様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

# 科学研究費助成事業　　研究成果報告書

令和　２ 年　 ６ 月　 １ 日現在

機関番号：５３９０１
研究種目：基盤研究(C)（一般）
研究期間：2016～2019
課題番号：１６Ｋ０２４３３

研究課題名（和文）人工知能による日本の歴史的典籍の自動翻刻システムの構築およびその活用に関する研究

研究課題名（英文）Study on the development and the utilization of automatic interpretation system of Japanese ancient documents

研究代表者
早坂　太一（Hayasaka, Taichi）

豊田工業高等専門学校・情報工学科・准教授

研究者番号：５０３１４０９２

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：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，古典籍画像から抽出した100万字を超えるくずし字データセットを用いて，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．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スタンドアロン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．

研究成果の学術的意義や社会的意義
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，歴史的典籍が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．

研究成果の概要（英文）：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 by using over 1 million kuzushiji characters as learning data from the Japanese ancient document images. 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.

研究分野：認知科学

キーワード：くずし字　文字認識　テキスト検出　深層学習　スタンドアロン

様　式　C－１９，F－１９－１，Z－１９（共通）

## １．研究開始当初の背景

「くずし字」とは草書体で書かれた文字のことを指すが，江戸時代末までは印刷されたものも含め書物や文書の多くがこの文字で書かれていた．19 世紀末以降に活版印刷が主流になると，出版物では楷書体が用いられるようになり，教育の現場においても明治 33 年に小学校で教授されるべき平仮名字体が１音価１字体に定められた．さらに，昭和 23 年 1 月 1 日に施行された戸籍法施行規則第 60 条で，人名表記での変体仮名の使用が制限されたことが表しているように，日本人にとって「くずし字」の日常性は次第に失われていった．また，書道人口もこの 20 年間で半減しており[1]，草書体を学ぶ人口が減少した結果，当然のことながらくずし字を読める人口も減っていった．こうした背景や近年の生涯学習ブームにより，くずし字翻刻に関する研究の需要がより高まってきている．

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」[2]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っている．

## ２．研究の目的

現行の研究の中で，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，進捗度の大きい分野であると考えられる．本研究では，その中でも，様々な分野で導入が進んでいる深層学習を用いたアプローチにより，くずし字の自動翻刻を行うコンピュータシステムの開発を行う．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，階層型ニューラルネットワークと同様に，一度モデルを構築しさえすれば自動翻刻に要する時間はごくわずかである．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，深層学習の導入によって，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．研究の方法

1) 国文学研究資料館が作成し，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[3]に，デジタルアーカイブシステム ADEAC [4]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う．
2) 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 アプリケーションを開発・公開する．
3) 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スタンドアロンシステムを開発する．

## ４．研究成果

1) 本研究では，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（以下，CNN; Convolutional Neural Network と略す）によりくずし字認識モデルを学習させた．くずし字を１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．

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 [5]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．

学習には国文学研究資料館が作成し，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[3]（以下，CODH データと略す），および ADEAC [4]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，計 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 1 に示す．また表 2 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』[9],

および写本である『御着　城御当日御規式帳』[10]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 110 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 26,000 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 2 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．

表 1　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 [3] | 和翰名苑[6] | ADEAC [4] | CODH データ [7] | 五體字類[8] | | | |
| 変体仮名 (割合) | 714,568 (96.84%) | 3,265 (0.44%) | 9,061 (1.23%) | 9,511 (1.29%) | 1,473 (0.20%) | 737,878 | 77 | 9,583 |
| 漢字 他 (割合) | 360,550 (94.93%) | — | 17,068 (4.49%) | 2,207 (0.58%) | — | 379,825 | 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 2　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 [9] | 認識率 | | （写本） ADEAC 御着城御当日御規式帳[10] | 認識率 | |
|---|---|---|---|---|---|---|
| | | 上記すべて学習に使用 | CODH データ[3]のみ使用 | | 上記すべて学習に使用 | CODH データ[3]のみ使用 |
| 変体仮名 | 9,852 | 98.15% | 98.08% | 3 | 100.00% | 100.00% |
| | | 99.39% | 99.37% | | 100.00% | 100.00% |
| 漢字 他 | 1,280 | 87.58% | 87.66% | 963 | 74.66% | 66.87% |
| | | 93.75% | 92.97% | | 82.87% | 74.14% |
| 小計 | 11,132 | 96.94% | 96.88% | 966 | 74.74% | 66.98% |
| | | 98.74% | 98.63% | | 82.92% | 74.23% |

2)　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 アプリケーションを公開した (http://vpac.toyota-ct.ac.jp/kuzushiji/)．ブラウザ画面の例を図 1 に示す．

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．

令和元年 11 月 16 日に名古屋市立桜丘中学校で出前授業を行い，開発したアプリケーションのデモンストレーションを行った (図 2)．動作に問題はなかったが，教室にインターネット環境はなく，複数台のタブレット端末と共に私たちが Wi-Fi ルータを用意しなければならなかった．くずし字教育に利活用するためには，まだまだ多人数が利用できるネットワーク環境の整備が不十分である教育機関やコミュニティ施設での利用には，そうした環境を事業提供者が用意するための相応の時間と費用がかかるという問題点がある．


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

図 2　出前授業におけるくずし字認識 WWW アプリケーションのデモンストレーションの様子

3)　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スタンドアロンシステムを開発した．図 3 にシステムの概要図を示す．図 3 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．Raspberry Pi は ARM プロセッサを搭載した学校教育向けのシングルボードコンピュータであり，教育目的以外にも IoT や組み込みシステム開発などで利用されている．

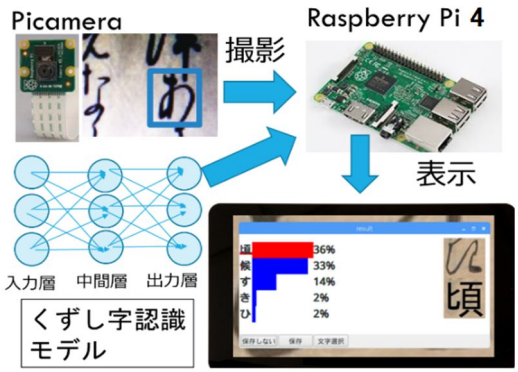
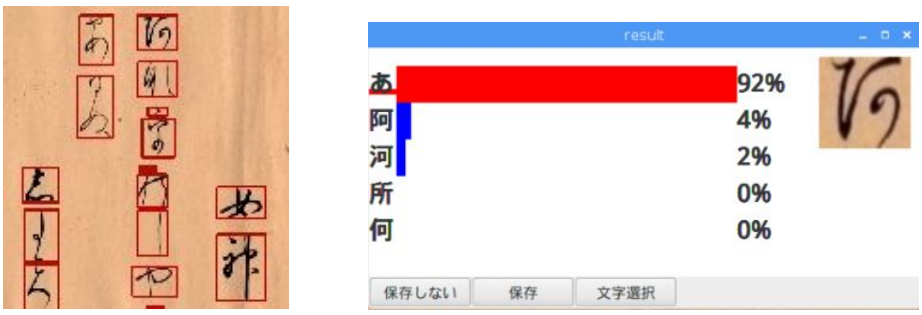



図 3　開発システムの概要図

ディスプレイの背面に設置した Raspberry Pi 専用のカメラモジュール Camera Module V2 から対象の古典籍画像を撮影できる．本システムでは解像度を 1920×1080 ピクセル，フォーカスを 30cm に設定した．入出力装置として Raspberry Pi 用の 7 インチタッチスクリーンディスプレイを使用した．さらに，本システムでは microSDHC 32GB のカードを使用することで，画像の読み込みおよびくずし字検出・解読結果の保存を可能とした．

1)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 を用いて認識する文字の選択や領域検出のための画像処理などを行う．まず，学習した CNN モデルについて，入力層のみ 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，モデルサイズを約 1/4 まで軽量化することができ，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．

古典籍画像に対する前処理には OpenCV3 を利用した．ここで，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，モルフォロジ処理を適用して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．本システムでは，100 文字あまりの古典籍画像 1 枚に対して，約 5 秒でくずし字領域の検出が可能である．図 4 (a)に示すような検出結果画像における赤色で囲まれた各矩形の内部をタッチすることで CNN モデルによる認識が行われ，図 4 (b)に示すように，確信度が高い順に五つの認識候補が表示される．




(a)　(b)

図 4　くずし字領域の自動検出および認識の例

システムの評価として，ADEAC [4]から提供のあった古典籍「御着　城御当日御規式帳」[10]（ルビや欄外文字を除く 939 文字）に対して，本システムによるくずし字領域の自動検出を行ったところ，適合率 92.6%，再現率 87.4%，F 値 89.9%（出力文字数 887，正解文字数 821）の精度を得た．残念ながら，1 行に収まりきらずに行末に付け足された文字や，ルビのような文字サイズの異なる文字についてはまだまだ検出精度が良くない．また，単文字

認識結果として，確信度が最上位の認識結果に対して適合率 59.2%，再現率 55.9%，Ｆ値 57.5%(正解文字数 525)，確信度が５位以内の認識結果に対して適合率 76.3%，再現率 72.1%，Ｆ値 74.2%（正解文字数 677）であった．検出領域の取り方に認識精度が左右されるため，本システムではウェブサービスである KuroNet[11]や表 3 に示した結果には追い付かなかったが，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によるスタンドアロンシステムでも，十分実用的な性能を示すことはできたと考えられる．

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インタラクティブセッション，および令和元年 12 月 14 日に大阪府茨木市で開催された人文科学とコンピュータシンポジウムデモンストレーション発表において，20〜60 歳代の男性 19 名および女性 1 名に対し，以下に示すような設問により，本システムに関する５段階評価のアンケートを行った．いずれの設問も「５」が最高評価で，「３」は「どちらとも言えない」という選択肢である．

1) 操作性：簡単に操作することができましたか？
2) 視認性：画面は見やすかったですか？
3) 認識精度：認識精度はいかがでしたか？
4) 将来性：（今後，利用する機会があれば）便利そうなシステムだと思いますか？

アンケート結果を図 5 に示す．デモンストレーションの際に操作を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，その他の項目は，80%前後の方々から４以上の高評価を頂いており，本システムの有用性を客観的に示すことができたと考えられる．

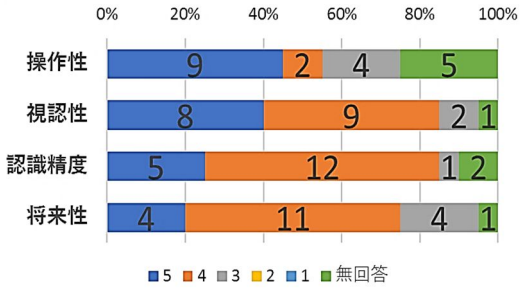



図 5　開発システムに対するアンケート結果

## 引用文献


[1] 日本生産施本部(編)，“レジャー白書 2013　やめる理由はじめる理由 — 余暇活性化への道筋，” 生産性出版日本生産施本部，2013

[2] 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.

[3] 人文学オープンデータ共同利用センター，“日本古典籍くずし字データセット（国文研所蔵・CODH 加工），” doi:10.20676/00000340, http://codh.rois.ac.jp/char-shape/, 2020 年 1 月 18 日参照.

[4] TRC-ADEAC 株式会社，“デジタルアーカイブシステム ADEAC，” https://trc-adeac.trc.co.jp/, 2020 年 1 月 18 日参照.

[5] 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:1408.5093, 2014.

[6] 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.

[7] 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（国文研等所蔵）二十一代集，” doi:10.20730/ 200007092, http://jcbsv.nii.ac.jp/oa/NIJL0-1/items/NIJL0002.zip, 2016 年 7 月 25 日参照.

[8] 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.

[9] 人文学オープンデータ共同利用センター，“日本古典籍くずし字データセット（国文研所蔵・CODH 加工）源氏物語，” http://codh.rois.ac.jp/char-shape/book/200003803/, 2020 年 1 月 18 日参照.

[10] 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, 2020 年 1 月 18 日参照.

[11] T. Clanuwat, A. Lamb, and A. Kitamoto, “KuroNet: Pre-modern Japanese Kuzushiji character recognition with deep learning,” Proceedings of the 15th International Conference on Document Analysis and Recognition, arXiv:1910.09433, 2019

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　0件／うち国際共著　0件／うちオープンアクセス　0件）

| | |
|---|---|
| １．著者名<br>早坂 太一 | ４．巻<br>28 |
| ２．論文標題<br>人工知能による日本の歴史的典籍の自動翻刻システムの構築 | ５．発行年<br>2018年 |
| ３．雑誌名<br>公益財団法人内藤科学技術振興財団研究成果論文集 | ６．最初と最後の頁<br>51-54 |
| 掲載論文のDOI（デジタルオブジェクト識別子）<br>なし | 査読の有無<br>無 |
| オープンアクセス<br>オープンアクセスではない、又はオープンアクセスが困難 | 国際共著<br>－ |

〔学会発表〕　計8件（うち招待講演　0件／うち国際学会　0件）

| |
|---|
| １．発表者名<br>竹内正広，早坂太一，大野亙，加藤弓枝，山本和明，石間衛，石川徹也 |
| ２．発表標題<br>ディープラーニングによるくずし字認識組み込みシステムの開発 |
| ３．学会等名<br>2019年度人工知能学会全国大会(第33回) |
| ４．発表年<br>2019年 |

| |
|---|
| １．発表者名<br>竹内正広，早坂太一，大野亙，加藤弓枝，山本和明，石川徹也 |
| ２．発表標題<br>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 |
| ３．学会等名<br>情報処理学会人文科学とコンピュータシンポジウム「じんもんこん2019」 |
| ４．発表年<br>2019年 |

| |
|---|
| １．発表者名<br>早坂太一，竹内正広，大野亙，加藤弓枝，山本和明，石間衛，石川徹也 |
| ２．発表標題<br>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 |
| ３．学会等名<br>第25回公開シンポジウム「人文科学とデータベース 」 |
| ４．発表年<br>2020年 |

１．発表者名
竹内正広，早坂太一，大野亙，加藤弓枝，山本和明

２．発表標題
RaspberryPiを用いたくずし字認識組み込みシステムの開発

３．学会等名
第17回情報科学技術フォーラム

４．発表年
2018年

１．発表者名
竹内正広，早坂太一，大野亙，加藤弓枝，山本和明，石間衛，石川徹也

２．発表標題
人工知能による日本の歴史的典籍の自動翻刻システムの構築

３．学会等名
第3ブロック専攻科研究フォーラム

４．発表年
2019年

１．発表者名
早坂 太一，大野 亙，加藤 弓枝，山本 和明

２．発表標題
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

３．学会等名
2017年度 人工知能学会全国大会（第31回）

４．発表年
2017年

１．発表者名
早坂太一，大野亙，加藤弓枝，山本和明

２．発表標題
ディープラーニングによる日本語の歴史的典籍におけるくずし字の認識およびWWWアプリケーション開発の試み

３．学会等名
電子情報通信学会パターン認識・メディア理解研究会

４．発表年
2016年

１．発表者名
早坂太一，大野亙，加藤弓枝，山本和明

２．発表標題
ディープラーニング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

３．学会等名
情報処理学会人文科学とコンピュータシンポジウム「じんもんこん2016」

４．発表年
2016年

〔図書〕　計0件

〔産業財産権〕

〔その他〕
豊田高専・くずし字翻刻WWWサービス
http://vpac.toyota-ct.ac.jp/kuzushiji/

６．研究組織

| | 氏名<br>（ローマ字氏名）<br>（研究者番号） | 所属研究機関・部局・職<br>（機関番号） | 備考 |
|---|---|---|---|
| 研究分担者 | 加藤　弓枝<br>(Kato Yumie)<br>(10413783) | 鶴見大学・文学部・准教授<br>(32710) | |
| 研究分担者 | 大野　亙<br>(Ohno Wataru)<br>(60321444) | 豊田工業高等専門学校・電気・電子システム工学科・准教授<br>(53901) | |
| 連携研究者 | 山本　和明<br>(Yamamoto Kazuaki)<br>(90249433) | 国文学研究資料館・古典籍共同研究事業センター・センター長<br>(62608) | |