Vol.2019-CH-119 No.4
2019/2/16

日本古典籍くずし字文書の文字列認識

佐藤 旭[1,a)]　小林 心[1,b)]　Nam Tuan Ly[1,c)]　Nguyen Tuan Coung[1,d)]　北本 朝展[2,3,e)]
中川 正樹[1,f)]

概要：日本古典籍の文書画像を機械認識する試みが始まっている．しかし，文字のくずしや接触，汚れや裏写りなどのために，その難易度は高い．未だに解読されていない古典籍も数多く，解読には多大な時間と労力を伴う．先行研究の機械認識は対象を仮名に限定したが，本稿では，漢字仮名交じり文の認識実験を報告する．方式としては，第 21 回 PRMU アルゴリズムコンテストの変体仮名認識で最優秀賞を獲得した認識モデルを土台にした．現時点では文脈処理は適用していないが，約 74 %の文字認識率を達成した．

Text Recognition of Japanese Classical Documents

1. はじめに

歴史文書のディジタル化は世界的に進展しており，我国でも近世までの古典籍をディジタルで収録・公開する研究開発が盛んになってきている [1], [2], [3]．しかし，画像収録に加えて，解読情報を付与することには多大の労力と時間を要する．そのため，文書画像を機械認識する試みも始まっており，第 21 回 PRMU アルゴリズムコンテストでは，変体仮名の認識が課題に設定された．これは，切り出された文字，特に，変体仮名を対象にしたものであったが，文字のくずしや接触，汚れや裏写り，自由度の高いレイアウト，罫線の存在と接触などのために，機械認識の難易度は高い．

本稿では，第 21 回 PRMU アルゴリズムコンテストの変体仮名認識で最優秀賞を獲得した認識モデルを土台に，漢字仮名交じり文の認識実験を報告する．オープンデータである仮名漢字交じり文字のデータセットから実験対象にする文字列画像のデータを用意し，それを対象画像とした．

[1] 東京農工大学
Tokyo University of Agricultute and Technology
[2] ROIS-DS 人文学オープンデータ共同利用センター
ROIS-DS Center for Open Data in the Humanities
[3] 国立情報学研究所
National Institute of Informatics
[a)] s183892r@st.go.tuat.ac.jp
[b)] dryassy1ac@gmail.com
[c)] namlytuan@gmail.com
[d)] ntcuong2103@gmail.com
[e)] kitamoto@nii.ac.jp
[f)] nakagawa@cc.tuat.ac.jp

2. 関連研究

今回の対象データは日本古典籍の文書画像である．今回の認識モデルは Deep Convolutional Recurrent Network (DCRN) と呼ばれる Convolution Neural Network (CNN) と Bidirectional Long Short Term Memory (BLSTM), Connectionist Temporal Classification (CTC) を用いたモデルである [4]．3 章の認識モデルで詳細に説明するが，CNN で特徴を抽出し，BLSTM で認識を行い，CTC で結合して出力するものである．第 21 回 PRMU アルゴリズムコンテストでは，単独文字（レベル 1），3 文字以下（レベル 2），複数行に渡る場合がある 16 文字以下（レベル 3）のそれぞれ短文，かつ仮名の文字画像を対象としている．ここで DCRN はレベル 2 とレベル 3 で 1 位となり最優秀賞を獲得した．Nam らは，コンテストの評価セットは非公開で使えないことから，公開セットの中から評価セットを用意し，レベル 2 で 68.4 %，レベル 3 で 17.43 %の文字認識率を達成している [5]．

山本らは，既存の OCR と寺沢らの文書画像検索技術 [6] を統合して，古典籍の仮名漢字交じりから文字切出し後の文字認識を実現し，約 80 %の精度を報告している [7]．

さらに，上記のコンテストでは，Neural Network による方式が多く提案され，レベル 1 で 97.2 %，レベル 2 で 87.6 %，レベル 3 で 39.1 %の文字認識率が報告された [8]．

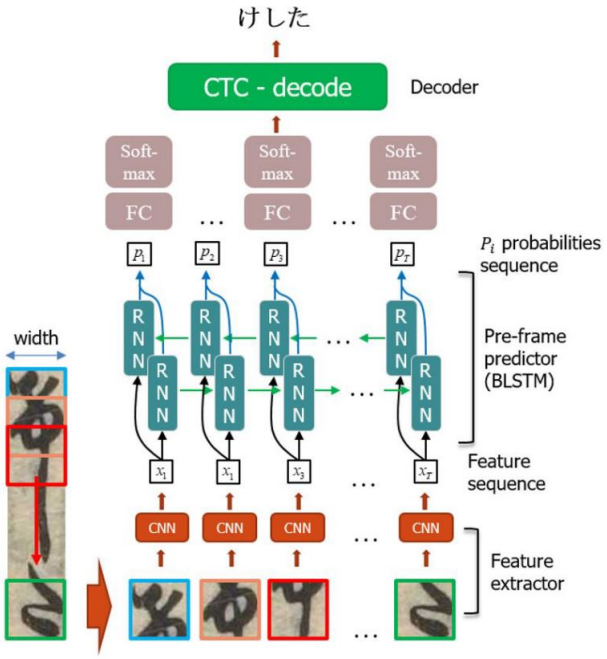

図 1　DCRN のネットワーク構造 [5]

3. 認識モデル

今回使用したモデルは Nam らによって提案された DCRN モデル [5] を今回のデータセット用に拡張したものである．図 1 のように，縦方向に画像を分割し，それをもとに特徴を CNN で出力する．これらの特徴を BLSTM で認識させ，文字のラベルの確率列を出力する．最後に CTC を用い全結合層でデコードして，最尤の文字ラベル列を出力する．CTC 及び BLSTM の認識層については参考文献を参照されたい [5]．

4. データセット

今回用いたデータセットは人文学オープンデータ共同利用センターの HP に公開されている，「日本古典籍くずし字データセット」[9] である．これは，「国文学研究資料館所蔵で日本古典籍データセットにて公開する古典籍，および国文学研究資料館の関係機関が公開する古典籍 15 点の画像データ」[9] に文字のアノテーションを行ったものである．これらは 3,999 種，403,242 文字と，判読できない文字（判読不能文字）や Unicode に含まれていない文字（合略文字）47 種が含まれている．もちろん，すべての文字が均等に含まれているわけではない．表 1 のように，多くのサンプルが含まれている字種から，1 つしかない字種で多種多様である．また，判読不能文字・合略文字には外接矩形が定義されておらず，左上の座標データだけが存在する．我々はこのアノテーションをもとに，次の手順にそって処理を行った．

- 文字のアノテーションをもとに，行に含まれる文字をすべて含む矩形を行として定義する．
- 判読不能文字・合略文字の左上の座標がいずれかの行に含まれていた場合は，アノテーションにおいて，それより上の文字の Unicode とそれより下の文字の Unicode の間に判読不能のコードを挿入する．
- ただし，挿入場所が行の先頭または末尾だった場合は追加しない．
- 以上の処理に問題がないか人間が確認し，ある場合は人力で行を含む矩形を修正する．

処理の例を図 2，切り出しの例を図 3 に示す．3999 種の文字と判読不能文字・合略文字を合わせ，4,046 クラスを含む 25,275 行データとなった．

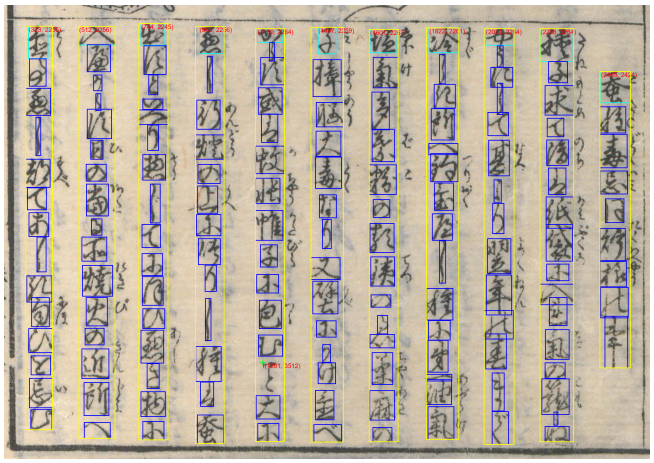

図 2　文書画像とアノテーションの例
水色及び青色は文字の外接矩形，緑色十字は判読不能文字・合略文字，黄色は生成した行の外接矩形を示す．

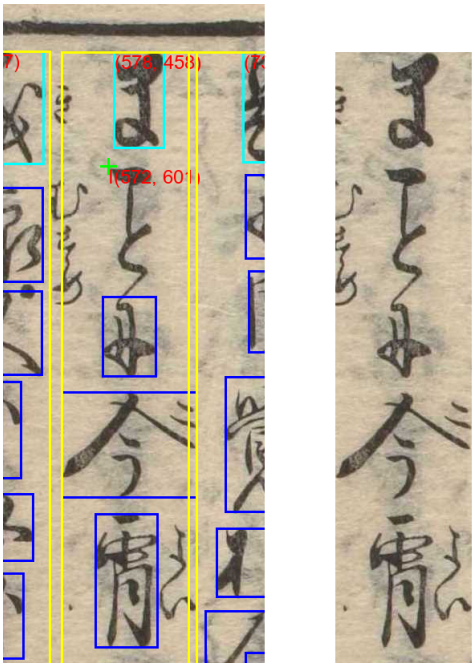

図 3　合略文字を含む行画像の作成例（一部）
左：アノテーション，右：切り出し結果

表 1　データベース内文字数

ID	Unicode	字種	画像データ数
1	U+306B	に	15982
2	U+306E	の	14337
3	U+3057	し	13386
…	…	…	…
3997	U+6E23	渣	1
3998	U+6ABB	檻	1
3999	U+83EA	若	1

5. 実験方法

今回は 5-fold のクロスバリデーションを行っている．対象古典籍を本単位で 5 分割し，4 つを訓練データ，残り 1 つをテストデータとし，検証データは訓練データの 1 割をランダムにサンプルした．

また, 評価方法として，Label Error Rate (LER) と Sequence Error Rate (SER) を用いた．LER は正しいラベルと結果の編集距離の平均を表し，SER は結果が正しいラベルと一致していない割合を表す．どちらも値が低いことが望ましい．今回の編集距離はレーベンシュタイン距離を用いており，式 1 により求められる．同様に SER は式 2 により求められる．今回は学習結果として LER が最も小さい epoch を t-epoch と定義する．

$$\mathrm{LER}(h, S') = \frac{100}{Z} \sum_{(x,z)\in S'} \mathrm{ED}(h(x), z) \qquad (1)$$

$$\mathrm{SER}(h, S') = \frac{100}{|S'|} \sum_{(x,z)\in S'} \begin{cases} 0 & (h(x) = z) \\ 1 & (otherwise) \end{cases} \qquad (2)$$

ここで，x は入力画像，z はラベル，S' はテストセット，h はパターン分類器，Z は S' 内の対象ラベルの長さ，$\mathrm{ED}(p, q)$ は p と q のレーベンシュタイン距離を表す．

6. 実験結果

実験結果を表 2 に示す．ここで，v は検証データ，t はテストデータを表し，それぞれの LER, SER を v-LER, v-SER のように結合して表現している．

表 2　実験結果

データセット	v-LER	v-SER	t-LER	t-SER	t-epoch
1	14.478	68.831	31.280	99.273	19
2	14.683	78.149	26.495	85.895	20
3	14.312	78.812	43.026	89.282	22
4	13.783	75.793	30.281	68.316	20
5	15.624	70.199	32.762	95.609	25
平均	14.576	74.357	32.769	87.675	21.2

すべてのテストには epoch25 回までで LER が最も良い

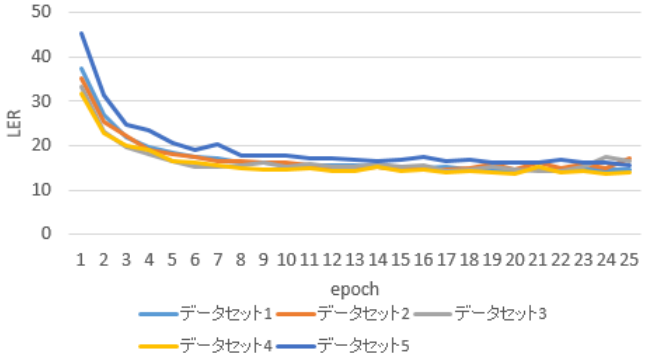

図 4　LER の遷移

図 5　左 3 枚:認識成功画像，右 3 枚:認識失敗画像

epoch（t-epoch）のパラメータを用いた．図 4 は各データセットにおける epoch 毎の LER の遷移を表している．グラフからわかるように，epoch を 25 回以上に増やしても LER が改善する見込みがなく，良くなっても過学習の可能性が高くなると考えたからである．

認識がうまくいった場合と行かなかった場合のラベル例を図 5 として示した．黒字が認識結果，赤字が正しいラベルとの違いを表している．この 6 枚が全データセットだったと仮定すると，LER と SER は式 3，4 よりそれぞれ 5，50 となる．また，旧字体や似ている字種，省略され人間でも認識しづらい文字は結果として誤認識されていることがわかる．

$$\mathrm{LER}(h, S') = \frac{100(0+0+0+2+3+2)}{25+11+15+26+20+27} = 5 \qquad (3)$$

$$\mathrm{SER}(h, S') = \frac{100}{6}(0+0+0+1+1+1) = 50 \qquad (4)$$

7. 考察

LER はある程度の精度を得ているが，SER は低い．これは，行の中のいずれかの文字の認識に失敗していること，行の中には学習データが少ない文字が含まれる場合が多い

図 6 最後の文字「じ」が上の字と被る例

こと，文脈処理を適用していないことによる．また，大きさが不揃いな文字が並ぶと誤認識を起こしやすい傾向も確認できた．サンプル数の少ない字種の文字画像データを人工的に補うこと，文脈処理を適用すること，文字サイズの変動に強い方式または学習データの利用が必要であろう．また，今回のデータベースには**図 6** のような画像もあり，単に縦に矩形として分割して読んでいく方式では対応が難しい．

8. おわりに

本稿では，特徴抽出のための CNN と文字ラベルの確率の列を出力する BLSTM，そして，最尤の文字列を出力する CTC を統合した方式（DCRN）による仮名漢字交じり文字の文字列認識を報告した．第 21 回 PRMU コンテストに比べ，長文であることと, 漢字が混じっていること, かつ文書画像データがクラス数に対し少ないと条件だったが，LER についてはある程度の精度を実現できることを示した．SER を高めるためには，サンプル数の少ない字種の文字画像の追加，人工的な補完，文脈処理の適用，文字サイズの変動に強い方式または学習データの利用が今後の課題である．

謝辞 データベースの画像チェックに協力してくださった中川研究室の森住啓，牛澤葵に深謝する．また，本研究は，ROIS-DS-JOINT 課題番号 027RP2018 の一部補助による．

参考文献

[1] Clanuwat, T., Lamb, A. and Kitamoto, A.: *End-to-End Pre-Modern Japanese Character (Kuzushiji) Spotting with Deep Learning*, 人文科学とコンピュータシンポジウム じんもんこん 2018, pp. 15–20 (2018)
[2] Clanuwat, T., Irizar, B.M., Kitamoto, A., Lamb, A., Yamamoto, K. and Ha, D.: *Deep Learning for Classical Japanese Literature*, Neural Information Processing Systems 2018 Workshop on Machine Learning for Creativity and Design (2018)
[3] 北本 朝展，本間 淳，Tarek Saier: *IIIF Curation Platform*：利用者主導の画像共有を支援するオープンな次世代 *IIIF* 基盤，人文科学とコンピュータシンポジウム じんもんこん 2018，pp.327–334 (2018)
[4] Nam, T.L., Nguyen, T.C., Nguyen, C.K. and Nakagawa M.: *Deep Convolutional Recurrent Network for Segmentation-free Offline Handwritten Japanese Text Recognition*, 14th IAPR International Conference on Document Analysis and Recognition (2017)
[5] Nguyen, T.H., Nam, T.L., Nguyen, C.K., Nguyen, T.C. and Nakagawa M.: *Attempts to recognize anomalously deformed Kana in Japanese historical documents*, The 4th International Workshop on Historical Document Imaging and Processing, pp.31–36 (2017)
[6] 寺沢憲吾, 川嶋稔夫郎: 文書画像からの全文検索のオンラインサービス，じんもんこん 2011 論文集，vol.8, p.329–334，情報処理学会 (2011).
[7] 山本 純子，大澤 留次郎: 古典籍翻刻の省力化：くずし字を含む新方式 OCR 技術の開発，情報管理，Vol.58，No.11，pp.819–827，科学技術振興機構 (2018).
[8] 電子情報通信学会 パターン認識・メディア理解（PRMU）研究会: 第 21 回 PRMU アルコン（オンライン），入手先 ⟨https://sites.google.com/view/alcon2017prmu/コンテスト結果⟩ (参照 2019–01–23).
[9] 人文学オープンデータ共同利用センター：日本古典籍くずし字データセット（オンライン），入手先 ⟨http://codh.rois.ac.jp/char-shape/⟩ (参照 2019–01–23).