. 詳細は 3.4.1 にて述べる.

3.4.1　ニューラルネットワークの Pre-training

吹き出し画像からの特徴量の抽出に用いる ResNet50 は, ImageNet で Pre-training したモデルをさらに Pre-training する. ここでは, 吹き出し画像を入力とし, 吹き出しのテール (話者がいる方向を示す, 鋭角な三角形であることが多い) の向きを推定するタスクに

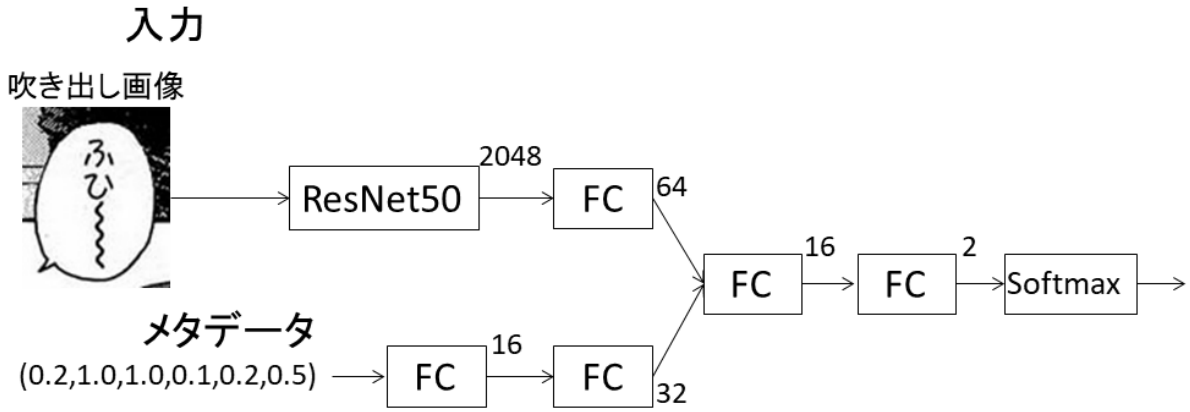


図 3.4: 本手法で用いるネットワーク．ResNet50 は最終層の全結合層を取り除いたものを用いる．

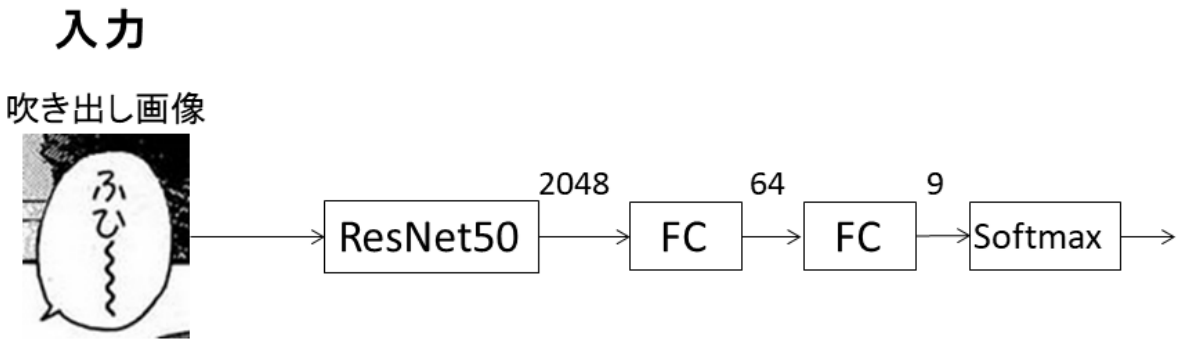


図 3.5: Pre-training で用いるネットワーク．ResNet50 は最終層の全結合層を取り除き, 新たに 2 層の全結合層を加えたものを用いる．

よって Pre-training を行う．ここで用いたネットワークを図 3.5 に示す．ResNet50 から最終層の全結合層を取り除き, 新たに 2 層の全結合層を付け加えたものを用いる．テールの向きの推定は右, 右下, 下, 左下, 左, 左上, 上, 右上の 8 方向もしくはいずれの向きでもない (テールが存在しない) のいずれかを選択する形で行う．つまり 9 クラス分類のタスクのとなる．こうして Pre-train したモデルから最終層を取り除いたモデルを学習に用いる．吹き出しのテールの向きを推定するタスクで Pre-training を行うことで, テールの向きに関係するような部分を中心に特徴抽出が行われると考えこのような設定での Pre-training を行った．

3.4.2　ニューラルネットワークのモデルと学習

3.4.1 にて Pre-training を行ったモデルから最終層を取り除き, 出力を 64 次元にしたものと, 6 次元のメタデータを 2 層の全結合層に入力して 32 次元にしたものを結合し, さらに 2 層の全結合層に入力してスコア計算を行う. 全結合層を通さない場合, 画像特徴量は 2048 次元, メタデータは 6 次元となるが, 次元数に差がありすぎると学習が収束しない可能性があるため, 全結合層を用いて次元数の調整を行った.

学習の際には, 吹き出しとキャラクターのペアが正しい組み合わせであるデータ, つまりキャラクターがその吹き出しの話者であるデータを正例, それ以外を負例とし, 2 値分類のタスクとしてトレーニングを行った.

3.5　スコアに基づく吹き出しの話者推定

各吹き出しに対して, 見開きページ内に存在する全てのキャラクターとペアを作る. 各ペアにおいてそれぞれメタデータから特徴量の抽出を行った上で 3.4.2 のネットワークを用いて, スコアの計算を行う. そして最もスコアが高くなったキャラクターをその吹き出しの話者として選択する. 話者推定の流れを図 3.6 に示す. 本手法ではしきい値処理などは行わず, 話者として選択するキャラクターは常に 1 人とした.

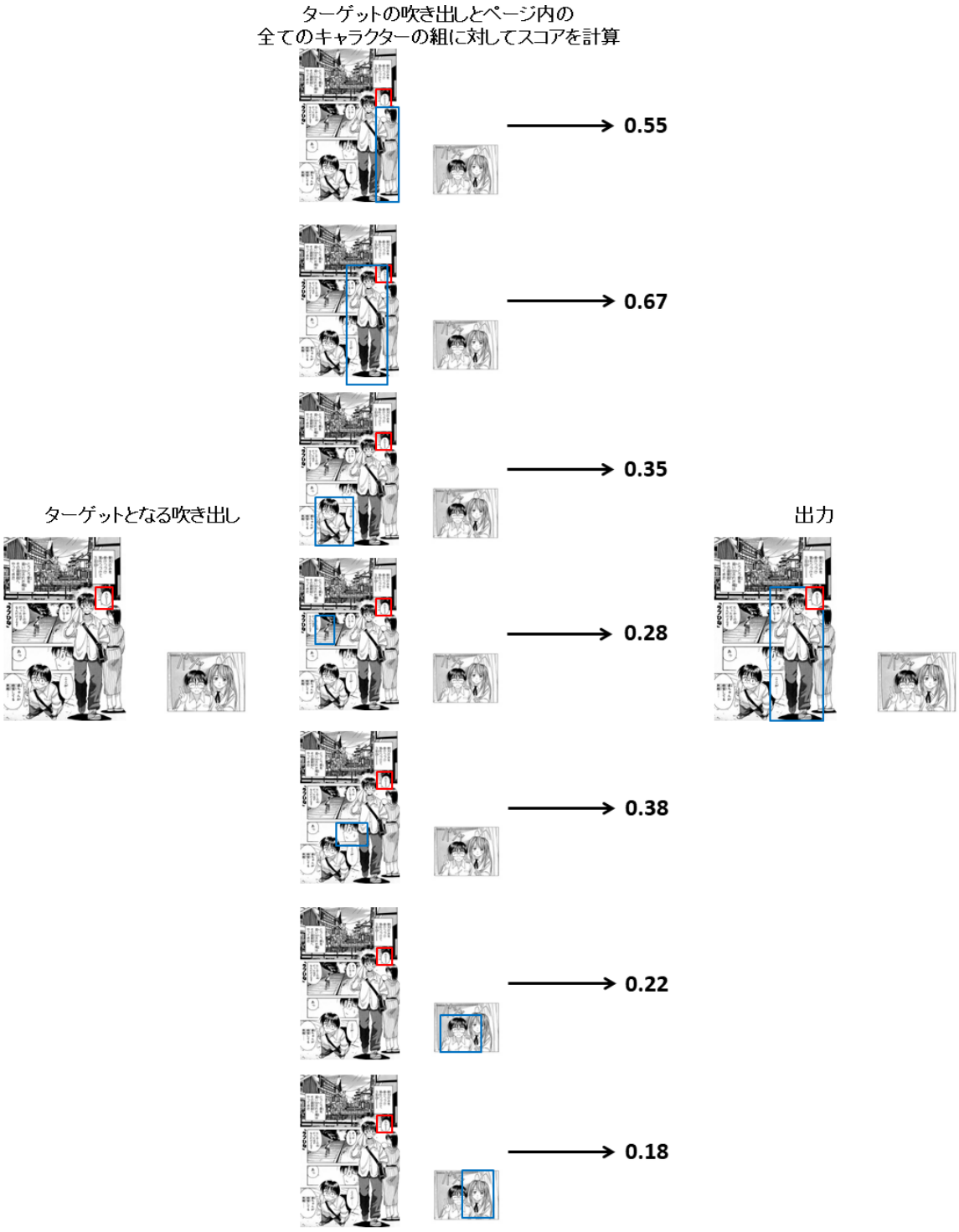


図 3.6: 話者推定の流れ．ターゲットとなる吹き出しと見開きページ内の全てのキャラクターの組に対してスコア計算を行い, 最もスコアの高かったキャラクターを話者として選択．©赤松健

第 4 章

実験

提案手法の有効性を確認するため，吹き出しの話者推定の精度を評価する実験を行った．

4.1 データセット

3.2 で述べた Manga109 データセット [9] に吹き出し領域のアノテーションを付与したものを実験に用いた．

表 4.1: 実験に用いた作品の一覧．

作品名	作者	年代	出版社	ジャンル	頁数	巻
ラブひな	赤松 健	1990 年代	講談社	ラブコメ	2192	1
青すぎる春	奥田 桃子	2000 年代	集英社	恋愛	210	
極限サイクロン	高波 伸	2000 年代	スクウェア・エニックス	スポーツ	115	
OL ランチ	さんり ようこ	2000 年代	竹書房	4 コマ	134	
デュアルジャスティス	竹山 佑右	2000 年代	幻冬舎	バトル	197	1
黒井戸眼科	平 雅巳	1990 年代	講談社	サスペンス	201	1
プラチナジャングル	篠原 正美	2000 年代	青葉出版	ミステリー	171	
とっておきの A・B・C	愛田 真夕美	1980 年代	白泉社	恋愛	188	1
征神記ヴァルナス	島崎 譲	2000 年代	講談社	バトル	200	2
ハイスクール奇面組	新沢 基栄	1980 年代	集英社	ギャグ	197	1
無敵冒険シャクマ	計奈 恵	1990 年代	エニックス	バトル	186	1
レヴァリアース	夜麻 みゆき	1990 年代	エニックス	ファンタジー	199	2
ラファエロ	里中 満智子	1990 年代	美術出版社	歴史	234	1
ARMS	加藤 雅基	1980 年代	東京三世	SF	162	
アンバランス・トーキョー	内田 美奈子	1990 年代	徳間書店	SF	176	

表 4.2: 実験に用いたデータの概要

	作品名	吹き出しの個数
学習データ	ラブひな, 青すぎる春, 極限サイクロン, OL ランチ, デュアルジャスティス	3136
検証用データ	黒井戸眼科, プラチナジャングル, とっておきの A・B・C, 征神記ヴァルナス, ハイスクール奇面組	737
テストデータ	無敵冒険シャクマ, レヴァリアース, ラファエロ, ARMS, アンバランス・トーキョー	809

実験に用いた作品を表 4.1 に示す．これらの作品のうち,5 冊の全てのページを学習データに,5 冊の 40 ページずつを検証用データに, 最後の 5 冊の 40 ページずつをテストデータに用いた. この際, 吹き出しの話者が 0 人のものは除外し, 話者が 1 人以上のデータのみを使用している. それぞれのデータの概要を表 4.2 に示す. また, 学習用データと検証用データは 3.4.1 で述べた Pre-training にも用いた.

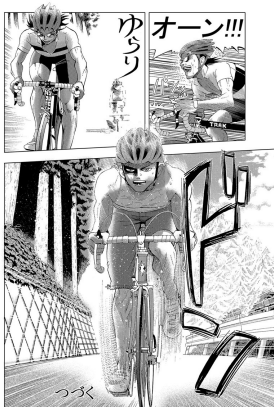
実験に用いた漫画の一部を図 4.1 および 4.2 に示す. それぞれ異なる作者の漫画であり幅広い作風となっている.

4.2　吹き出しのテールの向きの推定

まず, Pre-training に用いたタスクである吹き出しのテールの向きの推定の精度について評価を行った. 実験では, ResNet50 を学習用データの 5 冊の吹き出し画像を用いて学習を行い, 検証用データの 5 冊の漫画で最も高い精度となったネットワークを吹き出しの話者推定を行うネットワークの初期の重みに用いた. この検証用データで最も高い精度となったネットワークについて, テスト用データを用いてテールの向きの推定の精度の評価を行った (テスト用データはこの精度評価のみに使用しており, Pre-training および 3.4.2 で述べた学習に用いる Pre-training 済みモデルの選定には一切使用していない) . 吹き出しのテールの向きは 8 方向および向き無しの 9 通りであり, 出力結果と正解データが一致したものの割合 (Accuracy) を評価指標とした. 実験の結果を表 4.3 に示す.

(a) ラブひな

(b) 青すぎる春



(c) 極限サイクロン

(d) OL ランチ

(e) 黒井戸眼科

(f) プラチナジャングル

(g) 征神記ヴァルナス

(h) ハイスクール鬼面組

(i) 無敵冒険シャクマ

(j) ラファエル

(k) ARMS

(l) アンバランス・トーキョー

図 4.1: 実験に用いた漫画画像の一例. ©赤松 健, 奥田 洋子, 高波 伸, さんり ようこ, 平雅巳, 篠原 正美, 島崎 譲, 新沢 基栄, 計奈 恵, 里中 満智子, 加藤 雅基, 内田 美奈子

(a) ラブひな　(b) 青すぎる春　(c) 極限サイクロン　(d) OL ランチ

(e) 黒井戸眼科　(f) プラチナジャングル　(g) 征神記ヴァルナス　(h) ハイスクール鬼面組

(i) 無敵冒険シャクマ　(j) ラファエル　(k) ARMS　(l) アンバランス・トーキョー

図 4.2: 実験に用いたアノテーション付きのデータセットの一例. ©赤松 健, 奥田 洋子, 高波 伸, さんり ようこ, 平 雅巳, 篠原 正美, 島崎 譲, 新沢 基栄, 計奈 恵, 里中 満智子, 加藤 雅基, 内田 美奈子

表 4.3: テールの向きの推定精度評価. 常に“どの向きでもない (テールが存在しない) ”を選択し続ける場合と比較を行った.

手法	Accuracy
Choose no direction	33.7%
ResNet50	37.0%

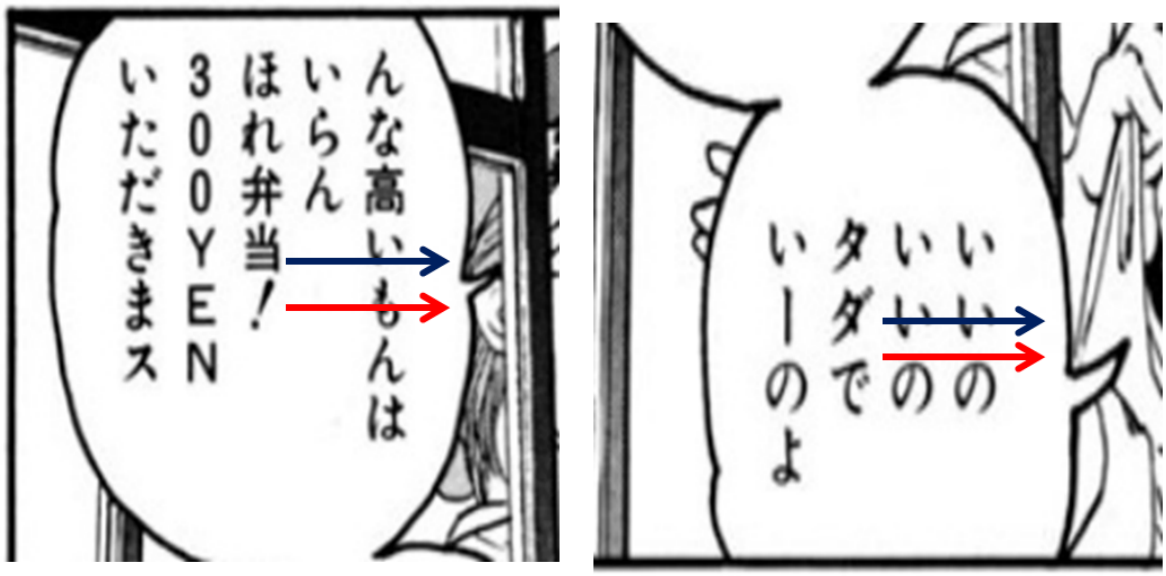


図 4.3: 吹き出し画像のテールの向きの推定に成功した例. 赤い矢印が正解, 青い矢印が出力結果. ©内田 美奈子

常に“どの向きでもない (テールが存在しない) ”を選択し続ける場合と比べて高い精度とはなっているものの, 十分な精度で吹き出しの向きが推定できているとは言えない. 吹き出しのテールの向き推定が成功例を図 4.3 に, 失敗例を図 4.4 に示す. 吹き出しのテール以外に起伏の少ない吹き出しにおいては話者推定に成功することが多い一方, 失敗例のように凹凸の激しい吹き出しにおいてはテールの向きの推定が困難であることが分かった.

図 4.4: 吹き出し画像のテールの向きの推定に失敗した例．赤い矢印が正解, 青い矢印が出力結果．ⓒ計奈 恵

4.3　吹き出しの話者推定

4.3.1　比較手法

比較する手法として，2 関連研究で述べた，吹き出しとキャラクターの距離に基づき話者を推定 [8], この手法に同じフレーム内に存在するキャラクターを優先的に選択する制約を加えた手法, 提案手法から画像特徴量を除いてメタデータのみを使用する手法を用いた．また, 文字情報を考慮しない場合の人手での話者推定とも比較を行った．

距離ベースの手法

比較手法として用いる吹き出しとキャラクターの距離に基づき話者を推定 [8] は, 吹き出し領域とキャラクター領域のアノテーションの中心間の距離に基づき, 距離が最も近いものを話者として選択するという手法である．この手法をベースに,「吹き出しと同じフレーム内に 1 人以上のキャラクターがいる場合はその中で最も距離が近いキャラクターを話者として選択し, 同じフレーム内にキャラクターが 1 人もいない場合は見開きページ全体の中から最も距離が近いキャラクターを選択する」という制約を加えたものを比較手法に用いた．

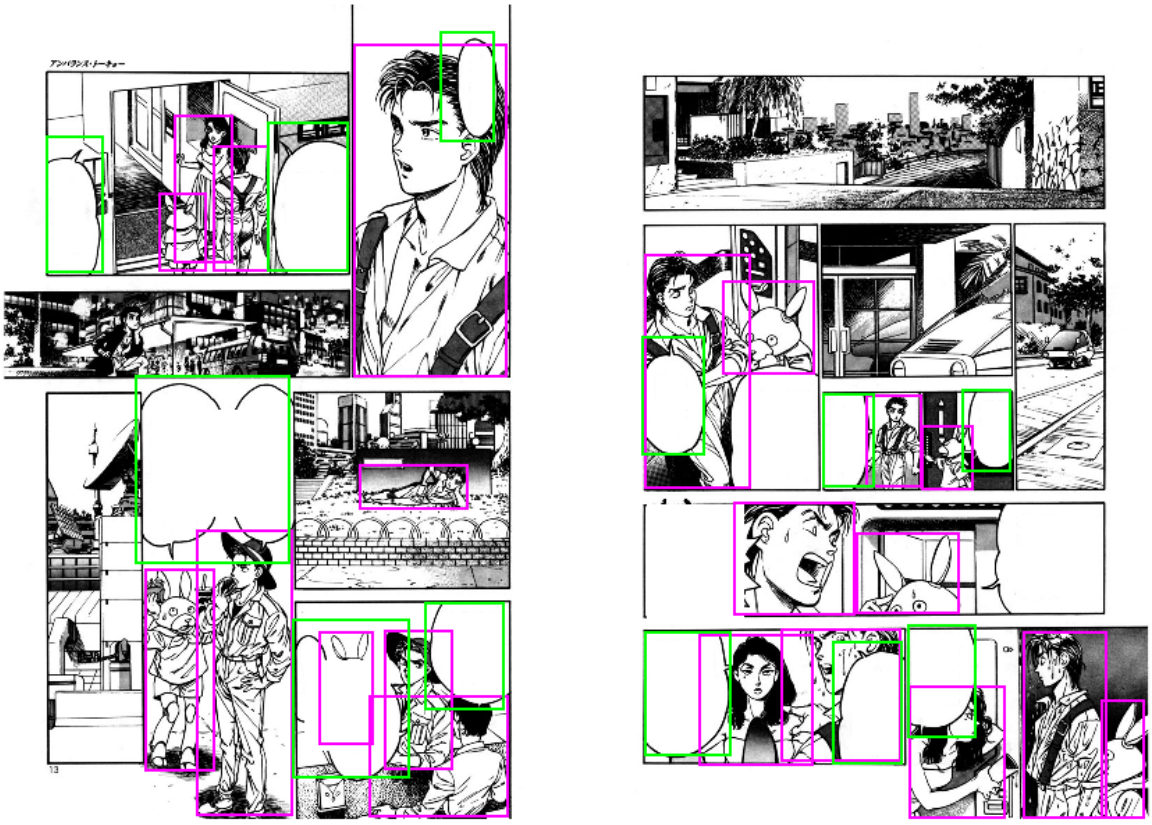


図 4.5: 人手での話者推定に用いる漫画画像 ©内田 美奈子

人手での話者推定

本研究で提案した手法では, 吹き出し内の文字の内容などは一切考慮していない. 吹き出し内の文字を考慮しない条件下での話者推定の精度の限界を調査するため, 文字を考慮しない条件下での人手での吹き出しの話者推定の精度とも比較を行った. 図 4.5 のように, 文字領域を白抜きにした漫画画像を用いて人手での話者推定を行った.

画像特徴量を除きメタデータだけを用いた手法

画像特徴量の有効性を検証するため, 入力から画像特徴量を取り除き, アノテーションデータから抽出した特徴量だけを使用する手法を比較手法に用いる. 画像特徴量を取り除くことに伴い, ネットワークの形を図 4.6 のように全結合層を 2 層だけ用いたものに変更した. 画像特徴量を用いた手法と比べて全結合層の数が少なくなっているのは, 画像特徴量の次元数に合わせて次元数を増やす必要がないためである.

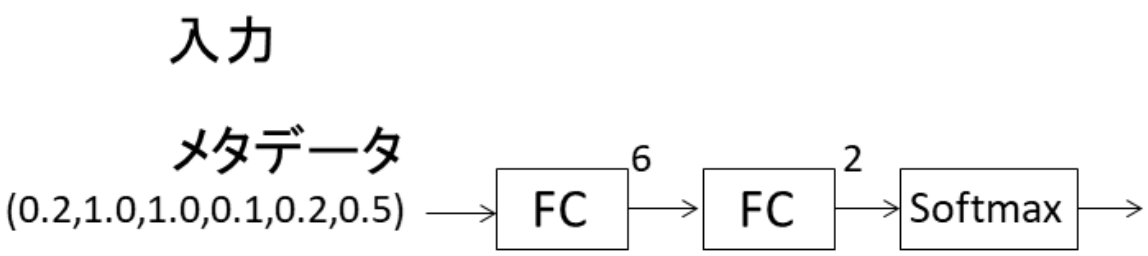


図 4.6: 画像特徴量を取り除き, メタデータのみでの話者推定に用いるネットワーク

4.3.2　評価手法

今回用いたデータセットにおいては, 吹き出しの話者が 2 人以上である場合もあるものの, ほとんどのデータにおいて話者は 1 人であるため, Top1-Accuracy を評価指標に用いる. Top1-Accuracy は, 上位 1 つの出力が正解データのいずれかと一致すれば正解, 不一致であれば不正解とし正解の割合をスコアとする評価手法である. 今回の提案手法においては, ネットワークが最も大きいスコアを出力したキャラクターが話者のいずれかであれば正解となる.

4.3.3　実験結果

提案手法および比較手法の話者推定の精度を表 4.4 に示す. ただし, 比較手法に用いられている Random+Frame は, 吹き出しと同じフレーム内に存在するキャラクターの中からランダムに話者を推定し, 同じフレーム内にキャラクターが存在しない場合は見開きページ内の全てのキャラクターからランダムに話者を推定する手法である.

学習ベースの手法を用いることで, 距離に基づいて話者を推定する手法と比べて話者推定の精度が向上することが分かった. 一方で, 今回提案した手法では吹き出し画像は話者推定の有効な手がかりとはならず, 吹き出し画像を用いることによる精度の向上は見られないという結果が得られた.

吹き出し画像とメタデータを用いる手法およびメタデータのみを用いる手法のいずれにおいても話者推定に成功した例を図 4.7 に示す. 距離に基づく手法では吹き出しから一番近いキャラが選択されてしまう例において, 提案手法では“面積の大きいキャラクターほど吹き出しの話者になりやすい”という傾向を考慮することができるため, 話者推定に成功していると考えられる. 逆に, 吹き出し画像とメタデータを用いる手法およびメタデータ

表 4.4: 話者推定の Top1-Accuracy での精度評価. Distance は 4.3.1 で述べた距離ベースの手法, Metadata only は 4.3.1 で述べたメタデータだけを用いた深層学習で予測する手法である.

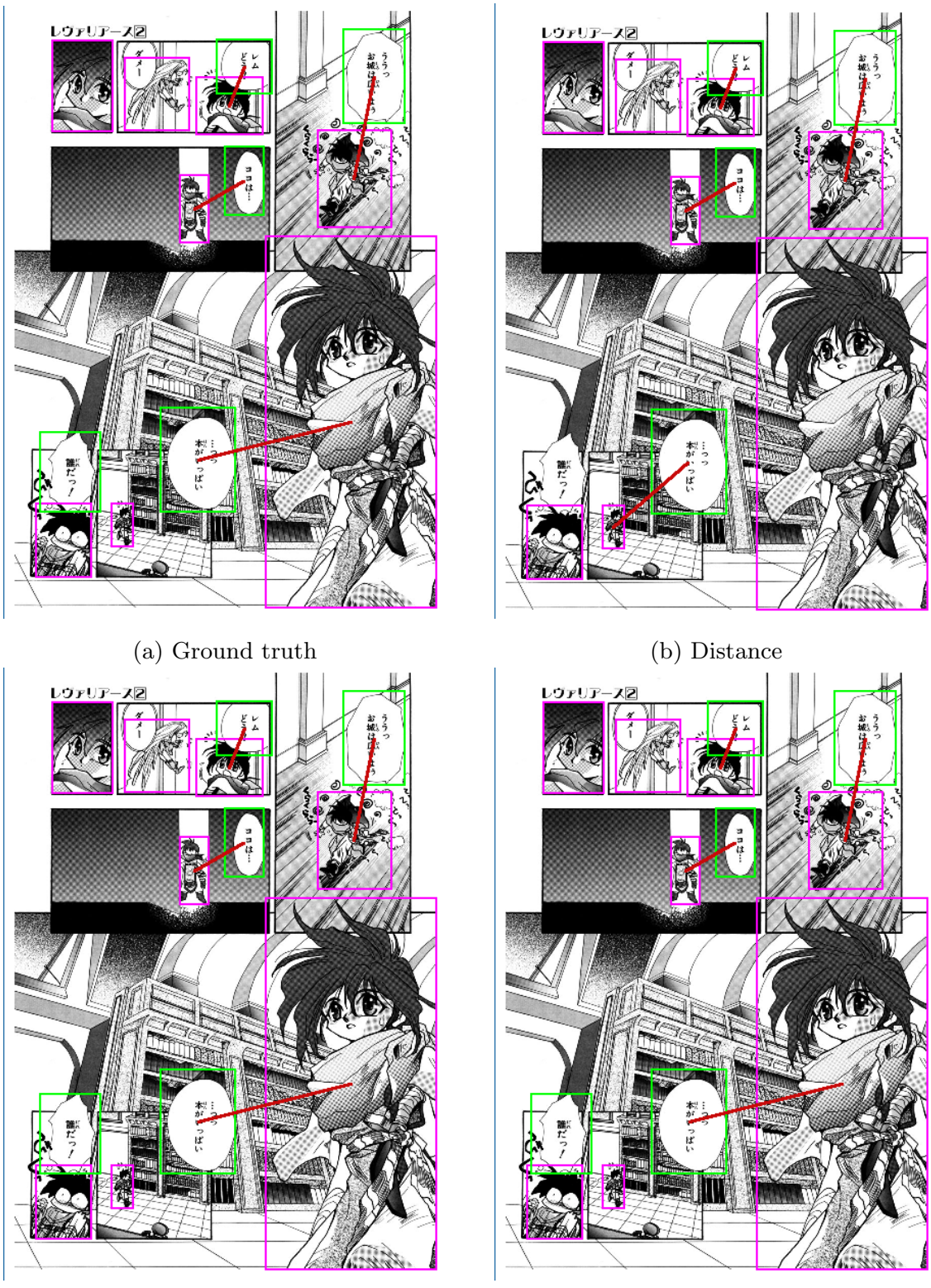
手法	Top1-Accuracy
Random	57.4%
Distance	77.9%
Human performance	91.3%
w/o balloon image	80.6%
w balloon image	79.2%

のみを用いる手法のいずれにおいても話者推定に失敗した例を図 4.8 に示す. 吹き出しのテールが明確に存在する例であるが, 吹き出し画像を用いる手法においても話者推定がうまくいっていないことから, テールの情報を適切に利用できていないと考えられる.

吹き出し画像とメタデータから抽出した特徴量の両方を用いた手法とメタデータから抽出した特徴量のみを用いた手法のいずれの場合においても, 人手での話者推定と比べて低い精度となった. 人手での話者推定の場合のみ正解している例を図 4.9 に示す. テールが明確であるため, 人手での話者推定は容易である一方, 提案手法ではこれらの吹き出しの話者は吹き出しから最も近いキャラクターが選択されてしまう. 提案手法は吹き出しとキャラクターの距離情報だけに基づいた手法と比べて精度が上がっているものの, 人手での話者推定には及ばないという結果となった.

また, 人手での話者推定がうまくいかなかった例を図 4.10 に示す. 人手での話者推定においては吹き出しのテールの位置や向きが重要な要素であるため, このようなケースでは正確な話者推定が困難となる. 文字情報を考慮しない話者推定には限界があることから, さらに高い精度での話者推定のためには文字情報も考慮できるようなシステムが必要となると考えられる.

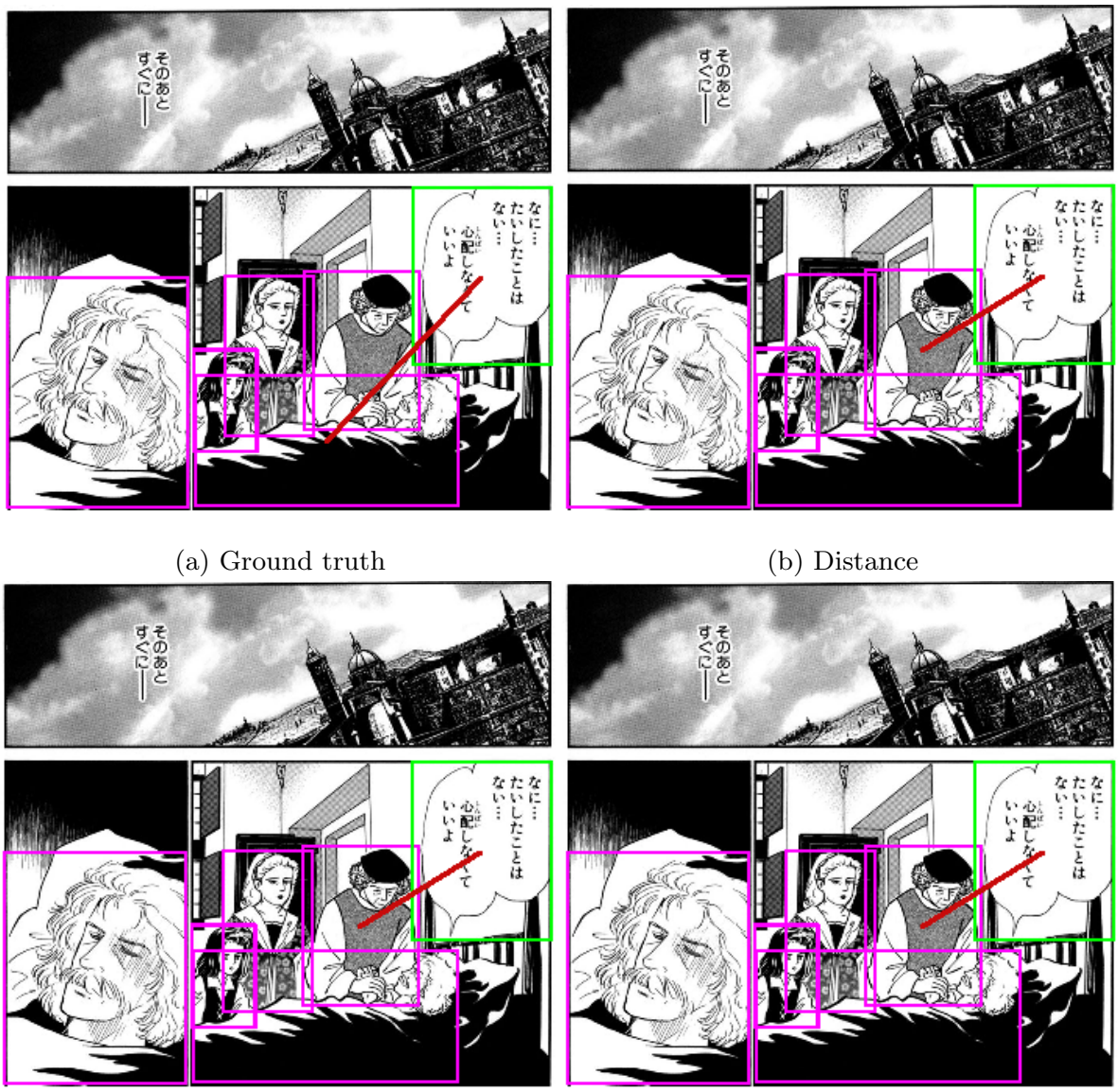
続いて, テストデータの 5 冊の漫画における漫画ごとの精度の差を図 4.11 に示す. 吹き出し画像とメタデータを用いた話者推定の精度は 73.5% から 90.4% と幅があり, 漫画によって話者推定の難易度は大きく異なることが分かる. いずれの漫画においても距離に基づいて話者推定を行う比較手法と大きな差はない一方, 人手による話者推定との精度の差は漫画によって大きく異なる. これは, 距離に基づく手法や吹き出し画像とメタデータに基づく手法はどちらも吹き出しとキャラクターの距離を重視して話者推定を行っている一方, 人手での話者推定は吹き出しとキャラクターの距離よりもテールの位置や向きとキャラク



(a) Ground truth　(b) Distance

(c) w/o balloon image　(d) w balloon image

図 4.7: 吹き出し画像とメタデータを用いる手法およびメタデータのみを用いる手法のいずれにおいても話者推定に成功した例. ©夜麻 みゆき



(a) Ground truth　(b) Distance

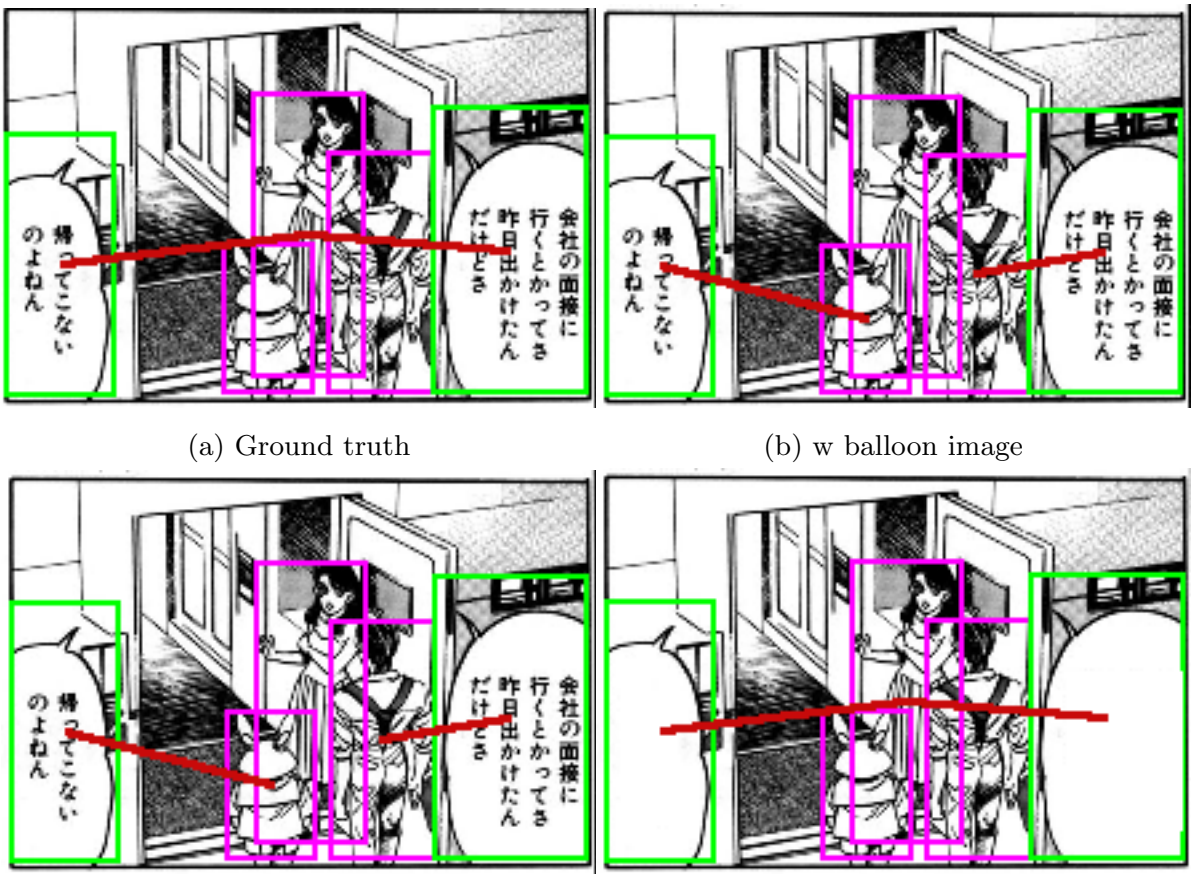
(c) w/o balloon image　(d) w balloon image

図 4.8: 吹き出し画像とメタデータを用いる手法およびメタデータのみを用いる手法のいずれにおいても話者推定に失敗した例. ©里中 満智子

ターの位置関係を重視して話者推定を行っているという違いに起因すると考えられる.

4.3.4　メタデータから抽出した特徴量の有用性の評価

今回用いた特徴量のうち, どの特徴量が有効に働いたのかを検証するため, メタデータから抽出した特徴量だけを用いた手法において, 特徴量を 1 つだけ取り除いたものについ



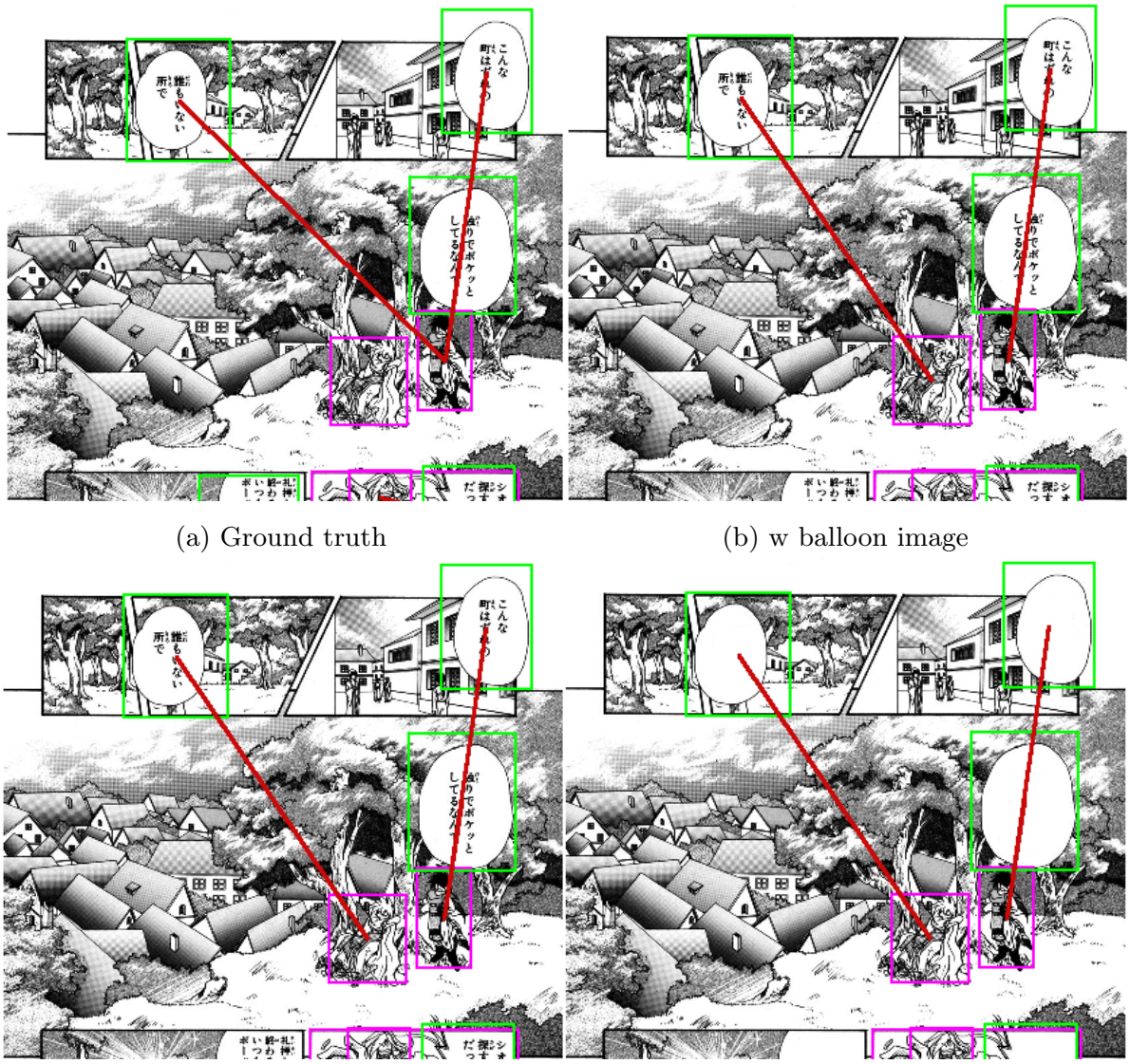
(a) Ground truth　(b) w balloon image

(c) w/o balloon image　(d) Human performance

図 4.9: 人手での話者推定のみ正解している例. ©内田 美奈子

て精度を検証した. 4.3.1 で述べた, メタデータから抽出した特徴量だけを用いたネットワークから入力を 1 次元減らしたネットワークを用いてそれぞれ学習およびテストを行った. その結果を表 4.5 に示す. ただし, Nothing はメタデータだけを用いた手法から何も取り除かないもの, -Distance は“吹き出しとキャラクターの距離”の情報を除いた手法, -Distance Rank は“吹き出しとキャラクターの距離の順位”の情報を除いた手法, -Frame は“吹き出しとキャラクターのペアが同じフレーム内に存在しているかどうか”の情報を除いた手法, -Character Size は“キャラクター領域の面積”の情報を除いた手法, -Balloon Size は“吹き出し領域の面積”の情報を除いた手法, -Angle は“吹き出しとキャラクターの位置関係”の情報を除いた手法である.

基本的にメタデータからいずれかの情報を取り除くと精度が低下するが, キャラクターの大きさの情報だけは取り除いても精度がほぼ変わらないという結果となった.“吹き出しとキャラクターの距離”の情報と“, 吹き出しとキャラクターの距離の順位”の情報は非常



(a) Ground truth　(b) w balloon image

(c) w/o balloon image　(d) Human performance

図 4.10: 人手での話者推定も失敗している例. ©夜麻 みゆき

に相関の強い特徴量であると考えられるが, どちらか片方を取り除くと精度が低下するという結果となった. また, “吹き出しとキャラクターのペアが同じフレーム内に存在しているかどうか” の情報を取り除くと特に大きく精度が下がったことから, フレームの情報は非常に重要であると考えられる. フレームの情報を取り除くことにより話者推定がうまくいかなくなった例を図 4.12 に示す. “同じフレーム内のキャラクターは話者になりやすい” という傾向を考慮できないため, 距離が近いキャラクターを優先的に選択した結果誤りが発生していると考えられる.

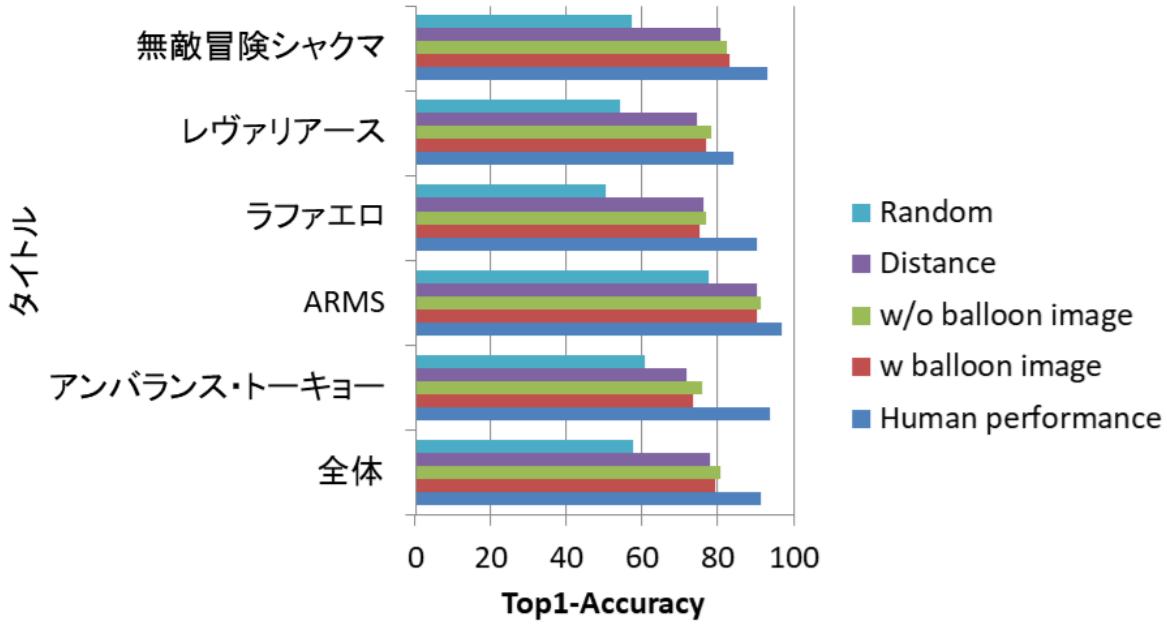


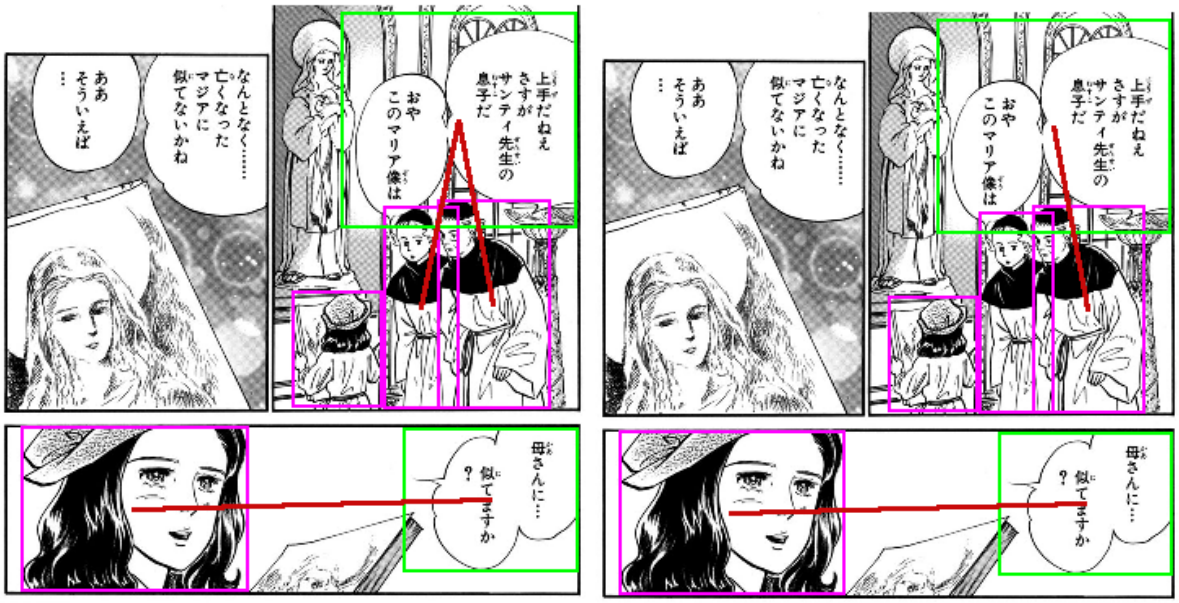
図 4.11: テストデータに用いた 5 冊の漫画における話者推定精度

表 4.5: 6 つの特徴量のうち 1 つを除いた場合の話者推定の Top1-Accuracy での精度評価.

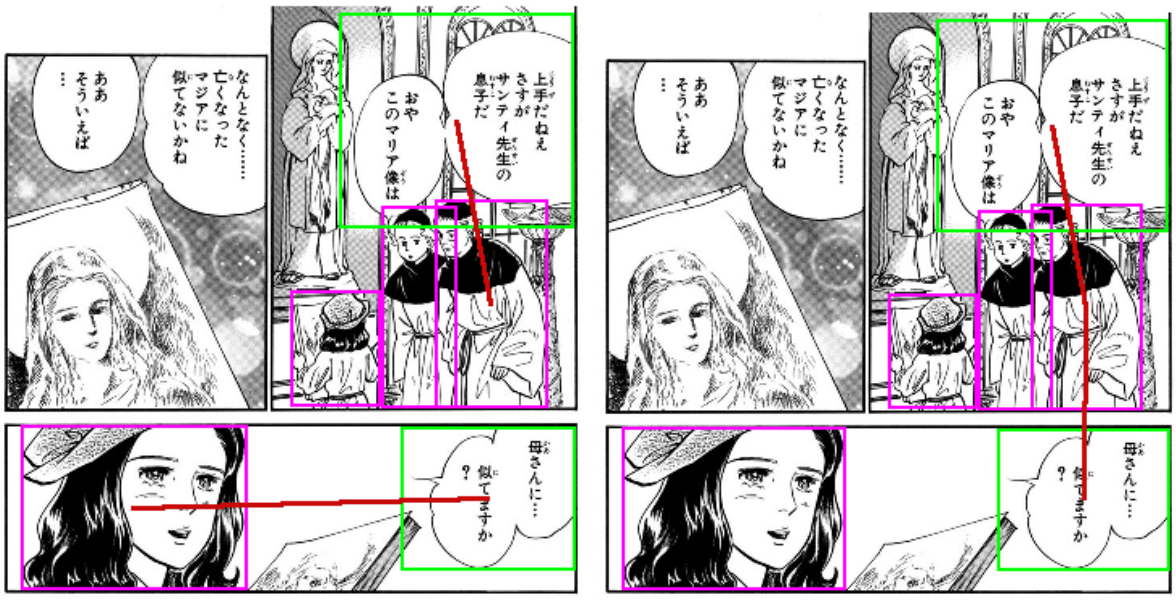
除いた特徴量	Top1-Accuracy
Nothing	80.6%
Distance	78.6%
Distance Rank	77.9%
Frame	72.4%
Character Size	80.3%
Balloon Size	75.1%
Angle	75.4%

キャラクターの面積の情報を取り除くことによって話者推定がうまくいかなかった例を図 4.13, キャラクターの面積の情報を取り除くことによって話者推定に成功した例を図 4.14 に示す. キャラクターの面積の情報を取り除くと, キャラクターの面積が考慮されなくなり, 吹き出しから近い位置にいるキャラクターが選択される傾向が強まると考えられる. それにより話者推定に失敗してしまうケースと成功するケースが同程度の数あり, 取り除いても精度がほとんど変わらないと考えられる.

吹き出しの面積の情報を取り除くことによって話者推定がうまくいかなくなった例を図

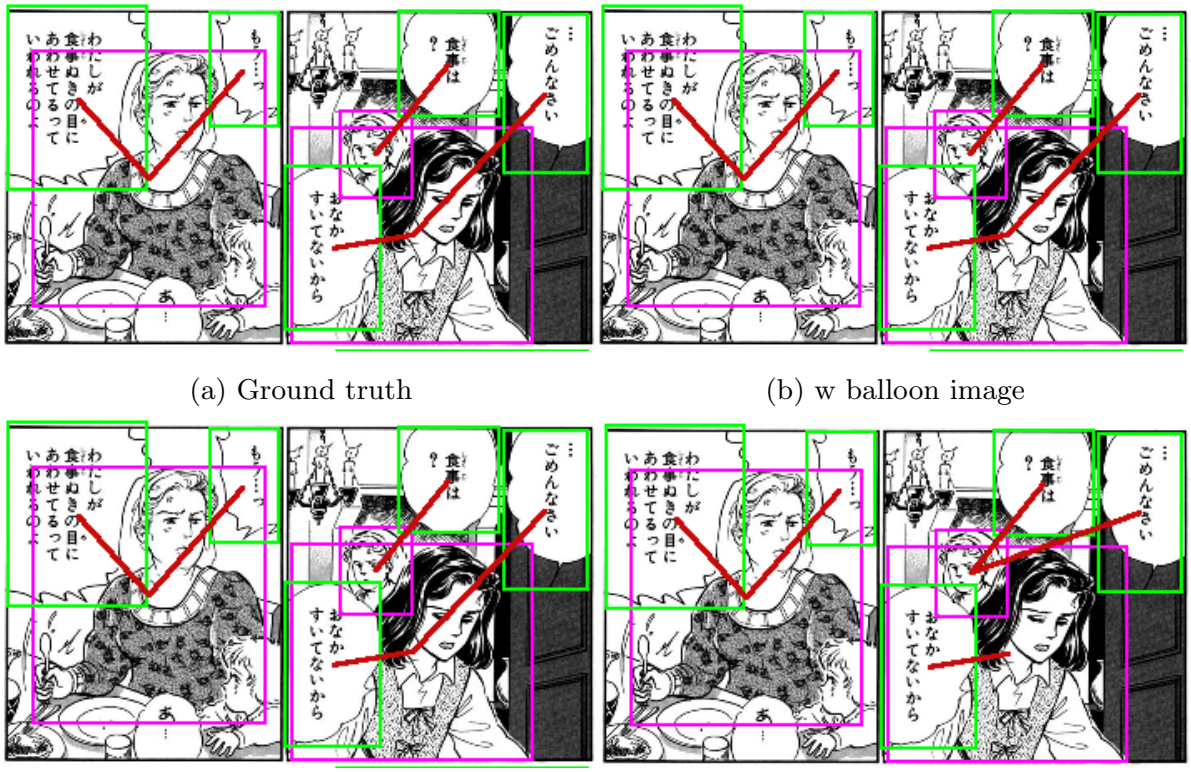


(a) Ground truth　(b) w balloon image



(c) w/o balloon image　(d) w/o balloon image and frame

図 4.12: 吹き出しとキャラクターが同じフレームに所属しているかどうかの情報を取り除くと話者推定に失敗する例. ©里中 満智子



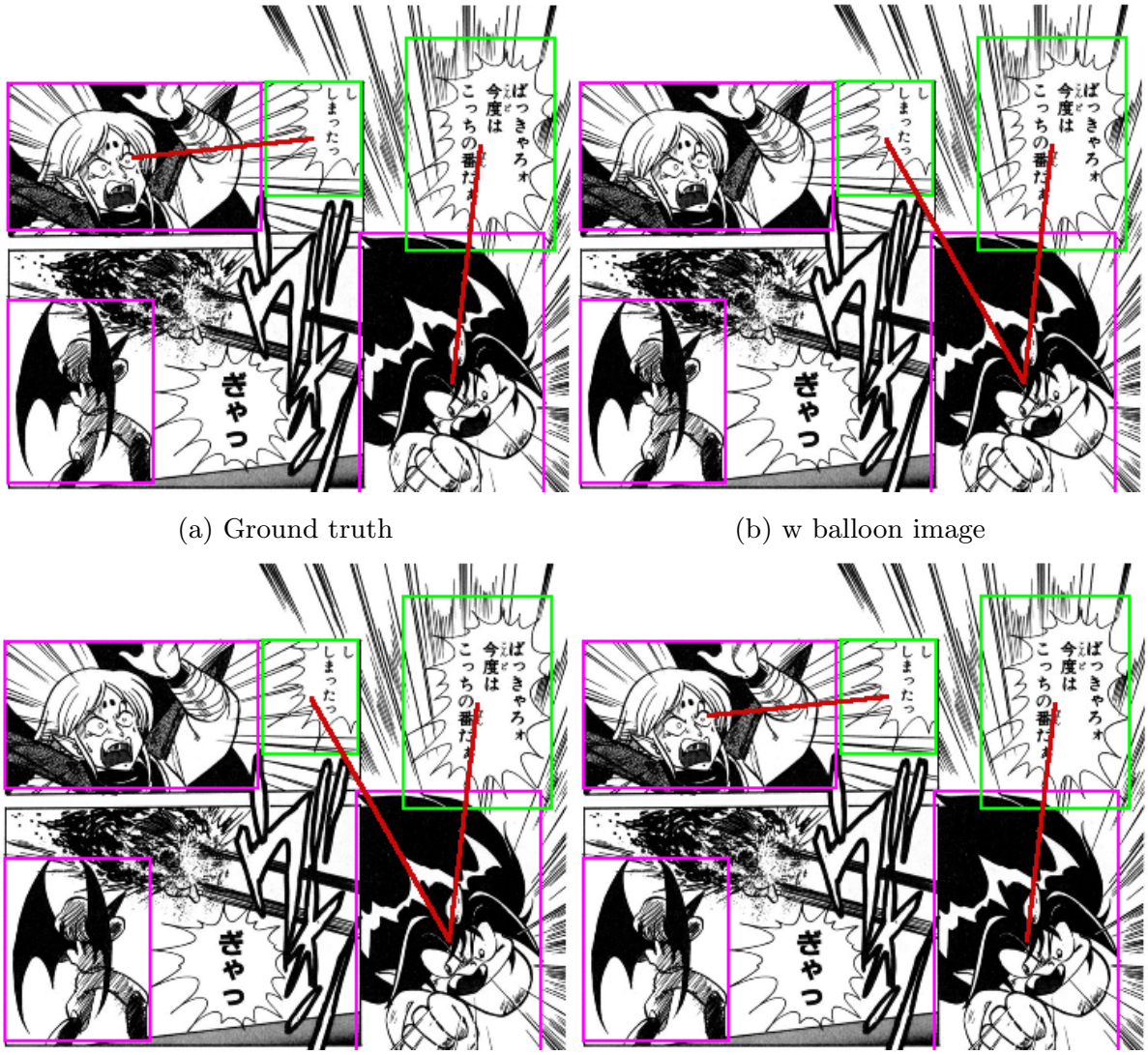
(a) Ground truth　(b) w balloon image

(c) w/o balloon image　(d) w/o balloon image and character size

図 4.13: キャラクターの面積の情報を取り除くと話者推定に失敗する例. ©里中 満智子

4.15 に示す. 吹き出しの話者には一般に, “面積の大きいキャラクターほど吹き出しの話者になりやすい”, “吹き出しの大きさと話者の大きさには相関がある (面積の大きい吹き出しの話者は面積の大きいキャラクターであることが多く, 面積の小さい吹き出しの話者は面積の小さいキャラクターであることが多い) ” という傾向が見られる. そのため, 吹き出しの大きさ情報を取り除いてしまうと, 面積の小さい吹き出しの話者キャラクターは面積が小さい可能性が高いということが考慮されず, “面積の大きいキャラクターほど吹き出しの話者になりやすい” という傾向だけが考慮された結果, 話者推定がうまくいかなかったケースが多かったのではないかと考えられる.

また, “吹き出しとキャラクターの位置関係 (向き) ” の情報は吹き出しの画像特徴量と組み合わせて利用することを想定して加えていた特徴量であるが, 単独でも精度に寄与していることが分かった. 吹き出しとキャラクターの位置関係の情報を取り除くことによって話者推定がうまくいかなくなった例を図 4.16 に示す. 吹き出しの話者は吹き出しより下側に存在することが多く, 吹き出しより上側に存在することは少ない. 吹き出しとキャラクターの位置関係の情報を取り除いてしまうとそのような事情が考慮できないため, “吹き出

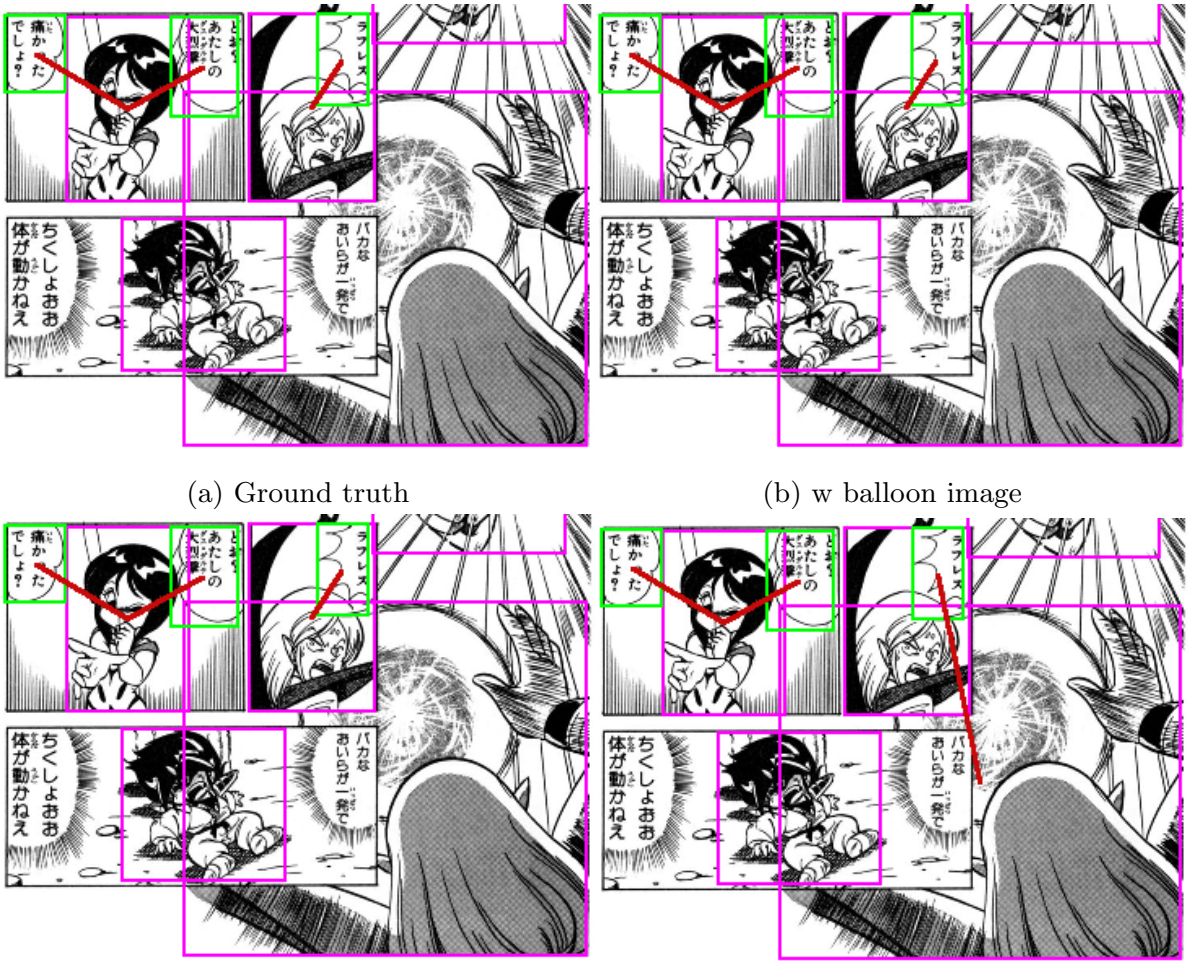


(a) Ground truth　(b) w balloon image

(c) w/o balloon image　(d) w/o balloon image and character size

図 4.14: キャラクターの面積の情報を取り除くと話者推定に成功する例. ©計奈 恵

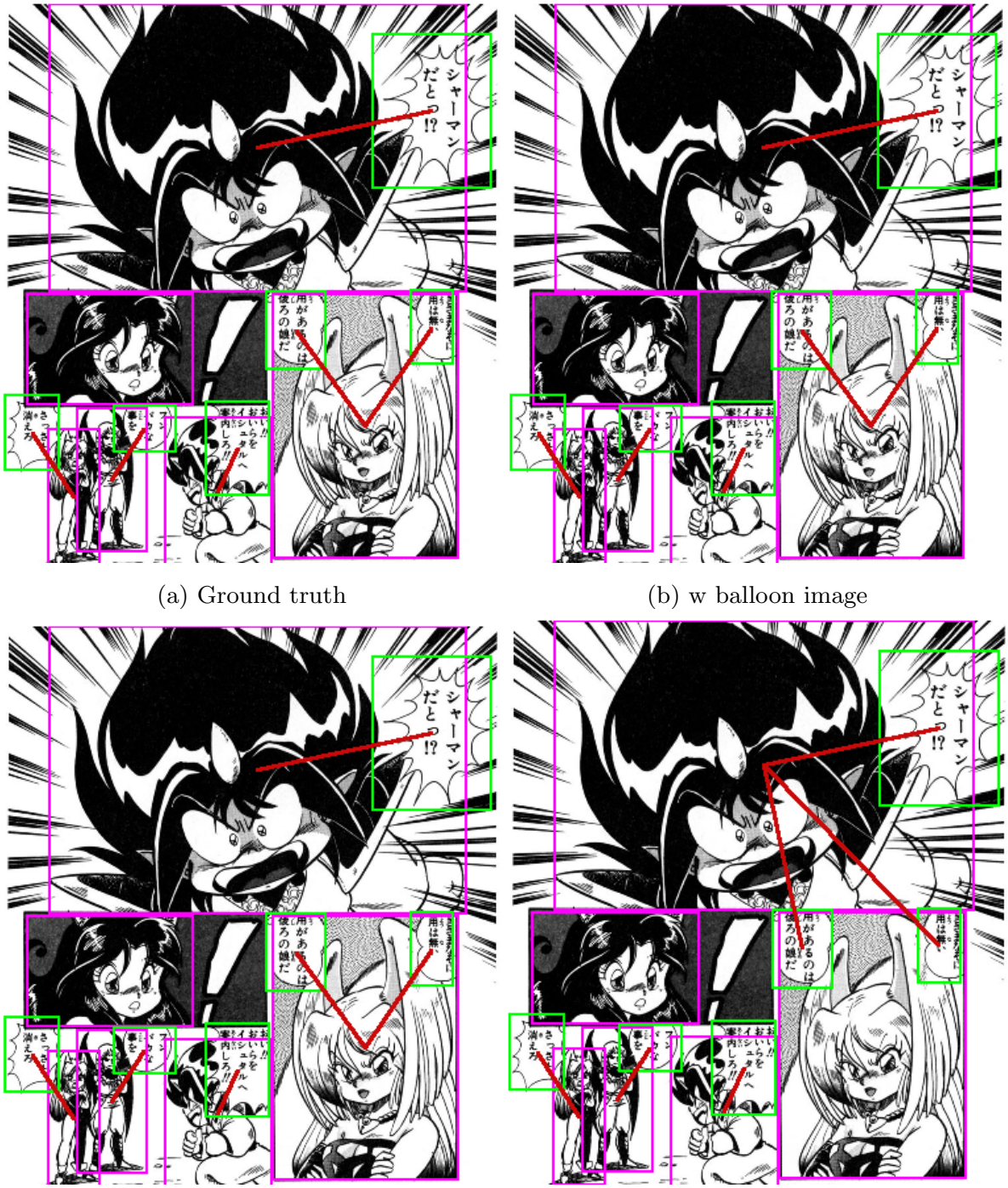
しと距離が近いキャラクターほど話者になりやすい”,“面積の大きいキャラクターほど吹き出しの話者になりやすい”といった傾向が重視されてしまい話者推定に失敗するケースが多いと考えられる.



(a) Ground truth　(b) w balloon image

(c) w/o balloon image　(d) w/o balloon image and balloon size

図 4.15: 吹き出しの面積の情報を取り除くと話者推定に失敗する例. ©計奈 恵



(a) Ground truth　(b) w balloon image

(c) w/o balloon image　(d) w/o balloon image and character angle

図 4.16: 吹き出しとキャラクターの位置関係の情報を取り除くと話者推定に失敗する例. ©計奈 恵

第 5 章

結論

5.1　まとめ

本研究では，漫画画像とメタデータを用いた学習ベースの手法で漫画中の吹き出しの話者推定を行う手法を提案し, その性能を評価した．本研究で提案した内容を以下に示す.

深層学習を用いて吹き出しの話者推定を行う手法を提案した．また, 特徴量として, 吹き出し領域の画像とメタデータを使用することを提案した. また, 学習ベースの手法を用いるにあたって必要となる吹き出しとその話者のデータセットを作成した.

これらの提案の有効性を確認するため，提案したデータセットを使用し評価実験を行った. 実験の結果, 従来の手法と比べて高い精度で吹き出しの話者を推定できることが分かった．一方で, 吹き出し画像を入力に用いることは精度の向上に寄与しないことが分かった. また, メタデータから抽出した 6 つの特徴量についてそれぞれ有効性を検証する実験を行った. 実験の結果, キャラクターの面積の情報は取り除いても話者推定の精度はほぼ変わらないが, それ以外の 5 つの特徴量は話者推定の精度に寄与していることが分かった.

5.2　今後の課題

吹き出しの話者推定における今後の課題を以下に示す.

1. データセットの大規模化
   本研究では, 学習ベースの手法で吹き出しの話者推定を行う手法を提案した. それに伴い, Manga109 データセットの一部に吹き出しのアノテーションを付与したデータセットを提案したが, 吹き出し領域のデータ数はおよそ 5000 個程度となっている. 実験においても, “面積の大きいキャラクターほど吹き出しの話者になりやすい”

という傾向について, 学習データの数が足りないために十分な学習ができていないと考えられる結果が見られた. データセットをより大規模なものとすることが話者推定の精度の向上につながる可能性がある.

2. ネットワークの構造や, 用いる特徴量の改善
   本研究では, メタデータからいくつかの特徴量を抽出し, 深層学習の入力に用いることを提案し, それぞれの特徴量が話者推定の精度にどのように寄与するかを検証した. しかし, 本研究で提案したもの以外にも話者推定に重要な特徴量が存在する可能性があるが, それについての検証は必ずしも十分ではない. また, ネットワークの構造についても, 検討の余地がある.
3. 話者が 0 人の場合や 2 人以上の場合への対応
   実世界での応用においては吹き出しの話者が 1 人も存在しない場合や, 逆に 2 人以上の話者がいる場合についても対応する必要があるが, 本研究で提案した手法ではそれらのケースには対応できていない. 話者が 0 人や 2 人以上の場合に対応できる手法についても検討していく必要がある.
4. 文字情報の考慮
   本研究では, 文字を隠した状態で人手での話者推定についても精度の評価を行い, 文字情報を考慮せずに行う話者推定の限界を示した. 今後, より高い精度での話者推定を行うためには文字情報についても考慮できるようなモデルを検討していく必要がある.

謝辞

Manga109 データセットをはじめとする漫画の画像データセットは本研究に欠かせないものでした．漫画の作者の皆さまには，その画像を研究に使用することを許可していただいたことに感謝いたします．

指導教官である相澤教授には，多数のアドバイスをいただきました．特に研究の方針や論文の構成についてなど, 重要な場面でのアドバイスを多数いただきました

山崎准教授にはミーティングでの発表などの場において多数のご指摘や意見をいただきました. 研究の方向性を決めていく上で山崎先生の意見には大いに助けになりました．

秘書の松林さんには様々な事務手続きでお世話になりました．

研究室の同期とは様々な場面で意見を出し合い, 何度も助けてもらいました. 特に同じく漫画を研究の題材とする小川君には非常に有意義な意見を多数もらいました．

漫画班の先輩である松井さん, 藤本さん, 後輩の成田くん, 大坪くん, 坪田くんとは漫画班ミーティングなどの場で有意義な議論ができました．

ここに名前を載せられなかった研究室のメンバーとも，研究室で切磋琢磨し, 助け合う仲間として様々な場面でお世話になりました．

参考文献

[1] マンガ図書館 Z, http://www.mangaz.com/.

[2] Krizhevsky. Alex, Sutskever. Ilya, and Hinton. Geoffrey. E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[3] Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari. Measuring the objectness of image windows. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2189–2202, 2012.

[4] Kohei Arai and Herman Tolle. Automatic e-comic content adaptation. *International Journal of Ubiquitous Computing*, 1(1):1–11, 2010.

[5] Kohei Arai and Herman Tolle. Method for real time text extraction of digital manga comic. *International Journal of Image Processing (IJIP)*, 4(6):669–676, 2011.

[6] Yuji Aramaki, Yusuke Matsui, Toshihiko Yamasaki, and Kiyoharu Aizawa. Text detection in manga by combining connected-component-based and region-based classifications. In *Image Processing (ICIP), 2016 IEEE International Conference on*, pages 2901–2905. IEEE, 2016.

[7] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Computer vision and image understanding*, 110(3):346–359, 2008.

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

[9] 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*, page 2. ACM, 2016.

[10] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[11] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[12] 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 *Document Analysis and Recognition (ICDAR), 2013 12th international conference on*, pages 1145–1149. IEEE, 2013.

[13] Yanagisawa Hideaki and Watanabe Hiroshi. Faster r-cnn を用いたマンガ画像からのメタデータ抽出. In *2016* 年映像情報メディア学会年次大会. ITE, 2016.

[14] Gang Hua and Amir Akbarzadeh. A robust elastic and partial matching metric for face recognition. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 2082–2089. IEEE, 2009.

[15] He. Kaiming, Zhang. Xiangyu, Ren. Shaoqing, and Sun. Jian. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[16] Samu Kovanen and Kiyoharu Aizawa. A layered method for determining manga text bubble reading order. In *Image Processing (ICIP), 2015 IEEE International Conference on*, pages 4283–4287. IEEE, 2015.

[17] 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 *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*, pages 761–765. IEEE, 2015.

[18] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[19] Luyuan Li, Yongtao Wang, Ching Y Suen, Zhi Tang, and Dong Liu. A tree conditional random field model for panel detection in comic images. *Pattern*

Recognition, 48(7):2129–2140, 2015.

[20] M. Lin, Q. Chen, and S. Yan. Network in network. In *International Conference on Learning Representations*.

[21] 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*, pages 21–37. Springer, 2016.

[22] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[23] Jiri Matas, Ondrej Chum, Martin Urban, and Tomás Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Image and vision computing*, 22(10):761–767, 2004.

[24] Yusuke Matsui, Kota Ito, Yuji Aramaki, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using manga109 dataset. *arXiv preprint arXiv:1510.04389*, 2015.

[25] Jean-Michel Morel and Guoshen Yu. Asift: A new framework for fully affine invariant image comparison. *SIAM Journal on Imaging Sciences*, 2(2):438–469, 2009.

[26] Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. 漫画物体検出に向けた検出器の並列化. In *2017* 年 第 *16* 回情報科学技術フォーラム. IPSJ, 2017.

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

[28] 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*, pages 91–99, 2015.

[29] Christophe Rigaud, Jean-Christophe Burie, Jean-Marc Ogier, Dimosthenis Karatzas, and Joost Van de Weijer. An active contour model for speech balloon detection in comics. In *2013 12th International Conference on Document Analysis and Recognition*, pages 1240–1244. IEEE, 2013.

[30] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[31] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going

deeper with convolutions. June 2015.

[32] Kohei Takayama, Henry Johan, and Tomoyuki Nishita. Face detection and face recognition of cartoon characters using feature extraction. In *Image, Electronics and Visual Computing Workshop*, page 48, 2012.

[33] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013.

[34] Yongtao Wang, Xicheng Liu, and Zhi Tang. An r-cnn based method to localize speech balloons in comics. In *International Conference on Multimedia Modeling*, pages 444–453. Springer, 2016.

[35] Yongtao Wang, Yafeng Zhou, and Zhi Tang. Comic frame extraction via line segments combination. In *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*, pages 856–860. IEEE, 2015.

[36] Hideaki Yanagisawa, Daisuke Ishii, and Hiroshi Watanabe. Face detection for comic images with deformable part model. In *4th IIEEJ International Workshop on Image Electronics and Visual Computing (October 2014)*, 2014.

発表文献

国際会議

[1] Azuma Fujimoto, Toru Ogawa, Kazuyoshi Yamamoto Yusuke Matsui, Toshihiko Yamasaki, Kiyoharu Aizawa. Manga109 Dataset and Creation of Metadata. ICPR MANPU 2016

国内会議

[2] 藤本東, 小川徹, 山本和慶, 松井勇佑, 山崎俊彦, 相澤清晴. Manga109 とそのメタデータ基盤の構築．画像の認識・理解シンポジウム 2016.

[3] 藤本東, 小川徹, 山本和慶, 松井勇佑, 山崎俊彦, 相澤清晴. 学術用アノテーション付き漫画画像データセットの構築と解析．映像情報メディア学会,・メディア工学研究会 2016.

[4] 相澤清晴, 山崎俊彦, 松井勇佑, 小川徹, 山本和慶, 大坪篤史, 成田嶺, 坪田亘記．漫画・イラストのマルチメディア処理に向けた基盤技術研究．情報処理学会・電子情報通信学会 情報科学技術フォーラム 2017.

[5] 山本和慶, 小川徹, 山崎俊彦, 相澤清晴．データドリブンなアプローチを用いた漫画画像中の吹き出しの話者推定． (発表予定) メディア工学研究会 2017.