

ADEAC の画像データを利用したくずし字認識 AI の開発と 組み込みシステムへの実装

Development of Embedded AI System for Recognition of *Kuzushiji* by Using the Image Data on ADEAC®

早坂 太一¹ 竹内 正広¹ 大野 亙¹

加藤 弓枝² 山本 和明³ 石間 衛⁴ 石川 徹也⁴

Masahiro Takeuchi¹, Taichi Hayasaka¹, Wataru Ohno¹,

Yumie Kato², Kazuaki Yamamoto³, Mamoru Ishima⁴, and Tetsuya Ishikawa⁴

¹ 豊田工業高等専門学校, 愛知県豊田市栄生町 2-1

¹National Institute of Technology, Toyota College, 2-1 Eisei, Toyota, Aichi

² 鶴見大学, 神奈川県横浜市鶴見区鶴見 2-1-3

²Tsurumi University, 2-1-3 Tsurumi, Tsurumi-ku, Yokohama, Kanagawa

³ 国文学研究資料館, 東京都立川市緑町 10-3

³National Institute of Japanese Literature, 10-3 Midori-cho, Tachikawa, Tokyo

⁴TRC-ADEAC株式会社, 東京都文京区大塚 3-1-1

⁴TRC-ADEAC, Inc., 3-1-1 Ohtsuka, Bunkyo-ku, Tokyo

概要: 古典籍におけるくずし字翻刻に関する研究では深層学習を利用したアプローチが盛んである。本研究では、ROIS-DS 人文学オープンデータ共同研究センターが公開している 100 万字を超えるくずし字データセットにデジタルアーカイブシステム ADEAC 内の古典籍画像から抽出した字形データ約2万6千字を加えて、16 ビット Unicode にくずし字を分類する畳み込みニューラルネットワークの学習を行った。その結果、版本のテストデータではそれほど影響が見られなかった認識率について、写本のテストデータに対しては明らかな向上が見られた。さらに、そのモデルをシングルボードコンピュータ Raspberry Pi に実装することで、複数のくずし字を一括して自動検出し、認識を行うことのできる組み込みシステムを開発した。インターネットへの接続を必要としないため、小中学校での教育や古民家での調査などの場面で手軽に利用でき、くずし字翻刻の支援ツールとして活躍することが期待できる。

Abstract: There are many approaches using deep learning in research on the interpretation of *kuzushiji* characters in Japanese ancient documents. In this study, we trained a convolutional neural network that classifies *kuzushiji* into 16-bit Unicode characters. In addition to the over 1 million characters published by Center for Open Data in the Humanities, we extracted approximately 26,000 *kuzushiji* characters as learning data from the Japanese ancient document images on ADEAC®, which is the system of digitalization and exhibition for archive collections. The recognition rate was clearly improved for the test data in handwritten documents, which was not significantly affected for the test data in printed documents. Furthermore, we developed the embedded system that can automatically detect and recognize multiple *kuzushiji* by implementing the deep learning model on the single-board computer Raspberry Pi. Since the system does not require the internet connection, it can be expected to play an active role as a support tool for interpreting *kuzushiji* in the situations such as education in elementary and junior high schools, surveys in old houses, and so on.

キーワード: 文字認識, くずし字, ADEAC, 深層学習, 組み込みシステム

Keywords: character recognition, *kuzushiji*, ADEAC, deep learning, embedded system

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、ROIS-DS 人文学オープンデータ共同研究センターが公開しているくずし字データセット[4]に、デジタルアーカイブシステム ADEAC [5]の古典籍画像から抽出した字形データを加えた学習データを用いて、くずし字を分類する畳み込みニューラルネットワークの学習を行い、その結果から、版本および写本に対する学習データの相違による認識精度を比較する。さらに、小型で比較的安価なシングルボードコンピュータ Raspberry Pi にその学習モデルを組み込んだくずし字の自動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(deep learning)は、第3次AIブームの中心的技術として様々な分野で導入が進んでいる。その原理は、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(多層パーセプトロン)が基になっている。深層学習により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畳み込みニューラルネットワーク(CNN; iConvolutional 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 Unicode 65,535 クラスに分類する学習を行った。CNN に漢字認識を行わせた先行研究[13]を参考にし、本研究で用いたネットワーク構造を表 1 に示す。

表 1 学習に用いたネットワーク構造(1+Max Pooling 層,²+ReLU 層,³+Dropout 層)

層の種類	ユニット数
畳み込み層 ^{1,2}	64
畳み込み層 ^{1,2}	128
畳み込み層 ^{1,2}	256
畳み込み層 ^{1,2}	512
全結合層 ^{2,3}	1,024
全結合層 ²	1,024
全結合層(出力)	65,535

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 [6]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×2 を搭載したワークステーション GDEP Deep Learning Box を利用した。

学習には、ROIS-DS 人文学オープンデータ共同研究センター(CODH; Center for Open Data in the Humanities)が公開しているくずし字データセット[4](以下、CODH データと略す)、および ADEAC [5]における「常総市デジタルミュージアム」「宮代町デジタル郷土資料」「京都女子大『山城国淀藩上月家文書』」から提供された 4,335 種類、1,117,703 文字のくずし字画像データを用いた。学習データの詳細を表 2 に示す。また、表 3 に学習には用いていないテストデータに対する認識率、すなわち確信度の高い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を示

表 2 学習データの詳細

	ダウンロード		手作業で抽出			小計	種類数	種類数／文字
	CODHデータ[4]	和翰名苑[7]	ADEAC[5]	CODHデータ[8]	五體字類[9]			
変体仮名 (割合)	714,568 (96.82%)	3,265 (0.44%)	9,209 (1.25%)	9,511 (1.29%)	1,473 (0.20%)	738,026	77	9,585
漢字 他 (割合)	360,550 (94.96%)	—	16,920 (4.46%)	2,207 (0.58%)	—	379,677	4,258	89
小計 (割合)	1,075,118 (96.19%)	3,265 (0.29%)	26,129 (2.34%)	11,718 (1.05%)	1,473 (0.13%)	1,117,703	4,335	258
種類数	4,299	48	1,638	417	48			

表3 テストデータに対する認識結果

(上段:確信度最上位, 下段:確信度 10%以上のすべての文字に対する認識率)

	(版本)CODH 源氏物語[10]	認識率		(写本)ADEAC 御着城御当日 御規式帳[11]	認識率	
		上記すべて 学習に使用	CODH デー タ[4]使用		上記すべて 学習に使用	CODH デー タ[4]使用
変体仮名	9,852	98.15%	98.08%	2	50.00%	100.00%
		99.39%	99.37%		100.00%	100.00%
漢字 他	1,280	87.58%	87.66%	964	74.07%	66.22%
		93.75%	92.97%		81.95%	73.34%
小計	11,132	96.94%	96.88%	966	73.91%	65.94%
		98.74%	98.63%		81.78%	73.19%

す。本研究では、CODH データのみを学習に用いた場合と、ADEAC における古典籍画像データから手作業で抽出した字形データをそれらに加えた場合とで、江戸時代の版本である『源氏物語』[10]、および写本である『御着 城御当日御規式帳』[11]をテストデータとしたときの認識率を比較した。

CODH オープンデータが 100 万字を超えるデータ数なのに対して、ADEAC からの字形データは 2 万 6 千字あまりと、ほとんど学習に影響がないように予想されるが、表 3 より、版本のテストデータではそれほど影響が見られなかった認識率について、写本のテストデータに対しては明らかな向上が見られていることがわかる。これは ADEAC のデータが全国各地の写本を主とするものであることによるものと考えられる。割合としては少なくとも、学習する字形データの時代や書き手、種類などのバリエーションを増やすことは、こうした深層学習によるくずし字翻刻アプローチには必要な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文字分のくずし
字を翻刻する 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.



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

このアプリケーションでは、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 秒である。

4. Raspberry Pi を用いたくずし字自動検出・認識システム

4.1 システムの概要

現状では、多人数が利用できる Wi-Fi 等のインターネット環境が整備されている小中学校やコミュニティ施設はそう多くはないため、古典籍に関する生涯学習等の講座を開催したい場合に、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。本研究では、インターネット環境を必要としないくずし字の自動抽出および認識を行う組み込みシステムを開発した。図 2 にシステムの概要図を示す。

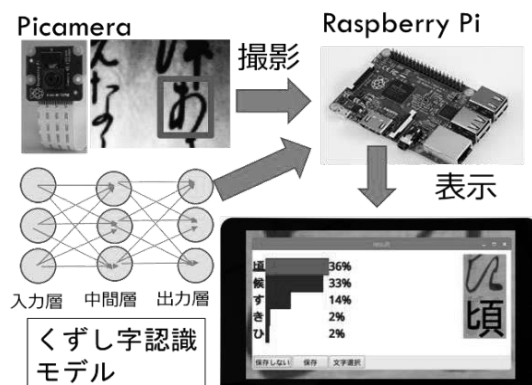


図2 開発システムの概要図

図2に示すシステムでは、ハードウェアとして小型で比較的安価な教育用シングルボードコンピュータ Raspberry Pi Model 4B を用いた。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できる。また、Raspberry Pi に挿入された SD カードに保存された画像の読み込みも可能である。

§3で述べた CNN モデルをシステムに搭載しているが、これはくずし字1文字に対して認識を行うものであるため、Python3 および OpenCV3 を用いて認識するくずし字の選択や領域検出のための画像処理などを行う。認識は機械学習用フレームワーク TensorFlow Lite [12]を用いて行われ、確信度の高い順に五つの認識結果の候補がタッチスクリーンに表示される。

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座標およびy座標を取得した後、その領域をリサイズし、2値化を行った後に、CNN モデルに入力している。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機能を実装した。

くずし字領域の自動検出にはモルフォロジ処理を利用した。入力画像に対して、図 3(a)に示すように縦方向に対して膨張処理を行い、上下の字を繋げることで、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)に検出されたくずし字領域の例を示す。

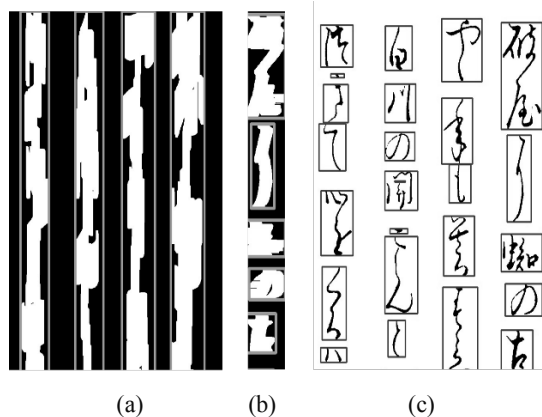


図3 文字領域の検出過程と結果例

また図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 CNN モデルで一度認識を行い、確信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 5 秒でくずし字領域の検出が可能である。

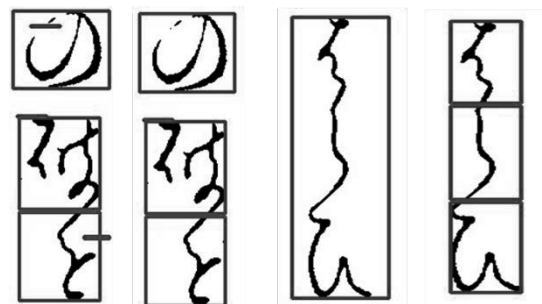


図4 検出した文字領域の分割・除外・統合の例

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 [12]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 0.08 秒とすることができた。

4.4 操作例

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。SDカードやUSBフラッシュメモリに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。

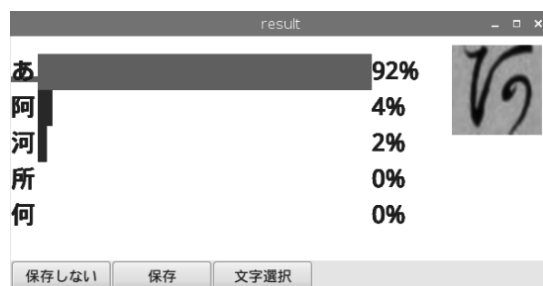


図 5 開発システムによるくずし字認識結果の表示例

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。



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「山水下」[13]より)

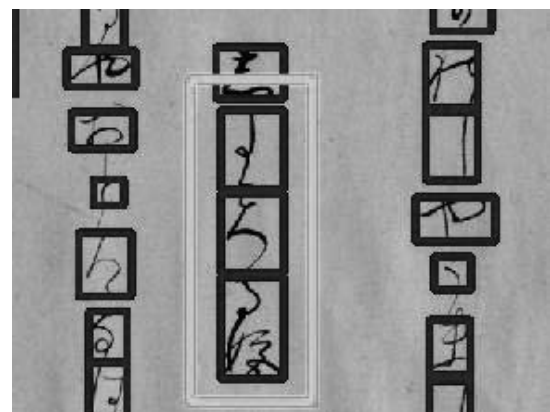


図 7 開発システムにおける編集領域の選択例

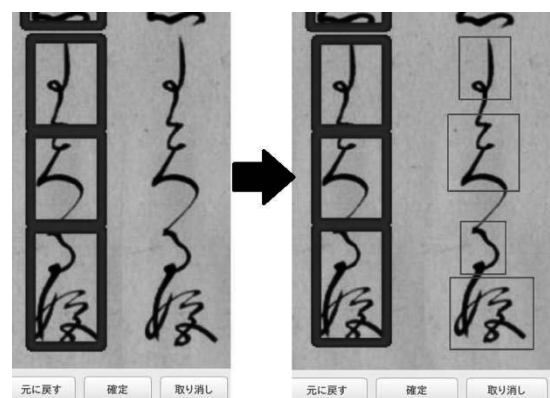


図 8 開発システムにおける検出領域の再選択の例

4.5 テストデータに対する評価

ADEAC から提供のあった写本「山下水」[13](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、62.2%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率は、確信度最上位の文字に対して 57.3%、確信度が第5位までのすべての文字に対して 78.4%であった。

5. あとがき

本研究では日本語の歴史的典籍の自動翻刻を目的として、深層学習により畳み込みニューラルネットワークモデルにくずし字の認識を行わせ、それをくずし字 1 文字認識用の WWW アプリケーションとして実装した。結果として、高性能なハードウェアを利用しなくても十分な演算速度による翻刻が実現できることを明らかにした。デジタルアーカイブシステム ADEAC における歴史的典籍から時代や書き手、種類などのバリエーションを増やした字形データを学習データとして用いることによって、今後もさらなる認識率の向上に繋がることが期待される。

また本研究では、その深層学習モデルを利用して、カメラで撮影した歴史的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。小中学校などの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。

本研究の成果が発展することにより、近い将来、翻刻作業に人間を必要としなくなるという指摘もあるが、例えば機械翻訳技術が急速に発展している現在においても「翻訳」という職業はなくならないように、歴史的典籍を持つ古人の心を伝えるためには、例えば文学研究者の力が必要となると考えられる。研究者のみならず一般の人々が歴史的典籍を判読することを支援することで、海外における日本の歴史的典籍の利用価値を高め、それらに記された知識の遺産を有効活用することを促すことができると期待される。その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。

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. Clauwat, 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, vol.2016, no.2, pp.7-12.

[4] 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/>, (参照 2019-06-07).

[5] 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).

[6] 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).

[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://jcbvsv.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. “源氏物語: 日本古典籍くずし字データセット”. <http://codh.rois.ac.jp/char-shape/book/200003803/>, (参照 2020-01-18).

[11] 京都女子大学・京都女子大学図書館. “淀藩士上月家文書 御着 城御当日御規式帳”, <https://trc-adeac.trc.co.jp/WJ11F0/WJJS07U/2672055100/2672055100200030/mp100030>, (参照 2020-01-18).

[12] TensorFlow Lite, <https://www.tensorflow.org/lite?hl=ja>, (参照 2020-01-18).

[13] 河本家住宅保存会・手銭記念館・島根大学附属図書館. “山陰地域史資料アーカイブ 河本家古典籍 山下水”, <https://trc-adeac.trc.co.jp/WJ11F0/WJJS07U/3290515100/3290515100200010/mp900000>, (参照 2020-01-18).