

Ⅱ－２ 研究成果の活用と関連研究

Ⅱ－２－１

変体仮名の認識

Hung Tuan Nguyen, Nam Tuan Ly, Kha Cong Nguyen, Cuong Tuan Nguyen, 中川正樹

１、はじめに

木簡の文字の認識と並行して、変体仮名の認識に取り組んだ。すなわち、電子情報通信学会パターン認識・メディア理解研究会（PRMU）が主催した「第21回アルゴリズムコンテスト」⁽¹⁾において、変体仮名を認識する課題に取り組み、難易度の高いレベル2とレベル3⁽²⁾で優秀な認識性能を示し、コンテストに応募されたアルゴリズムのうち、最も優秀と判定されたアルゴリズムを考案した者1件に贈呈される「PRMU アルゴリズムコンテスト最優秀賞」を受賞した（2017年12月16日）。また、情報処理学会人文科学とコンピュータ研究会から、CH 賞を受賞した（2018年1月27日）これに先立ち、11月に開催された4th International Workshop on Historical Document Imaging and Processing（HIP 2017）での発表⁽³⁾に対しても、最優秀論文賞を受賞した。

(1) アルゴリズムコンテストの概要：パターン認識・メディア理解分野の若手研究者・学生の育成および研究会活動の活性化を目的として、毎年開催されています。提示される課題には、代表的・基礎的な研究課題が取り上げられ、応募されたアルゴリズムは、その性能・独創性・処理時間の観点で評価されます。（詳細：<https://sites.google.com/view/alcon2017prmu>）

(2) 第21回アルゴリズムコンテストの課題とレベル：図40～42のように、古典籍画像の指定領域に含まれるくずし字を認識して、コードを出力します。

1. 課題の難易度

外接する長方形に含まれる文字数に応じて課題の難易度を設定しています。レベル1は1文字、レベル2は縦方向の3文字、レベル3は縦横方向の3文字以上の文字を含んでいます。

2. 認識対象の文字

認識対象の文字は変体かな50種程度です。漢字は含みません。

(3) Hung Tuan Nguyen, Nam Tuan Ly, Cong Kha Nguyen, Cuong Tuan Nguyen, Masaki Nakagawa: Attempts to recognize anomalously deformed Kana in Japanese historical documents, *Proc. of the 2017 Workshop on Historical Document and Processing*, pp. 31-36, Kyoto, Japan (11.2017).

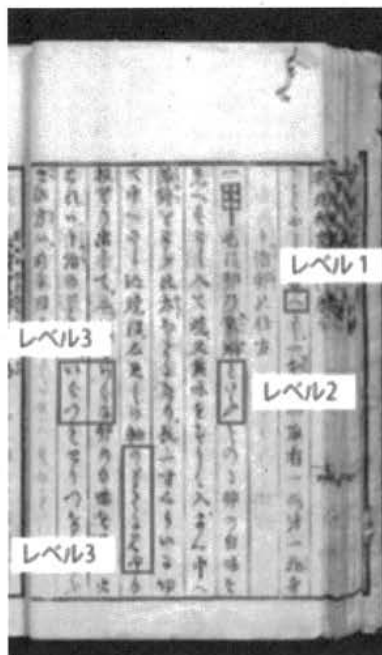


図40 課題の3レベル。レベル1は1文字、レベル2は縦方向の3文字、レベル3は縦横方向の3文字以上

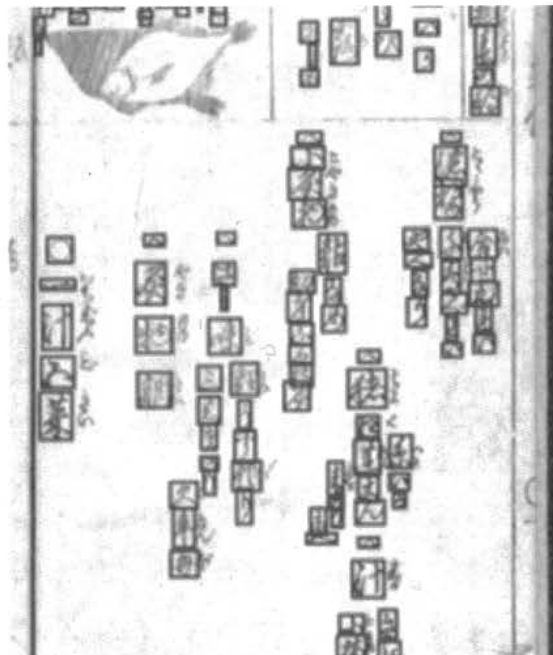


図41 人文学オープンデータ共同利用センターが公開している日本古典籍字形データセット

か[ka]



う[u]



に[ni]



異なる字種にも関わらず似た字形

同一文字の異なる字形

図42 人間でも読みにくい文字の組

2. 方式

レベル2のためには、これまでに提案されているいくつかのニューラルネットワークを検討し、Convolutional Neural Network (CNN: 畳込みニューラルネットワーク)、Bidirectional Long Short-term Neural Network (BLSTM: 双方向長・短期記憶ニューラルネットワーク)、そして、Connectionist Temporal Classification (CTC: コネクショニスト時系列識別法)を3層に組み合わせ、Deep Convolutional Recurrent Network (DCRN: 深層畳込み再帰ネットワーク)を構成する(図43)。第一層では、事前に学習させたCNNによって縦書きのくずし字から特徴の列を抽出し、第二番目の再帰層ではBLSTMによって候補文字と確率の組の列に変換し、3番目の層のCTCで文字列に変換する。

レベル3では、X-Y カット法とボロノイダイアグラムを使って文字行を切り出し、行を

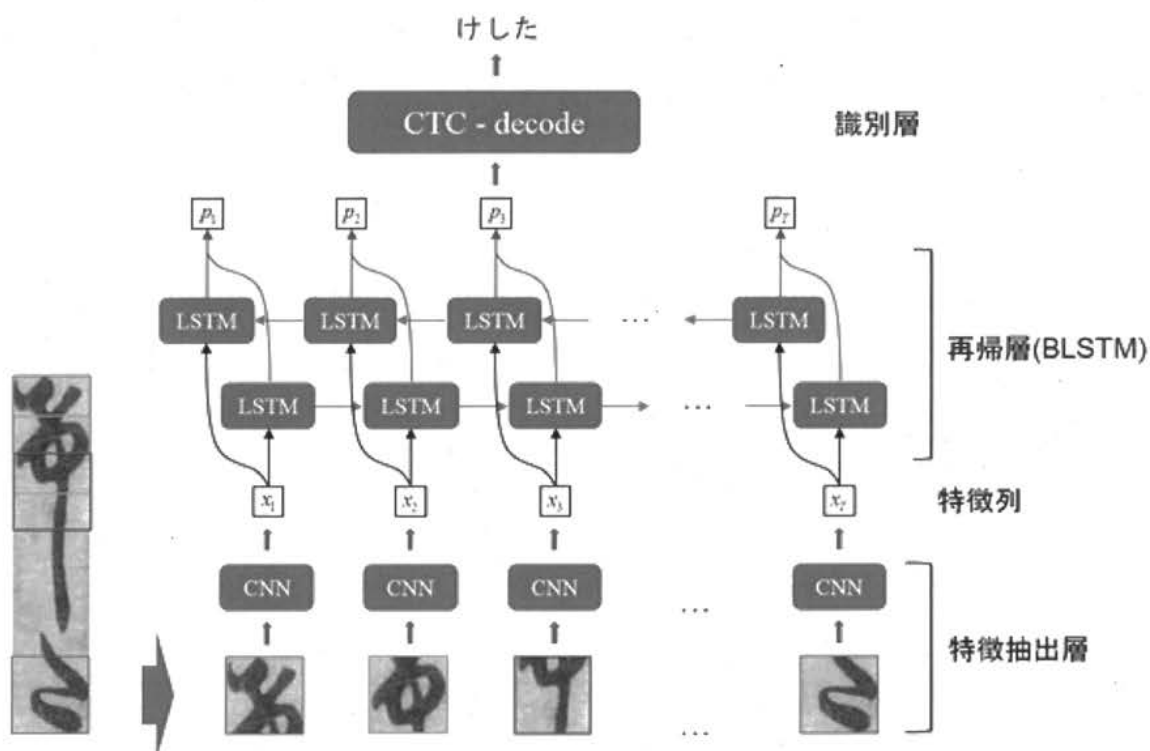


図43 DCRNの構成

一列につないでから⁽¹⁾、レベル2の方式を適用する。X-Yカット法は、縦方向、あるいは、横方向への射影によって空白を見つけて切り、次に、他の方向への射影において空白を見つけて切るという作業を、切れなくなるまで繰り返す手法である。ボロノイダイアグラムは、複数の黒画領域に等距離の白画素内の点をつないでできる分割図である。斜めや複雑に空間がある場合でも切り出せるというメリットがあり、X-Yカット法で分離できないものに、ボロノイダイアグラムを利用する。

3、各レベルにおける結果

・レベル2

- 認識率：87.6% (23チーム中1位)
- 認識時間：1文字あたり2.16秒 (23チーム中10位)

(1) Kha Cong Nguyen, Cuong Tuan Nguyen and Masaki Nakagawa: A segmentation Method of Single- and Multiple-Touching Characters in Offline Handwritten Japanese Text Recognition, *IEICE Trans. on Inf. & Syst.*, Vol. E100-D, No. 12, pp.2962-2972, Dec 2017.

・レベル 3

- 認識率：39.1% (23チーム中 2 位)
- 認識時間：1 文字あたり 0.43 秒 (23チーム中 5 位)

4、審査員の講評

1 点から 5 点の採点で、3 人の評価者から平均して、新規性 4 点、信頼性 5 点、明瞭さ 4.33 点を得ました。明確な文字の切出しを必要としないこと、種々の構成を検討していること、レベル 3 のための複数行の検出と統合、総合的な構成が評価されています。

5、今後の展開

古典を電子的に保存・公開し、手軽に分析できるようにすることは、歴史研究には不可欠であり、古典籍で使われるくずし字を正確に認識する技術開発を進めることで、歴史研究の発展に寄与できると考えられる。

今回、レベル 3 の認識率が低かったのは、1 文字でも誤認識すると、全体として誤認識になるためである。文字列のなかに混同しやすい文字があると、こうした結果になる。これを改善するためには、当時の言語統計から文字と文字のつながりやすさ（文脈）を利用する方法が効果的であり、それによってレベル 3 の認識率はレベル 2 に近くなることが予想される。データをさらに大量に蓄積することで、漢字を含めて、認識率はさらに高まることが期待される。