Vol.2020-CH-122 No.4
2020/2/1

深層学習とデータ合成を用いた古文書画像からの行切り出し

犬塚 直人（芝浦工業大学大学院理工学研究科 システム理工学専攻）
鈴木 徹也（芝浦工業大学大学院理工学研究科 システム理工学専攻）

概要：我々は手書き変体仮名の翻刻作業を支援するために, 機械学習を用いた行切り出しを導入しようと考えている. しかし, 機械学習のための大量のアノテーション付き文書画像データの手作業による構築は大変手間のかかる作業である. そこで, 我々はデータ合成によるアノテーション付き文書画像データの構築法を模索している. 本研究では現状のデータ合成システムについて報告する. また機械学習を用いて実際の古文書画像から行を切り出す際に, データ合成したアノテーション付き文書画像が学習データとして有効であるかを確かめるべく実際の古文書画像との比較実験を行った. 合成画像を元に学習したモデルによる行切り出し結果の Intersection Over Union は, 実際の古文書画像を元に学習したモデルのそれと同程度であった.

Text Line Segmentation for Japanese Historical Document Images Using Deep Learning and Data Synthesis

1. はじめに

我々は手書き変体仮名翻刻支援システムの開発に取り組んでいる. そのシステムの中で, 文書画像からの行切り出し処理が必要とされている. 我々はその処理に機械学習によって得られたモデルを利用する予定である.

そこで問題となるのは, 機械学習のための大量のアノテーション付き文書画像データの構築である. 手作業によるその構築は大変手間のかかる作業である.

現在, 我々はデータ合成によるアノテーション付き文書画像データの構築法を模索している. それが可能になれば, 多様な文書画像を低コストで大量に作成できる. 本稿ではその現状について報告する.

以降、本論文の構成は次のようになっている. 第 2 節で関連研究を紹介する. 第 3 節で今回我々が対象とする文書の特徴とその特徴を持った文書画像の合成法とを説明する. 第 4 節ではその文書画像合成法の評価実験について述べる. 第 5 節でまとめと今後の課題について述べる.

2. 関連研究

2.1 手書き変体仮名翻刻支援システム

我々が開発に取り組んでいる手書き変体仮名翻刻支援システムについて簡単に紹介する [6], [7], [8], [9], [10], [11]. 名前の通り, 手書き変体仮名で記述された書物の翻刻を支援するシステムである. このシステムの特徴的な点は, 文書画像からの複数通りの文字の切り出し方, 切り出された文字の複数通りの読み順, 切り出された各文字の複数通りの認識結果の組み合わせの中から最適な組み合わせを出力する点である.

システムの構成は主に画像解析器, 制約充足器そしてそれらを結合するグラフィカルユーザインタフェース（Graphical User Interface, GUI）からなる. 画像解析器は文字切り出し, 読み順の決定, 各文字の認識を行う. その結果を元に構成された制約充足問題を制約充足器が解くことで翻刻結果を得る. 制約充足器は, 中古和文用の形態素解析辞書 [13], [14], [15] を参照しながら, 単語の出現コストと単語の接続コストとの総和が最小となるような組み合わせを探索する. 利用者は GUI を通して画像解析器と制約充足器とを利用する.

現在, GUI と制約充足器との実装は進んでいるが, 画像解析器は未実装である. 画像解析器の最初の処理として文書画像からの行切り出しを予定している.

2.2 文書画像解析のための学習データ構築

機械学習を用いた文書画像解析での問題の一つは学習データの構築である. 大量の文書画像への手作業によるアノテーション付けは大変手間のかかる作業である. ここではその学習データの構築に関連する研究を三つ紹介する.

ただしいずれも日本語で書かれた古文書を対象としたデータ合成に関する研究ではない．

Capobianco らは歴史的文書画像の生成ツールを提案した [1]．歴史的文書画像を生成する手順は次のようになる．まず少数の文書画像を用意し，その文書画像からページの背景画像を抽出する．利用者がその少数の文書画像を参考にし，文書の構造を XML で記述する．次に，抽出したページの背景画像，文書構造を記述した XML，フォント，辞書とを生成器に与える．その結果，歴史的文書画像が生成される．多様なページを生成するために，行の高さや項目の繰り返し回数をランダムに決めたり，その行を生成する確率を指定することができる．さらにページの回転やノイズを加えてデータ拡張も可能である．

Pondenkandath らは深層ニューラルネットワークを用いた歴史的文書の合成法を提案した [4]．その手法ではまず LaTeX を用いて文書を作成する．次にその文書の画像を深層ニューラルネットワークに与え，手書きの歴史的文書風の画像を生成する．Pondenkandath らは画像変換に用いる深層ニューラルネットワークとして，CycleGAN と Neural Style Transfer との二通りを実験により比較した．その結果，Pondenkandath らは CycleGAN の方が有望であるとしている．

青池らは機械学習に利用可能な資料レイアウトデータセットを構築しそれを公開した [12]．対象となった文書は，国立国会図書館デジタルコレクションのデジタル化資料である．そのレイアウトデータセットは，文書レイアウト認識のための機械学習モデルで得た推定レイアウトを，アノテーションツールで修正することで構築された．精度の向上をはかるため，概ね 100 枚から 200 枚処理するごとに，それまでに得られたレイアウトを学習データに加えて機械学習モデルを更新した．

3. データ合成法

本研究で対象とする文書の特徴，その特徴を持った文書のモデル，そして文書モデルをもとに手書き古文書風の画像を合成する方法について順に説明する．

3.1 対象とする文書の特徴

本研究では文献 [2] の伊勢物語を対象とする．縦書きの文字だけからなる文書で比較的扱いやすいのがその理由である．この文書のレイアウトについて我々は次の点に注目した．

特徴 1　和歌や短歌などの出典を表す注釈（図 1（1））
特徴 2　行末での折り返し（図 1（2））
特徴 3　ルビや補足（図 1（3））
特徴 4　異なる行の文字同士の重なり（図 1（4））
特徴 5　行の中心線が傾いている行（図 1（5））
特徴 6　他の段落に補足的に記述された段落（図 1（6））

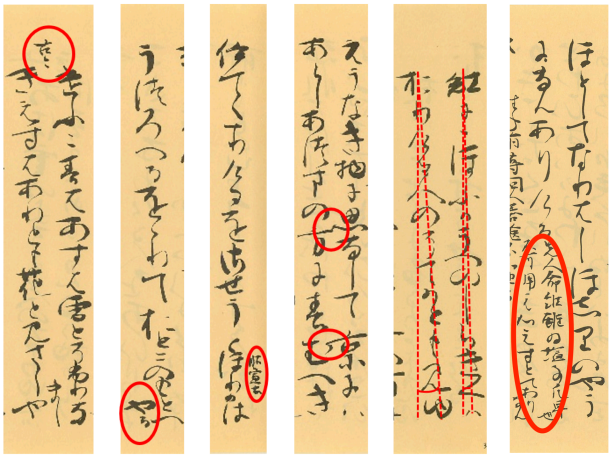

図 1　対象とする文書の特徴（文献 [2] から引用した文書画像に加筆した）

特徴 7　連綿（つづけ字）

3.2 文書モデル

第 3.1 節で挙げた特徴を持った文書のモデルを考案した．ただしこのモデルは特徴 7 の連綿については考慮していない．文書モデルは行を表す line，段落を表す paragraph，文書全体の書式を表す format から構成されている．以下，この順に説明する．

3.2.1 行 line

line は一行を表す．line はその行の文字のフォント幅，行の縦方向の長さ，行の中心線，ルビ，折り返し部分を属性に持つ（図 2）．行の中心線は折れ線で表現される．行内の文字はこの折れ線上に配置される．これによって第 3.1 節で指摘した文書の特徴 5（図 1（5））に対応する．文字の縦方向の間隔は format で指定される．ルビは line であり，行内の何文字目のルビかを指定できる．ルビは指定した文字の右側に配置される．これによって第 3.1 節で指摘した文書の特徴 3（図 1（3））に対応する．折り返し部分は line であり，上の余白を指定できる．折り返し部分は本体の行に接するように配置される．これによって第 3.1 節で指摘した文書の特徴 2（図 1（2））と特徴 4（図 1（4））に対応する．

3.2.2 段落 paragraph

paragraph は一つの段落を表す．上下左右の余白，上左右に配置する注釈，行間，段落の本体である行の列を属性に持つ（図 3）．上左右に配置する注釈はそれぞれ paragraph である．これらによって，第 3.1 節で指摘した文書の特徴 1（図 1（1））と特徴 6（図 1（6））に対応する．左右の注釈は，段落の本体と接するまで近くに配置される．段落の本体を構成する行は，それぞれが接するまで近くに配置される．ただし行頭の文字とその直前の行との間には行間以上の空間を空けることにする．これらによって特徴 4（図 1（4））に対応する．

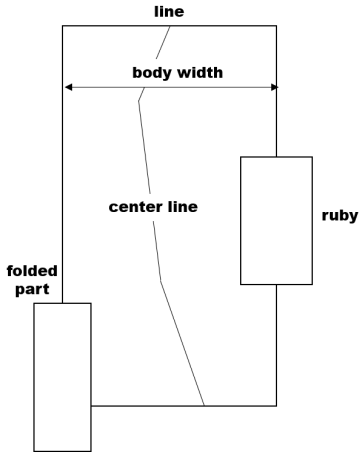

図 2　line

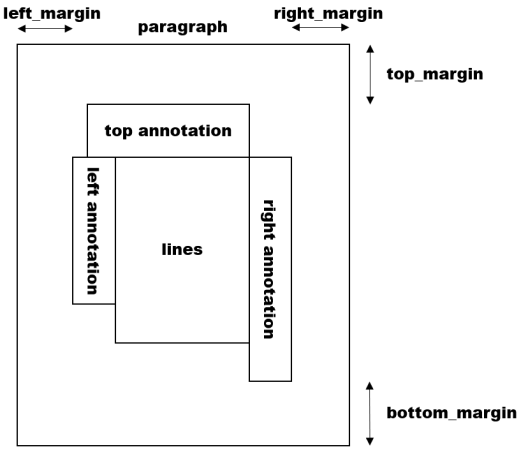

図 3　paragraph

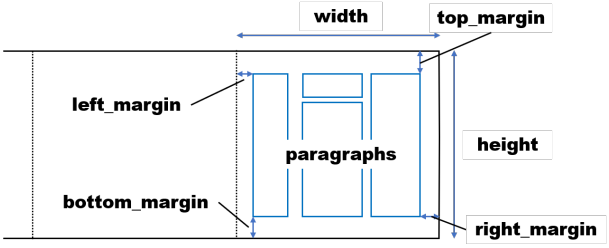

図 4　format

3.2.3　全体の書式 format

format は文書全体の書式を表し, ページの幅と高さと上下左右の余白, そして段落の列を属性にもつ（図 4）.

3.3　システム

データ合成システムは Python で記述された 4 つのプログラム fonts, format, typeset, print で構成されている. データ合成システムの概要を図 5 に示す.

3.3.1　fonts プログラム

fonts プログラムは, Kuzushiji-MNIST[3] から各平仮名について変体仮名のフォントデータを n 個ずつ抽出する. 各フォントデータを白地に黒文字に変換し, 文字の上下の白地の部分は削除する. そして抽出した各フォントデータをファイル（fontdata ファイル）に出力する. フォントデータはグレイスケールであるが, フォントデータを二値化することもできる.

3.3.2　format プログラム

format プログラムは第 3.2 節で示した文書モデルに基づいた文書の書式を Python の辞書オブジェクトとして生成し, ファイル（format ファイル）に出力する. 乱数を用いて, 多様な段落や行を生成する.

3.3.3　typeset プログラム

typeset プログラムは fontdata ファイルと format ファイルとを入力とし, 組版結果をファイル（typesetting ファイル）に出力する. typeset プログラムは, ランダムに平仮名を選択し, その平仮名のフォントを fontdata file からランダムに抽出する. そしてそのフォントデータに基づき format ファイルで指定された行に文字を配置してゆく. また typeset プログラムは文字を配置した行をページに区切って配置する.

3.3.4　print プログラム

print プログラムは fontdata ファイルと typesetting ファイルを用いて, 文書画像とそのアノテーション（各行を囲む矩形）を生成する.

4.　実験

4.1　目的

データ合成した古文書画像で学習した物体検出アルゴリズムが, 実際の古文書画像からの行切り出しに有効であるかを確かめるべく実験を行った.

4.2　方法

比較実験を行うために以下の 2 種類の画像群を用意した.

画像群 1　文献 [2] から抽出した画像群

画像群 2　文献 [2] の特徴を考慮してデータ合成した画像群

画像群 1 は大津の手法を用いて二値化し, 二値化した画像の幅を 512pixel, 高さを 512pixel に変更した.

画像群 1 のアノテーションは手作業で行った. 図 6 のように, バウンディングボックスを用いて行の最小領域をアノテーションした. 本実験では通常の行に加えて注釈, 行末の折り返し, ルビも行とした.

画像群 2 は以下の設定でデータ合成した.

(1) 各平仮名について変体仮名のフォントデータを 100 個ずつ抽出する
(2) フォントデータを二値化する
(3) フォントデータの幅, 高さを一定確率で伸縮させる
(4) 上下左右の余白, 行のフォント幅は文献 [2] を参考に設定する
(5) 文字間隔を 0 から 10 の間でランダムに変化させる

画像群 2 の例を図 7 に示す.

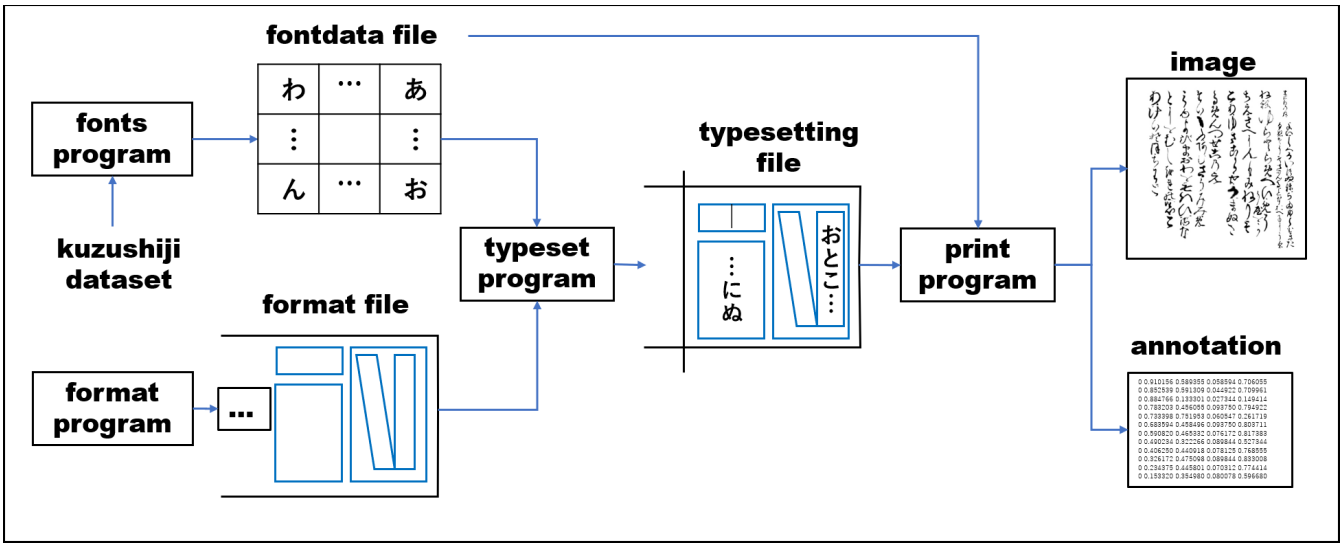

図 5　データ合成システムの概要

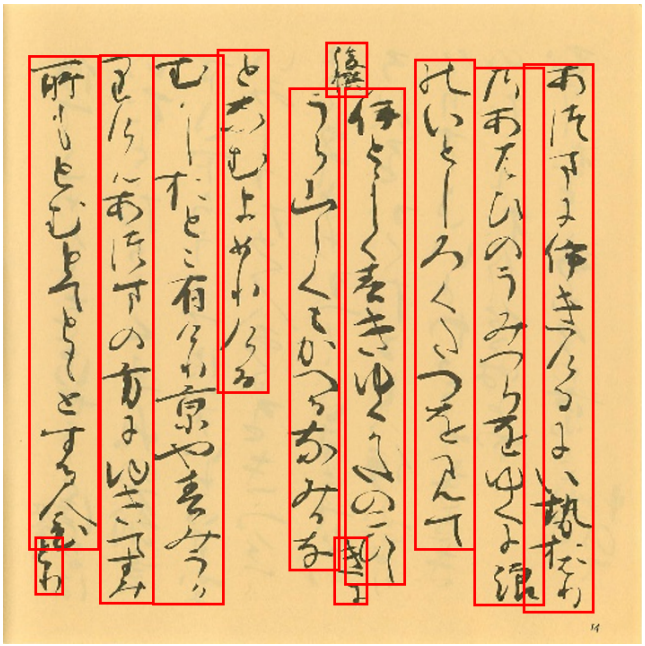

図 6　アノテーション例 (文献 [2] から引用した文書画像に加筆した)

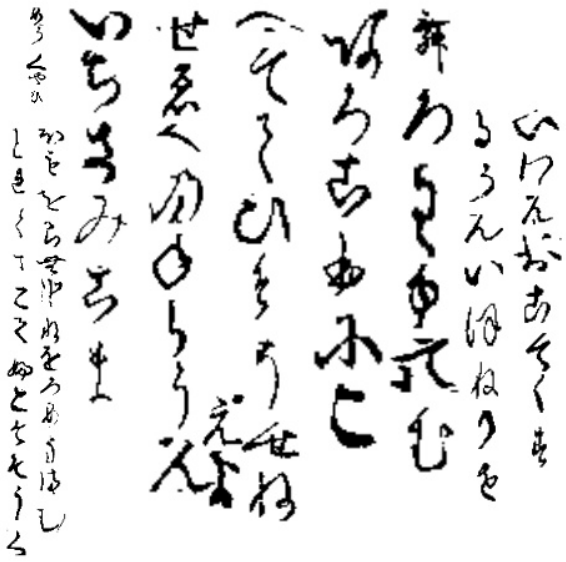

図 7　合成した画像の例

データセットを作成するために, 表 1 のように画像群を分割した. 学習用データセットはモデルの学習時に使用する訓練用データセットと, 学習の結果を評価する検証用データセットから構成されている. テスト用データセットは学習済みモデルの最終的な行切り出し結果を評価するために用いる. 訓練用データセット, 検証用データセットとテスト用データセットの比は 2:1:1 である.

物体検出アルゴリズム YOLO[5] を用いて, 以下の 2 種類のモデルを学習して作成した. 学習率は 0.001, バッチサイズは 16, エポック数は 4000 とした.

モデル 1　画像群 1 で学習したモデル

モデル 2　画像群 2 で学習したモデル

.

モデル 1 のテスト用データセットには画像群 1 を用いた. モデル 2 のテスト用データセットには画像群 1 と画像群 2 を用いた.

本実験では評価指標に Intersection Over Union(IoU) の平均（Mean IoU）を用いる. IoU は正解領域と予測領域の重なりの割合である. 正解領域を A, 予測領域を B とすると IoU は以下の式で表される.

$$Intersection\,Over\,Union\,(IoU) = \frac{|A \cap B|}{|A \cup B|}$$

4.3　結果

表 2 がそれぞれのモデルの Mean IoU である. また, それぞれのモデルの画像群 1 に対する行切り出しの例を図 8, 図 9 に示す.

表 1　学習用データセットとテスト用データセットの画像枚数

	訓練用データセット	検証用データセット	テスト用データセット
画像群 1（文献 [2]）	83	41	42
画像群 2（合成画像）	243	121	122

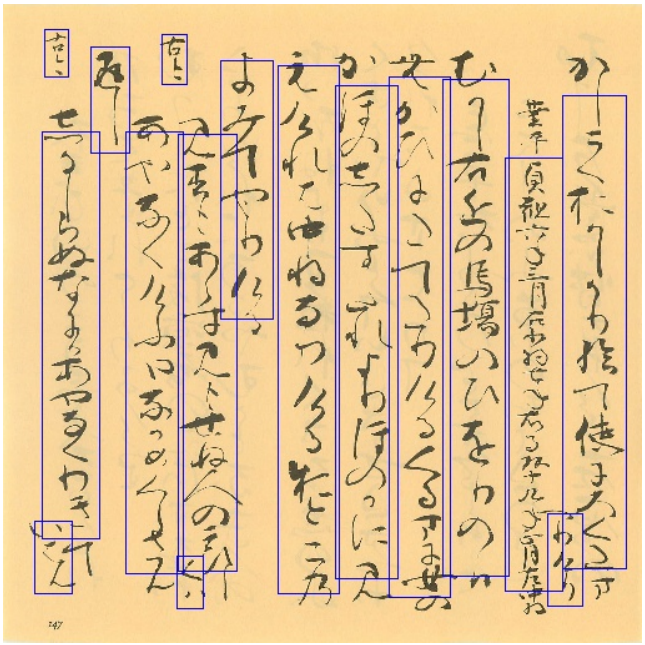

図 8　モデル 1 の行切り出し例

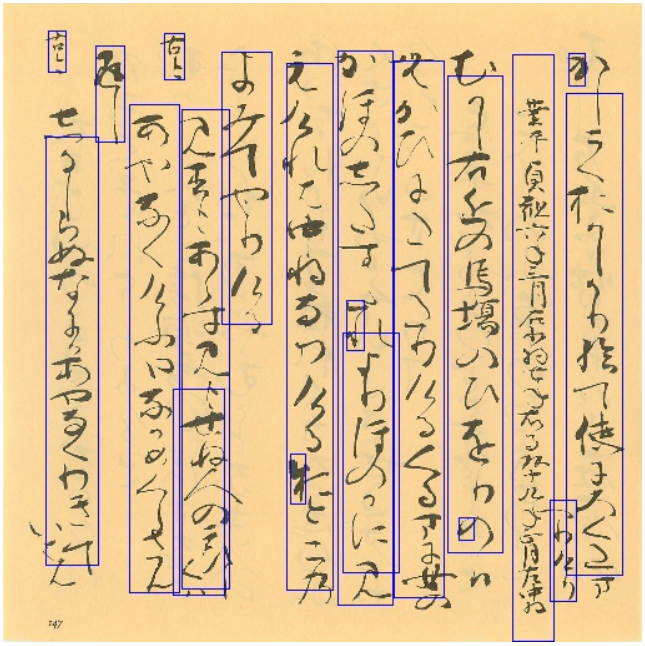

図 9　モデル 2 の行切り出し例

テスト用データセットと学習用データセットが同種のとき, いずれのモデルの Mean IoU も 0.85 を上回った.

テスト用データセットが画像群 1 のとき, モデル 2 の Mean IoU は 0.86 であり, モデル 1 の Mean IoU とほぼ等しい結果となった.

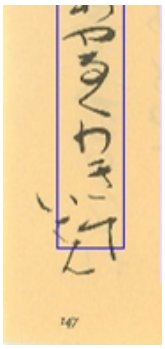

図 10　モデル 2 の行切り出し失敗例 1

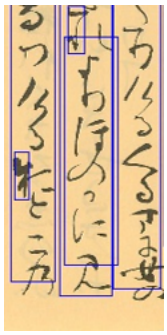

図 11　モデル 2 の行切り出し失敗例 2

4.4　評価

テスト用データセットと学習用データセットが同種のとき, いずれのモデルの Mean IoU も 0.85 を上回り, 概ね正しく切り出すことができた.

テスト用データセットが画像群 1 のとき, モデル 1 とモデル 2 の Mean IoU の差は 0.001 だった. 画像群 2 が文献 [2] の特徴を考慮した結果であると考えられる. しかし, モデル 2 は図 10 のような行末での折り返しを切り出せなかった. また図 11 のように文字の一部や行末の文字列を誤って切り出している場面が確認された. これは行末の文字列を行末での折り返し, 文字の一部をルビとして誤認識したためだとと考えられる.

画像群 1 に対するモデル 1 とモデル 2 の Mean IoU はほぼ同値であったが, 図 10, 11 のようなモデル 2 の行切り出し失敗例は IoU にあまり影響を与えないため, 行切り出しの精度を測定するより適した指標が望まれる.

5.　おわりに

本研究ではアノテーション付き古文書画像の合成方法を提案した. データ合成した古文書画像（合成画像）で学習した物体検出アルゴリズムが, 実際の古文書画像からの行切り出しに有効であるかを確かめるべく実験を行った. 実際の古文書画像で学習したモデルによる行切り出し結果の Mean IoU は, 合成画像で学習したモデルのそれと同程度であった. しかし, データを合成する際には対象画像に合わせてパラメータを設定する必要がある,

今後は, 行切り出しの精度を正確に計測するために,IoU と他の評価指標を併用する, モデルの行切り出し結果に対する後処理を行うなどの課題がある. またデータ合成システムの文字切り出しへの応用も考えられる.

表 2 Mean IoU

	学習用データセット	テスト用データセット	Mean IoU
モデル 1	画像群 1	画像群 1	0.8655
モデル 2	画像群 2	画像群 1	0.8646
モデル 2	画像群 2	画像群 2	0.9029

参考文献

[1] Capobianco, S. and Marinai, S.: DocEmul: A Toolkit to Generate Structured Historical Documents, *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, Vol. 01, pp. 1186–1191 (online), DOI: 10.1109/ICDAR.2017.196 (2017).

[2] 鈴木知太郎：御所本伊勢物語冷泉為和筆宮内庁書陵部蔵影印本，笠間書院 (1994).

[3] Clanuwat, T., Bober-Irizar, M., Kitamoto, A., Lamb, A., Yamamoto, K. and Ha, D.: Deep Learning for Classical Japanese Literature (2018).

[4] Pondenkandath, V., Alberti, M., Diatta, M., Ingold, R. and Liwicki, M.: Historical Document Synthesis with Generative Adversarial Networks, *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, Vol. 5, pp. 146–151 (online), DOI: 10.1109/ICDARW.2019.40096 (2019).

[5] Redmon, J., Divvala, S., Girshick, R. and Farhadi, A.: You Only Look Once: Unified, Real-Time Object Detection (2015).

[6] Sando, K., Suzuki, T. and Aiba, A.: A Constraint Solving Web Service for a Handwritten Japanese Historical Kana Reprint Support System, *Agents and Artificial Intelligence - 10th International Conference, ICAART 2018, Funchal, Madeira, Portugal, January 16-18, 2018, Revised Selected Papers* (van den Herik, H. J. and Rocha, A. P., eds.), Lecture Notes in Computer Science, Vol. 11352, Springer, pp. 422–442 (online), DOI: 10.1007/978-3-030-05453-3_20 (2018).

[7] Sando, K., Suzuki, T. and Aiba, A.: A Constraint Solving Web Service for Recognizing Historical Japanese KANA Texts, *Proceedings of the 10th International Conference on Agents and Artificial Intelligence, ICAART 2018, Volume 2, Funchal, Madeira, Portugal, January 16-18, 2018* (Rocha, A. P. and van den Herik, H. J., eds.), SciTePress, pp. 257–265 (online), DOI: 10.5220/0006709702570265 (2018).

[8] Yamazaki, A., Sando, K., Suzuki, T. and Aiba, A.: A Handwritten Japanese Historical Kana Reprint Support System: Development of a Graphical User Interface, *Proceedings of the ACM Symposium on Document Engineering 2018*, New York, NY, USA, Association for Computing Machinery, (online), DOI: 10.1145/3209280.3229117 (2018).

[9] 山藤一輝，鈴木徹也，相場　亮：手書き変体仮名認識システム-制約解消器の Web サービス化-，第 79 回全国大会講演論文集，Vol. 2017, No. 1, pp. 949–950 (2017).

[10] 山藤一輝，鈴木徹也，相場　亮：手書き変体仮名翻刻支援システム-複数通りの切り出し方を考慮した文字の配置方法-，第 81 回全国大会講演論文集，Vol. 2019, No. 1, pp. 705–706 (2019).

[11] 山藤一輝，山崎敦史，鈴木徹也，相場　亮：手書き変体仮名認識システム-グラフィカルユーザインターフェースの開発-，第 80 回全国大会講演論文集，Vol. 2018, No. 1, pp. 629–630 (2018).

[12] 青池　亨，木下貴文，里見　航，川島隆徳：機械学習のための資料レイアウトデータセットの構築と公開，じんもんこん 2019 論文集，Vol. 2019, pp. 115–120 (2019).

[13] 小木曽智信：中古仮名文学作品の形態素解析，日本語の研究， Vol. 9, No. 4, pp. 49–62（オンライン），DOI: 10.20666/nihongonokenkyu.9.4_49 (2013).

[14] 小木曽智信，小町　守，松本裕治：歴史的日本語資料を対象とした形態素解析，自然言語処理，Vol. 20, No. 5, pp. 727–748（オンライン），DOI: 10.5715/jnlp.20.727 (2013).

[15] 小木曽智信，小椋秀樹，田中牧郎，近藤明日子，伝康晴：中古和文を対象とした形態素解析辞書の開発，情報処理学会研究報告. 人文科学とコンピュータ研究会報告， Vol. 85, pp. D1–D8（オンライン），入手先 ⟨https://ci.nii.ac.jp/naid/110008003480/⟩ (2010).