3Q1-2in1

深層学習による変体仮名翻刻アプリケーション開発の試み

Trial Production of Application Software for Machine Transcription of *Hentaigana* by Deep Learning

早坂 太一[*1]　大野 亙[*1]　加藤 弓枝[*1]　山本 和明[*2]
Taichi HAYASAKA, Wataru OHNO, Yumie KATO, and Kazuaki YAMAMOTO

[*1]国立高等専門学校機構 豊田工業高等専門学校
National Institute of Technology, Toyota College

[*2]人間文化研究機構 国文学研究資料館 古典籍共同研究事業センター
Center for Collaborative Research on Pre-modern Books, National Institutes for the Humanities, National Institute of Japanese Literature

Effective utilization of "Pre-modern Japanese book database" constructed by the project supervised by Center for Collaborative Research on Pre-Modern Texts, National Institute of Japanese Literature, will push forward the development of the inter-field study. It may become an obstacle for the researchers with a little knowledge of classical literature, however, because historical Japanese texts have been written by *Kuzushiji* (*Hentaigana* and cursive kanji). In this article, we report an attempt of recognizing *Hentaigana* by deep learning, which is the artificial intelligence technology regarded throughout the world. Using the convolutional neural networks, we obtained a rate of correct distinction of *Hentaigana* in several pre-modern texts in open database. Furthermore, we developed the WWW software application to recognize *Hentaigana*.

1. はじめに

近年，くずし字に関する研究が注目される契機となったのは，国文学研究資料館により平成 26 年度より開始された「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」[1]である．この計画では，研究基盤整備として，約 30 万点の歴史的典籍を画像データ化し，既存の書誌情報データと統合させた「日本語の歴史的典籍データベース」の構築を行うことになっている．あらゆる分野の書籍が含まれる，膨大な画像データを有効活用できれば，例えば，津波や噴火などの天変地異の歴史を教訓とした防災研究のように，人文科学のみならず，自然科学系分野を融合させた研究の展開も期待される．しかしながら，いかに資料が集積されたとしても，多くの研究者にとっては，それらに書かれている文字が「くずし字」であることが障壁となる．

現行のくずし字翻刻に関する研究の中で，コンピュータ技術を利用した研究は，最も先行研究の蓄積があり，進捗度の大きい分野であると考えられる．特に，凸版印刷株式会社による「くずし字を高精度でテキストデータ化する OCR 技術の開発」[2]は記憶に新しい．同社が提供する「高精度全文テキスト化サービス」を，寺沢ら[3]が開発した文書画像検索システムと組み合わせることで，くずし字で記されている古典籍のOCR技術を開発したものである．このシステムについては，国文学研究資料館の協力の下で動作検証が行われている．テキストデータ化済みの文献を，OCR処理に用いるくずし字データベースとして使用することで，くずし字で記された文献を 80%以上の精度でテキストデータ化することが可能であることが発表されている．

本研究では，様々な分野で導入が進められつつある深層学習(deep learning)を利用した，くずし字翻刻のための人工知能を構築することを目的としている．深層学習は，一度モデルを構築しさえすれば，翻刻に要する時間はごく僅かである．また，学習に用いる文字画像を多数用意する必要はあるが，学習後のモデルには，それぞれの典籍やそれらが書かれた時代で異なる可能性のあるくずし字の特徴が反映されているため，OCR技術のように，翻刻の際に膨大なデータベースを用意する必要はない．つまり，人工知能技術の導入によって，一般的に普及している携帯情報端末でも動作する，小規模なアプリケーション・ソフトウェアとして，「いつでも／どこでも／誰でも自動翻刻」を実現することが可能になると考えられる．

連絡先：早坂太一，豊田工業高等専門学校 情報工学科，〒471-8525 愛知県豊田市栄生町 2-1，TEL(0565)36-5861，FAX(0565)36-5926，hayasaka@toyota-ct.ac.jp

2. 深層学習による変体仮名の認識

2.1 データセット

本研究では，変体仮名画像を，それぞれ平仮名「あ」「い」…「ゑ」「を」「ん」の 48 クラスに分類する学習を行った．ここで，濁点が画像中に含まれていても，分類上は考慮しない．

変体仮名を一文字ずつ，アスペクト比を保持しつつ 64×63 ピクセルの大きさにリサイズし，ネガ・ポジを反転した JPEG 形式のグレイスケール画像として用意し，学習用，学習途中のテスト用，および学習後のテスト用にそれぞれ分類した．学習に用いるデータとして，『五體字類』[4]の 1,473 文字，『和翰名苑』仮名字体データベース[5]の 3,265 文字，および正保4年(1647)に出版された『古今和歌集』[6]から 3,933 文字，『日本古典籍字形データセット』[7]から 53,254 文字，計 61,925 文字の変体仮名画像を用意した．学習途中のテストに用いるデータは，慶長年間頃に出版された『平治物語』巻一[8]から最初の 150 字の変体仮名画像を，学習後のテストに用いるデータは，承応3年(1654) に出版された『源氏物語』桐壺[9]から 10,026 字の変体仮名画像を，それぞれ切り出して利用した．

なお，Data augmentation のため，後述するネットワークモデルへは，62×62 ピクセルに切り取られたものが入力される．

2.2 ネットワークモデル

本研究では，文献[10]に採用されている Convolutional Neural Network (CNN) 構造 M5 および M6 を参考にし，三つの畳み込み層と二つの全結合層による CNN モデルを，繰り返し回数 450,000 回で事前学習させた後，そのネットワークを初期解として，四つの畳み込み層と二つの全結合層による CNN モデルを，繰り返し回数 310,000 回で学習させたものを採用した．

2.3 数値解析環境

本研究における一連の数値計算は，代表的な深層学習用ライブラリである Caffe[11]を用いて行われた．GPGPU (General-Purpose computing on Graphics Processing Units) による計算機環境として，OS は Ubuntu 14.04，CPU は Intel Core i7 6950X 10core/20thread 3.0GHz，GPU は nVidia GeForce GTX 1080 8.0GB を搭載したパーソナルコンピュータであるトーワ電機 GU-1100 を利用した．

2.4 認識結果

学習した CNN に，テストデータの変体仮名を入力し，認識させた結果を表 1 に示す．表 1 において，「第一候補」とは，softmax 関数で計算される平仮名の認識確率が最上位であったもの，「10%以上」とは，認識確率は最上位ではないが，その値が 10%以上であったものの割合を，それぞれ示している．

表 1. CNN による変体仮名の認識結果

	第一候補	*10%以上*
『平治物語』巻一	91.3%	2.0%
『源氏物語』桐壺	95.2%	1.2%

『源氏物語』桐壺 10,026 字のデータの中で，「第一候補」および「10%以上」を合わせて 85%未満の認識率であった平仮名は「え」(30.3%)，「こ」(82.2%)，および「せ」(82.6%)であった．特に低い「え」については，他の仮名と混同されている傾向（例えば「ん」と認識される率が 37.1%）があった．

3. WWWアプリケーションの実現

古典籍の画像データを読み込み，マウス等で選択された1文字分の変体仮名を翻刻する WWW アプリケーションを作製した．スマートフォンにおけるブラウザ(Safari)画面の例を図 1 に示す．

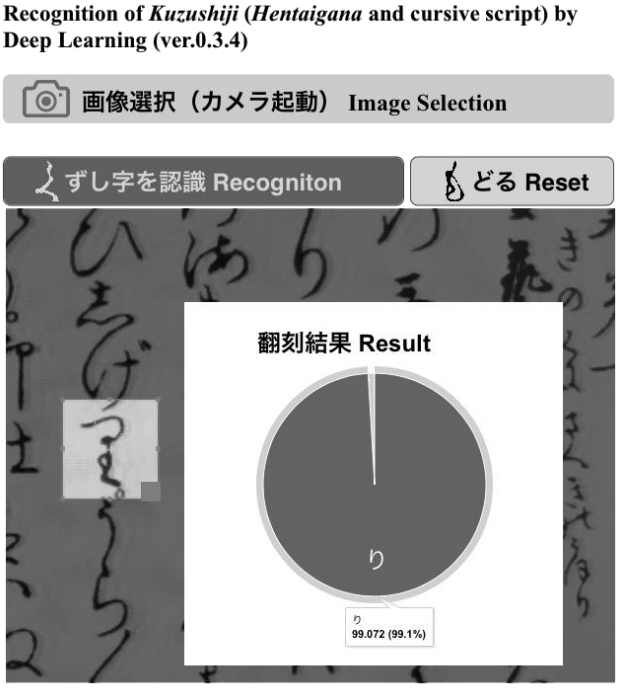

図 1. WWWアプリケーションによる変体仮名翻刻の例

読み込まれた画像に対し，openCV 2.4 を利用して画像処理を施し，学習された CNN に入力することで，平仮名ごとの認識確率が円グラフとして表示される．プログラミング言語は java script および python2.7 を，API として jQuery 3.1.1 (Cropper 2.3.4 プラグインを含む)および Google Chart を使用した．

WWW サーバのハードウェアとして Apple Mac Mini を用い，GPU ではなく，CPU による演算を行わせた．表示については，クライアント側の計算機環境に依存するが，サーバ側で1文字あたりの認識にかかる時間は約 0.4 秒であった．高性能なハードウェアや GPGPU を利用しなくとも，十分な演算速度による翻刻が実現できることが確認できた．

4. むすび

本研究の成果は，海外を含む様々な地域および分野の研究者が，日本に膨大に残る歴史的典籍を判読することを支援する「夢の技術」へと進展していくと考えられる．このことは，日本の歴史的典籍の海外における利用価値を高めることにも繋がる．また，研究者のみならず，一般の人々でも，本研究の成果を利用して，歴史的典籍に記された知識の遺産を有効活用することが期待される．このように，持続可能な社会を実現するためにも，本研究が果たす役割は少なくないと考えられる．

謝辞

本研究は JSPS 科研費 JP16K02433 の助成，および平成 28 年度内藤科学技術振興財団研究助成を受けたものです．

参考文献

[1] 国文学研究資料館：歴史的典籍に関する大型プロジェクト，https://www.nijl.ac.jp/pages/cijproject/，2015 年 10 月 14 日参照．

[2] 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, 2015.

[3] 寺沢憲吾，川嶋稔夫：文書画像からの全文検索のオンラインサービス，人文科学とコンピュータシンポジウム論文集，情報処理学会シンポジウムシリーズ，vol.2011, no.8, pp.329-334, 2011.

[4] 法書会編：五體字類，http://www.let.osaka-u.ac.jp/~okajima/PDF/5tai/, 2015 年 11 月 12 日参照．

[5] 岡田一祐：『和翰名苑』仮名字体データベース，https://kana.aa-ken.jp/wakan/, 2016 年 8 月 16 日参照．

[6] 人文学オープンデータ共同利用センター：日本古典籍データセット（国文研所蔵），二十一代集，http://codh.rois.ac.jp/pmjt/book/200007092/, 2017 年 2 月 15 日参照．

[7] 人文学オープンデータ共同利用センター：日本古典籍字形データセット（国文研所蔵・CODH 加工），http://codh.rois.ac.jp/char-shape/, 2016 年 12 月 8 日参照．

[8] 国立国会図書館：国立国会図書館デジタルコレクション 平治物語，http://dl.ndl.go.jp/info:ndljp/pid/2544708, 2016 年 1 月 14 日参照．

[9] 人文学オープンデータ共同利用センター：日本古典籍データセット（国文研所蔵），源氏物語，http://codh.rois.ac.jp/pmjt/book/200003803/, 2017 年 2 月 15 日参照．

[10] Y. Zhang: Deep convolutional network for handwritten Chinese character recognition, CS231n Course Project Reports, Stanford University, 2015.

[11] Y. Jia, *et al.*: Caffe: Convolutional architecture for fast feature embedding, arXiv preprint arXiv:1408.509. 2014.