H-036

接続部分の特徴に着目したくずし字認識の一手法
A Method of Kuzushiji Character Recognition Using Connections Feature

一色 昂祐‡　村木 祐太‡　小堀 研一‡
Kosuke Isshiki　Yuta Muraki　Kenichi Kobori

1. はじめに

国文学研究資料館は平成26年度より，日本語の歴史的典籍を国際的に共同研究するために「日本語の歴史的典籍の国際共同研究ネットワーク構築計画」に取り組んでいる．

しかし，日本語の歴史的典籍は「くずし字」(変体仮名と筆記体で書かれた文字)で書かれているため，その解読には専門的な知識と多大な時間と労力が必要となる．そのため，くずし字の自動翻刻が試みられている．

くずし字の自動翻刻においては，連綿と呼ばれる 1 文字ごとに線を切らずに続けて書く文字が障害となっている．

そこで本研究では，くずし字で書かれた歴史的典籍の見開き 1 ページを入力として連綿部分に着目し，くずし字特有の文字間の接続部分の特徴を利用して，文字を 1 文字に切り出す．そして，切り出された文字を畳み込みニューラルネットワーク(CNN)により識別することで自動翻刻を行う手法を提案する．

2. 提案手法

提案手法の手順を以下に示す。

(1)　前処理
(2)　行の抽出
(3)　文字領域の統合
(4)　文字領域の切り出し
(5)　CNNを用いた学習器による識別

2.1 前処理

歴史的典籍には紙面の劣化や染み，裏写りなどのノイズが存在している．これらは文字の切り出し，翻刻に影響を与えるため前処理によって除去を行う．

まず，入力画像に対してエッジ保存平滑化フィルタで平滑化を行う．次に，適応的ヒストグラム平坦化を用いて文字領域の強調を行う．そして，メディアンフィルタ差分画像を作成する．最後に wolf の手法[1]による二値化を行うことで，ノイズを除去した二値画像を作成する．

2.2 行の抽出

歴史的典籍には，本文以外に漢字の振り仮名や意図的に書かれた文字でない点や線が存在する場合がある．そのため本文の行のみを抽出する．

まず，二値画像の垂直方向の文字領域画素の数を求め，行が存在する領域ではその値は大きくなるため，平均以上の値を持つ領域を抽出する． 次に，抽出した領域の幅は本文の行においては近い値になるため，幅ごとの領域の数を要素とする配列の最大和の部分配列を求めることで，本文+の行を抽出する．そして，抽出した各行に接する文字領域をその行に属する文字領域とする． 結果を図 1 に示す．

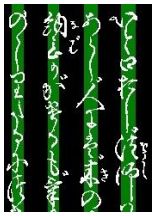

図 1 行の抽出結果

2.3 文字領域の統合

文字領域はひとつの文字領域が分離している場合があるため，文字領域の統合を行う．まず，各文字領域の凸包を求め，凸包同士が重複していれば統合を行う．次に，文字領域は偏と旁で左右に分離しており，凸包が重複しない場合があるため，Y 座標に重複があり，文字領域間の X 軸方向の距離が閾値未満であれば統合を行う．

2.4 文字領域の切り出し

文字領域には図 2 のように複数の文字領域が結合した連綿の文字領域が存在している．そのため，このような文字領域を 1 文字の文字領域に分離することで切り出しを行う．

図 2 連綿の例

まず，文字領域が連綿かどうかの判定を行う．連綿は複数の文字が繋がっているため，高さが大きくなる．そのため，式(1)を用いて文字領域に含まれる推定文字数を求める．

$$N = \left\lfloor \frac{h_i}{\bar{h}} + 0.5 \right\rfloor \tag{1}$$

式(1)において，Nは推定文字数，h_iはある文字領域の高さ，$\bar{h}$は文字領域の平均の高さである．N が 1 の文字領域を切り出し，2 以上の文字領域を連綿と判定する．

連綿の分離には，連綿の文字間の接続部分における以下の特徴を利用する．

①　X 軸方向に他の文字領域が存在しない
②　左下方向への傾斜を持つ
③　線幅の変化が小さい

2.4.1 接続部分の候補領域の抽出

まず①の特徴から， X 軸方向に走査を行い，文字領域を構成する画素の連続領域が 1 か所のみ存在する領域を接続部分が存在する可能性がある領域として抽出する．次に抽出した領域を②，③の特徴を利用するために細分化を行う．

細分化には水戸らの手法[2]を用いる．水戸らの手法は，線分を計算により推定した角度と線幅を用いてセグメンテーションを行う手法である．図 2 から抽出した領域を図 3(a)の赤色部分，水戸らの手法を用いて細分化を行った結果を同図(b)に示す．同図(b)の各領域を接続部分の候補領域とする．また，各領域対して主成分分析を行うことで，候補領域の傾斜方向を求める．

(a) 接続部分の候補領域　(b) 水戸らの手法の結果

図 3 接続部分の候補領域の解析

2.4.2 連綿の分離

2.4.1 項で得られた接続部分の候補領域を用いて連綿の分離を行う．

まず，連綿の文字領域において等間隔に文字が存在していると仮定した場合の推定分離位置を，推定文字数と文字領域の高さから求める．次に，接続部分の候補領域と推定分離位置の Y 軸方向の距離を求める．そして，候補領域ごとに分離の優先度を傾斜角度，長さ，推定分離位置との距離から求め，優先度が最大の接続部分の候補領域で分離を行う．この処理を，全ての文字領域で推定文字数が 1 になる，もしくは接続部分の候補領域がなくなるまで行う．最後に，分離の結果を 1 文字の文字領域として切り出す．図 4 に例を示す．

図 4 連綿の分離の例

2.5 CNN を用いた学習器による識別

2.4 節で得られた文字領域を，CNN を用いた学習器によって識別を行う．学習のデータセットには人文学オープンデータ共同利用センターが公開している機械学習用くずし字データセット[3]を用いた．

3. 実験と考察

「徒然草」，「源氏物語」，「人倫重寶記」の一部を入力画像として，切り出し精度と翻刻精度の検証を行った．結果を表 1 に示す．

表 1 切り出し精度と翻刻精度

	切り出し精度	翻刻精度
徒然草	79.46%	62.94%
源氏物語	77.96%	65.09%
人倫重寶記	75.38%	53.26%

実験結果より，切り出し精度と翻刻精度で「徒然草」と「源氏物語」の精度が入れ替わっていることが分かる．また，「人倫重寶記」においては翻刻精度が低い．これは，含まれる文字種の違いが原因であると考えられる．今回用いたデータセットは平仮名が 49 クラス，漢字は 3832 クラスである．これにより，漢字が多い「人倫重寶記」では翻刻精度が低くなり，平仮名の多い源氏物語では翻刻精度が高くなったと考えられる．そのため，データセットを含めた学習器の調整が必要であると考えられる．

切り出し精度は平均で 77.60%であった．しかし，図 4(a), (b)のように正しく切り出しが行われない場合がある．同図内の水色線が提案手法による切り出し結果，赤線が正しい切り出し位置を示している．

(a)　(b)

図 4 正しく切り出しが行われない例

まず同図(a)は上部の 2 文字が小さいため，一番下の文字領域を切り出したことで，残りの文字領域の文字数が 1 と推定され，上部の文字領域で切り出しが行われていない．そのため，文字数の推定を平均の高さ以外から求める必要があると考えられる．

次に同図(b)は振り仮名が本文と繋がっているため行の抽出において除去されず，本来統合されない文字領域が統合されたことで，接続部分の候補領域が存在せず切り出しが行われていない．そのため，行の抽出において本文と振り仮名の境界を求め，分離する必要があると考えられる．

4. おわりに

本研究では，連綿の接続部分の特徴に着目して歴史的典籍を翻刻する手法を提案した．提案手法では，接続部分の候補領域を解析し，切り出しの優先度を求めることで切り出しを行い，切り出された文字領域を CNN を用いた学習器で識別することで歴史的典籍の翻刻を行った．今後の課題として学習器の調整と，文字領域の文字数の推定手法の変更，繋がった本文と振り仮名を分離することが挙げられる．

参考文献

[1] Wolf, Christian, J-M. Jolion, and Francoise Chassaing. "Text localization, enhancement and binarization in multimedia documents." Object recognition supported by user interaction for service robots. Vol. 2. IEEE, 2002.

[2] 水戸 三千秋, 中島 望夢李, 四海 飛鳥, 尾郷晴美 “主成分分析を応用した線画像の線分認識”, 電気関係学会九州支部第 62 回連合大会講演論文集, 2009.1(2009): pp.434-434.

[3] 機械学習用くずし字データセット．人文学オープンデータ共同利用センター< http://codh.rois.ac.jp/kmnist/>(参照 2020.06.14)

†大阪工業大学 情報科学研究科
〒573-0196 大阪府枚方市北山 1-79-1
TEL.072-866-5301
Email: kobayashi_katuki@ggl.is.oit.ac.jp