

くずし字の検出および認識を行う組み込みシステムの開発

竹内 正広・早坂 太一・大野 亙 (国立高等専門学校機構 豊田工業高等専門学校)
加藤 弓枝 (鶴見大学) 山本 和明 (国文学研究資料館) 石川 徹也 (TRC-ADEAC 株式会社)

国文学研究資料館により整備されている日本語の歴史的典籍の画像および書誌情報の統合データベースを有効活用できれば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が、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。本研究では、深層学習を用いてくずし字の自動翻刻用 WWW アプリケーションとして実装されている学習モデルを応用し、小型で比較的安価なシングルボードコンピュータ Raspberry Pi を用いて開発されたくずし字の自動領域検出および認識を行う組み込みシステムについて報告する。本システムはスマートフォンを持ち込むことができない小・中学校や普段モバイル機器を持ち歩かない高齢者の方々がくずし字に触れたいという場面で支援ツールとして活躍することが期待される。

Development of Embedded System for Extraction and Recognition of *Kuzushiji*.

Masahiro Takeuchi / Taichi Hayasaka / Wataru Ohno (National Institute of Technology, Toyota College),
Yumie Kato (Tsurumi University), Kazuaki Yamamoto (National Institute of Japanese Literature),
and Tetsuya Ishikawa (TRC-ADEAC, Inc.)

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 *Kuzushiji* by deep learning. Using Raspberry Pi which is the low-cost small single board computer, we developed the embedded system with enough recognizing rate and speed. In addition, manifold images of *Kuzushiji* written in a text image of the spread can be extracted. Our system will be effective for school children or elderly people not possessing mobile devices who want to come in touch with *Kuzushiji*.

1. まえがき

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データベースの構築を行っている。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、例えば津波や噴火等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、コンピュータ技術によるくずし字自動翻刻に関する研究は最も先行研究の蓄積があり(例えば[2][3])、進捗度の大きい分野であると考えられる。著者らの研究グループは深層学習を用いたアプローチの先駆け[3]としてくずし字の自動翻刻の試みを行っている。本稿では、小型で比較的安価なシングルボードコンピュータ Raspberry Pi に深層学習

で得られたモデルを組み込んだくずし字の検出および認識を行うシステムについて報告する。

2. 人工知能によるくずし字翻刻

コンピュータ技術によるくずし字自動翻刻に関する研究に分類される本研究において用いる深層学習はヒト脳内における多数の神経細胞による情報のやりとりを数式によりモデル化した階層型ニューラルネットワークが基になっている。深層学習により翻刻を行うモデルを構築するには GPGPU (General-Purpose computing on Graphics Processing Units) といった計算機技術の導入を必要とするが、一度モデルを構築しさえすれば階層型ニューラルネットワークと同様に自動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古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、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、クラウドコンピューティングや第 5 世代移動通信システム (5G) に頼

表1 学習データの詳細
Table 1 Details of trained data for CNN.

	ダウンロード		手作業で抽出			計	文字種類数	1文字あたりの種類数
	CODH文字[5]	和翰名苑[7]	ADEAC[6]	CODH書籍[8]	五體字類[9]			
変体仮名 (割合)	424,751 95.68%	3,265 0.74%	4,911 1.11%	9,511 2.14%	1,473 0.33%	443,911 62.47%	77	5,516
漢字他 (割合)	257,813 96.65%	0 0.00%	6,723 2.52%	2,207 0.83%	0 0.00%	266,743 37.53%	4,573	56
計 (割合)	682,564 96.05%	3,265 0.46%	11,634 1.64%	11,718 1.65%	1,473 0.21%	710,654	4,650	147
文字種類数	4,604	48	1,335	417	48	4,650		

ることのない、一般的に普及している携帯情報端末やパーソナルコンピュータ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。

3. 深層学習によるくずし字認識モデルの学習

本研究では、深層学習で得られたくずし字認識モデルを用いる。このモデルは4層の畳み込み層と3層の全結合層から構成されるCNN(Convolutional Neural Network)である。くずし字を一文字ずつ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を入力とし、16ビットのUnicode(65,535クラス)に分類する学習を行った。

本研究における一連の数値計算は画像認識用として代表的な深層学習用ライブラリであるCaffe[4]を用いて行われた。計算機環境として、OSはUbuntu 14.04 LTS, CPUはIntel Core i7 6900K 8core/16thread 3.2GHz, 主メモリは128GB(16GB×8) DDR4-2133, GPUはnVidia GeForce 1080Ti 11GBを搭載したワークステーション G-DEP DeepLearning Boxを利用した。

学習には4,650種類の710,654文字(平均147文字/種類)のくずし字画像データ(CODHオープンデータ[5]およびデジタルアーカイブシステムADEAC[6]からの提供)を用いた。そのうち変体仮名(Unicode x3042~x3093)は77種類の443,911文字、漢字・カタカナ・記号等は4,573種類の266,743文字である。学習データの詳細を表1に示す。学習には用いていないテストデータである『源氏物語絵巻』桐壺[10]の11,286文字(変体仮名10,026文字、漢字他1,260文字)に対して認識精度が第一位であった文字が、あらかじめわかっている翻刻結果と同じであった割合(認識率)は93.3%(変体仮名95.7%,漢字他74.0%)であ

表2 『源氏物語絵巻』桐壺[10]における認識率の低い変体仮名

Table 2 Hentaigana with lower recognition rates in Genji Monogatari.

仮名	文字数	認識率	仮名	文字数	認識率
ぜ	4	25.0%	る	248	81.5%
え	89	41.6%	み	176	83.0%
ぐ	40	70.0%	こ	212	84.0%
せ	128	79.7%	ぞ	21	85.7%
ご	30	80.0%	ゑ	9	88.9%
ゐ	16	81.3%	ゆ	47	89.4%

った。認識率が90%未満であった変体仮名を表2に示す。濁音以外については著者らの先行研究[3]よりはるかに上回る改善が図られている。また、濁点の有無と認識率とに相関は見られなかった。

漢字・記号等について頻出する文字種に対する認識率を表3に示す。表3における12文字種で全体の約64%を占め、平均認識率は全体とほぼ同じ73.2%であった。変体仮名と比較して認識率が物足りない原因の一つは文字の種類数に対して学習に用いる入力サンプル数が不足していることであると考えられる。

学習したモデルを利用して、古典籍の画像データを読み込み、マウスや指で選択された1文字分の変体仮名を翻刻するWWWアプリケーションが2016年より公開されている[3](<http://vpac.toyota-ct.ac.jp/kuzushiji/>)。ブラウザ画面の例を図1に示す。画像ファイルの形式にこだわることなく、スマートフォン等で撮影した画像で手軽にくずし字を調べることができる。WWWサーバとしてApple Mac Miniを用い、GPUではなくCPUによる認識を行っている。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文字あたりの分類にかかる時間は約0.4秒であった。

表 3 『源氏物語絵巻』 桐壺[10]における頻出率の高い漢字および記号に対する認識率

Table 3 Recognition rates of kanji and symbols with more frequent appearance in *Genji Monogatari*.

漢字等	文字数	認識率	漢字等	文字数	認識率
給	217	47.0%	思	26	84.6%
御	160	98.1%	聞	23	65.2%
ゝ	121	48.8%	世	22	100.0%
人	75	100.0%	宮	22	95.5%
心	50	98.0%	く	21	66.7%
と	44	77.3%	侍	21	81.0%

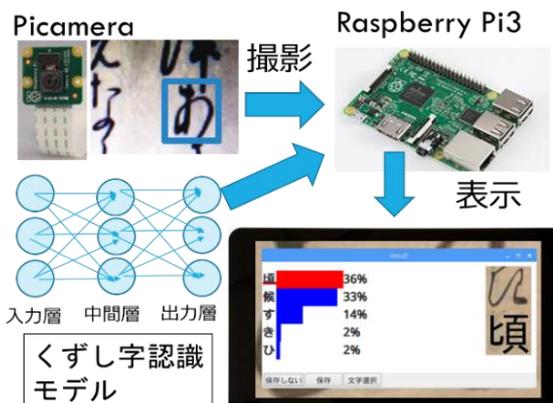


図 2 開発システムの概要図

Figure 2 Overview of the developed system.

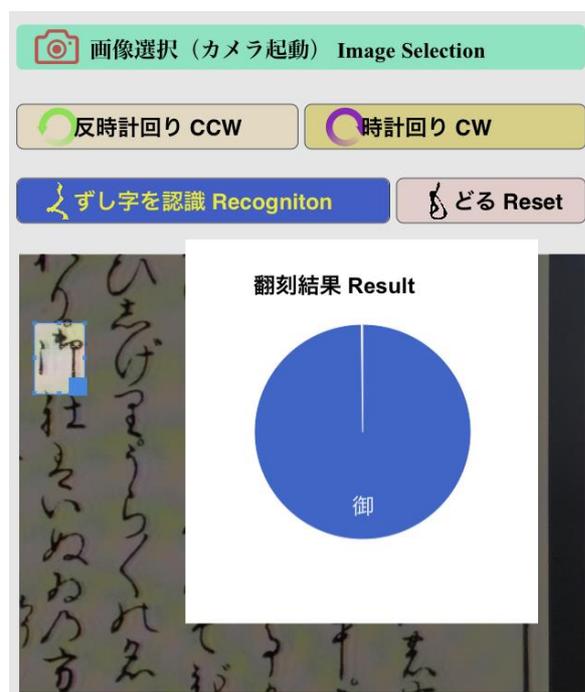


図 1 CNN によるくずし字認識用

WWWアプリケーションのスクリーンショット例

Figure 1 Screenshot example of WWW application software to recognize *Kuzushiji* by trained CNN.

4. Raspberry Pi を用いたくずし字自動抽出・認識システム

4.1 システムの概要

図 2 に本研究で開発したシステムの概要を示す。ハードウェアとして教育用シングルボードコンピュータ Raspberry Pi Model 3B を用いる。ディスプレイの背面に設置した Raspberry Pi 用カメラモジュールから対象の古典籍画像を撮影する。

深層学習により得られた、本研究で用いるモデルはくずし字 1 文字に対して認識を行うものであるため、Python3 および OpenCV3 を用いて認識する文字の選択や、領域検出のための画像処理な

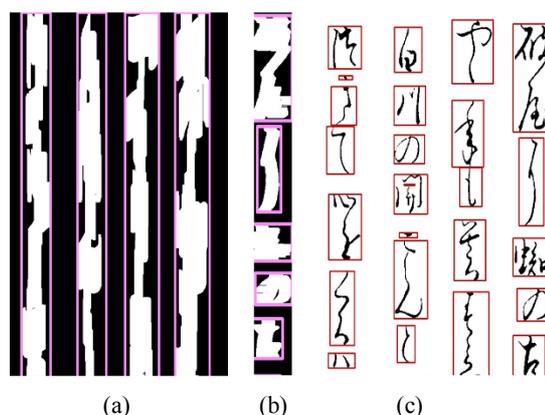


図 3 文字領域の検出過程と結果例

Figure 3 Extraction process of *Kuzushiji* areas.

どを行う。認識は機械学習用ライブラリ TensorFlow Lite を用いて行われ、結果は 7 インチのタッチスクリーン付きディスプレイに表示される。

4.2 くずし字領域の検出

古典籍画像に対する前処理には OpenCV3 を利用した。1 文字の選択では、タッチスクリーンに触れてから離すまでの x-y 座標を取得した後その領域を 62×62 ピクセルにリサイズし、2 値化を行った後に、学習モデルに入力する。しかしながら、1 文字ずつ手作業で抽出する方法では手間がかかるため、古典籍画像から自動でくずし字領域を検出することを試みた。

入力画像について、OpenCV3 のモルフォロジ処理を利用した。画像に対して、図 3(a)に示すように縦方向に対して膨張処理を行い、上下の字をつなげることで、1 行ごとの領域検出を行う。次に、図 3(b)に示すように横方向に膨張処理を行い、それぞれの行から 1 文字ずつの領域を検出する。図 3(c)に検出例を示す。

この他に、図 4 に示すように極端に小さい領域は除外すると共に、領域候補ごとに認識を行い、

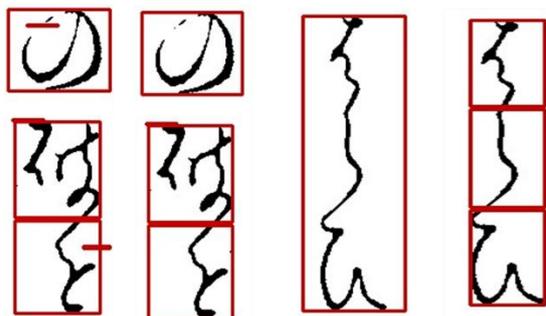


図4 検出した文字領域の分割・除外・統合の例
Figure 4 Examples of splitting / excluding / merging extracted Kuzushiji areas.

精度の高いものを採用することで、領域の統合または分割を行っている。100文字あまりの古典籍画像に対して約10秒でくずし字領域の検出が可能である。

4.3 モデルの量子化

本研究では、Raspberry Piによる推論用の深層学習用ライブラリとしてモバイル端末や組み込み機器向けのTensorFlow Liteを使用した。TensorFlow Liteでは、学習済みモデルの重み等を、浮動小数点形式から固定小数点形式に量子化し、モデルサイズを小さくすることで、処理能力がやや劣る機器での演算および読み込みにかかる時間を短縮することができる。本研究では、入力のみ32ビットの浮動小数点形式とし、他のパラメータは8ビットの固定小数点形式に量子化した。その結果、認識精度は量子化前と同程度のままモデルサイズは約1/4に軽量化され、1文字あたりの認識にかかる時間は約0.4秒とすることができた。

4.4 実行例

本システムでは、GUIアプリケーションから、カメラによる古典籍画像撮影、ファイル読み込み、文字領域抽出、手動での切り取り、自動認識のそれぞれの処理を行うことができる。

カメラモジュールは焦点距離を約30cmとし、3秒間のプレビュー後に古典籍画像を撮影するよう設定した。保存済みの画像ファイルの読み込みを行うことも可能である。文字認識後は図5に示すように結果を表示し、認識結果の保存や候補となる5文字の中から適切な認識結果を選択することも可能である。その後、古典籍画像中のくずし字を再びタッチすることで、認識結果を何度でも参照することができる。また、認識結果および検出したくずし字の座標は、テキストデータとして翻刻された対象の画像を選択することで、画像と同時に読み込まれ、いつでも再参照することができる。

認識を行った文字および検出を行った文字の確認をする場合は、ディスプレイを2本指でタッチすることで、図6に示すように、認識を行った

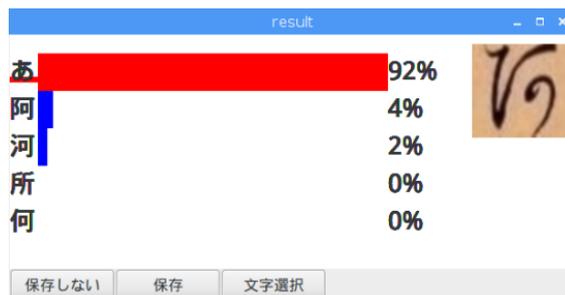


図5 開発システムによるくずし字認識結果の表示例
Figure 5 Example of recognition results of Kuzushiji by the developed system.



図6 開発システムによるくずし字の自動抽出および認識の実行例（河本家住宅保存会・島根大学附属図書館所蔵 河本家古典籍「山下水」[6]より）

Figure 6 Example of automatic extraction and recognition of Kuzushiji areas by the developed system.

文字は翻刻されて緑色枠で、文字検出を行った結果の領域は青色枠で囲まれて表示される。

検出した領域に対してユーザーが編集を行いたい場合は、図6に示す翻刻画像の左下にある領域編集ボタンにより可能である。図7に示すように変更したい領域を領域がすべて緑色枠に収まるように選択する。その後、図8に示すように領域の再選択を行い、確定ボタンを押すことで任意の領域に変更することができる。

5. システムの評価

デジタルアーカイブシステム ADEAC [6]から提供のあった古典籍「山下水」(2,712文字)に対して本システムによるくずし字領域の自動抽出を行ったところ、55.3%の精度を得た。また、正しく抽出できた文字領域に対する認識精度は62.1%であった。

また、令和元年6月6日に新潟市で開催された人工知能学会全国大会(第33回)インタラクティブセッションにおいて、20~60歳代の男性17名および女性1名に対し、本システムに関する5段階評価のアンケートを行った。その結果を図9に示す。デモンストレーションの際に本システムの操作は主に著者らが行ったため、操作性について

は無回答が多かったが、その他の項目については80%前後の方々から4以上の高評価を頂いた。

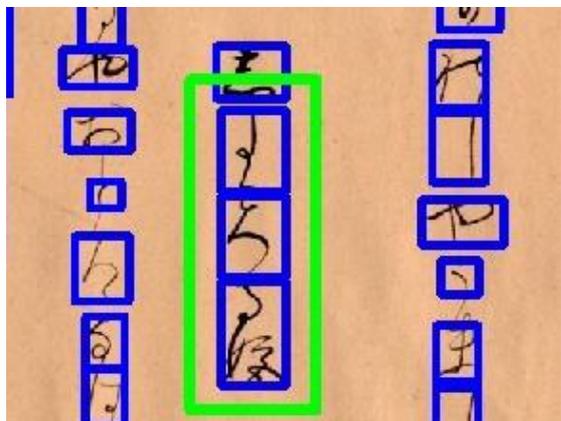


図7 開発システムにおける編集領域の選択例
Figure 7 Editing area selection in the developed system.

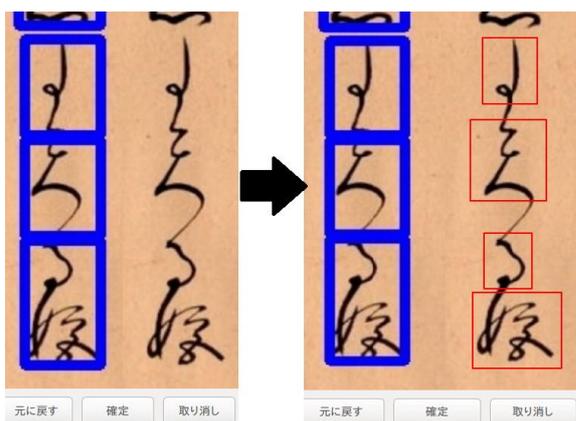


図8 開発システムにおける検出領域の再選択の例
Figure 8 Extracted area re-selection in the developed system.

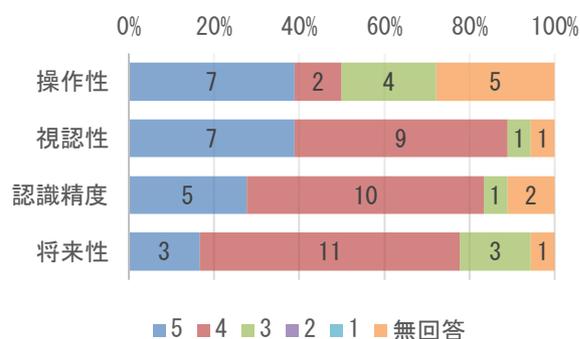


図9 開発システムに対するデモンストレーション後のアンケート結果

Figure 9 Questionnaire results after the demonstration of the developed system.

6. あとがき

本研究では、カメラで撮影した古典籍画像等からくずし字の自動検出および認識を可能とするシステムを開発した。Raspberry Piのような小型で比較的安価なシングルボードコンピュータでも素早い認識を行わせることができた。

首都圏にある学校等と異なり、地方の小中学校やコミュニティ施設には、多人数が利用できるWi-Fi等のインターネット環境はそれほど整備されていないため[11]、古典籍に関する生涯学習等の講座を開催したい場合には、そうした環境を事業提供者が用意することは相応の時間と費用がかかる。そうしたネットワーク環境のない場所や普段モバイル機器を持ち歩かない高齢者の方々でも、くずし字に触れたい場面で支援ツールとして活躍することができるよう、操作性やくずし字領域の検出精度をより高めることが今後の課題である。

謝辞

5年間にわたる本研究の実施にあたり、古典籍から数万にわたるくずし字の抽出にあたっていただいた豊田工業高等専門学校の学生の皆さんに感謝いたします。また本研究は JSPS 科研費 JP16K02433 および TRC-ADEAC 株式会社からの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/>, (参照 2015-10-14).
- [2] T. Clanuwat, M. Bober-Irizar, A. Kitamoto, A. Lamb, K. Yamamoto, and D. Ha. “Deep learning for classical Japanese literature”. arXiv:1812.01718, (参照 2019-07-03).
- [3] 早坂太一, 大野互, 加藤弓枝, 山本和明. 深層学習による変体仮名の翻刻および WWW アプリケーション開発の試み. 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, 2016, vol.2016, no.2, pp.7-12.
- [4] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, “Caffe: Convolutional architecture for fast feature embedding.”. arXiv:1408.509, (参照 2015-11-12).
- [5] 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/charshape/>, (参照 2019-06-07).
- [6] デジタルアーカイブシステム ADEAC. <https://trc-adeac.trc.co.jp/>, (参照 2019-07-12).
- [7] 岡田一祐. “『和翰名苑』仮名字体データベース”. <https://kana.aa-ken.jp/wakan/>, (参照 2016-08-16)

- [8] 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://jcbsv.nii.ac.jp/oa/NIJL0-1/items/NIJL0002.zip>, (参照 2016-07-25).
- [9] 法書会編. “五體字類”. <http://www.let.osaka-u.ac.jp/~okajima/PDF/5tai/>, (参照 2015-11-12).
- [10] 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/>, (参照 2017-02-15).
- [11] 望月陽一郎. “「小中学校で1人1台の端末」なんて夢のまた夢. そもそもネットにつながらない”. Forbes JAPAN, <https://forbesjapan.com/articles/detail/27818>, (参照 2019-10-24).