

古典籍画像に基づくくずし字認識A Iの改良 およびICT活用教育ツールへの適用

早坂 太一^{*1}

松田 訓典^{*2}

竹内 綾乃^{*2}

中西 智子^{*2}

山本 和明^{*2}

^{*1} 豊田工業高等専門学校

^{*2} 国文学研究資料館

国文研 DDH プロジェクト
キックオフシンポジウム
「A I×人文学ーデータ駆動による未来形成ー」
令和6年12月1日 @一橋講堂

あらまし：早坂、山本らは、国文学研究資料館による「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]で整備された歴史的典籍データベースを有効活用し、深層学習を用いたアプローチの先駆けとして、小型で比較的安価なシングルボードコンピュータに、深層学習で得られたAI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を実現している[2]。本シンポジウムでは、ブラウザ上で動作するWWWアプリケーションに加えて、複数人での作業実現を可能とするべく、既存のシステムに共同翻刻機能を実装したWebシステムについて報告する。

※本研究は国文学研究資料館の「国文研プロジェクト型共同研究」として実施されたものです。

RaspberryPiによるくずし字認識システム



河本家住宅保存会・島根大学附属図書館蔵 河本家古典籍「山下水」



御書城御当日御規式帳（京都女子大）

カメラで撮影された古文書画像をタッチスクリーン付きの7インチディスプレイに表示し、一度に複数のくずし字の抽出および認識が可能である。抽出にかかる時間は1ページあたり約5秒で、認識にかかる時間は1文字あたり約0.08秒と、手のひらサイズのコンピュータとしては十二分なパフォーマンスを披露することができた。本システムのAIは、全国各地の図書館や博物館とデジタルアーカイブシステムの構築を手掛けるTRC-ADEAC(株)と共同で開発し、京都女子大学所有の写本におけるくずし字を80%以上の精度で認識した。

M5StickVによるくずし字認識システム

AI推論に適したマイクロコントローラを搭載した小型デバイスを用いて、変体仮名を認識できるシステムで、Wi-Fiを必要とせず、数千円程度で実現できる。

国立国会図書館デジタルコレクション 平治物語



くずし字認識モデル

- 4層の畳み込み層と3層の全結合層から構成されるCNNである。
(畳み込みニューラルネットワーク)
- 62×62ピクセルJPEG形式のグレー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。
- 計算機はG-DEP DeepLearning Box(OS: Ubuntu 14.04 LTS, GPU: nVidia GeForce 1080Ti 11GB)を利用し、学習フレームワークはCaffeを用いた。
- 学習データ一覧

	ダウンロード		手作業で抽出		五體字類	小計	種類数	文字数 (総数)
	CODH文字	和籙名宛	ADEAC※1	CODH書影※2				
変体仮名※3	714,568	3,265	9,029	9,511	1,473	737,846	77	9,582
(割合)	96.85%	0.44%	1.22%	1.29%	0.20%	66.20%		
漢字他	360,550	—	14,027	2,207	—	376,784	4,320	87
(割合)	95.49%	—	3.72%	0.59%	—	33.80%		
小計	1,075,118	3,265	23,056	11,718	1,473	1,114,630	4,397	253
(割合)	96.46%	0.29%	2.07%	1.05%	0.13%			
種類数	4,299	48	1,604	417	48			

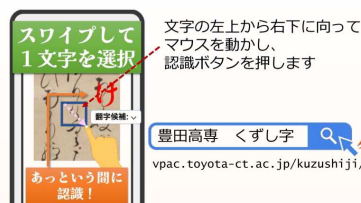
- テストデータに対する認識結果

	CODH 源氏 物語	認識率		(写本) 御書城 御当日 御規式帳	認識率	
		上記すべて 学習に使用 精度10%以上	CODH文字 のみ使用 精度10%以上		上記すべて 学習に使用 精度10%以上	CODH文字 のみ使用 精度10%以上
変体仮名※3	9,852	98.29%	98.08%	2	100.00%	100.00%
漢字他	1,280	86.25%	87.66%	964	73.65%	66.22%
小計	11,132	92.66%	92.97%	966	81.47%	73.34%
		96.91%	96.88%		73.71%	65.94%
		98.65%	98.63%		81.47%	73.19%

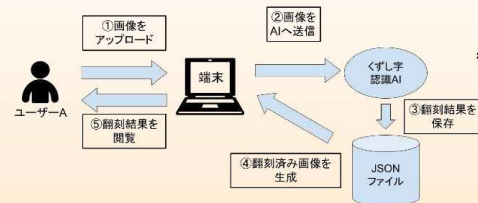
写本に対する認識結果の差が際立つ

※1：常陸市デジタルミュージアム
赤穂遺士、小野寺・近松岡氏書翰
宮代町デジタル郷土資料
元禄6年騎西落着落論議許状
深川はなみ集
巨勢帖
玉宛集
徳川歴代将軍朱印状
京都女子大・山城国淀上戸家文書
大橋院様御代付候御書
組々御代付候御書之留
※2：古今和歌集、金葉和歌集
※3：Unicode x3042~x3093

WWWアプリケーション(1) 1文字ごと



WWWアプリケーション(2) 1行ごと

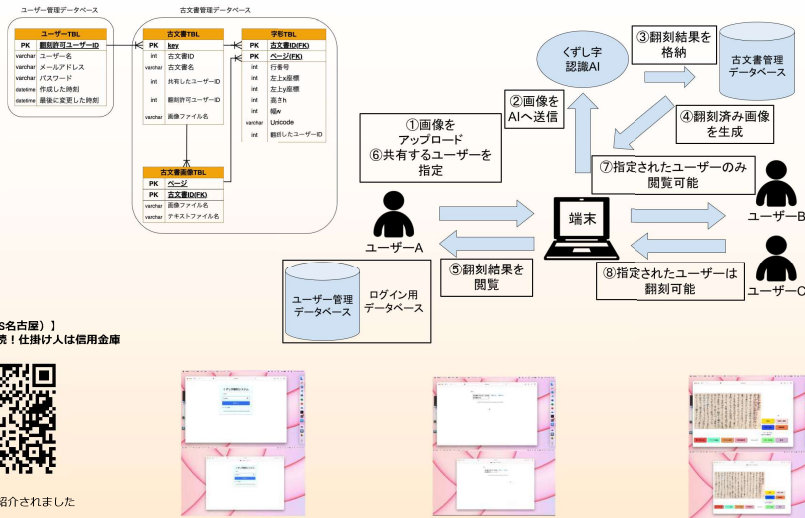


【大河ドラマは家康（LBS名古庫）】
徳川家康の手紙をAIが解読！仕掛け人は信用金庫



テレビ愛知で紹介されました

WWWアプリケーション(3) 共同翻刻機能付き



① ユーザがそれぞれログイン

① 上のユーザ名はuser1、下のユーザ名はuser2である。

② ホーム画面から「画像閲覧」ボタンをクリックすることでユーザー間で共有された画像を選択し、翻刻を行うことができる。このページには古文書管理データベースに格納された古文書名と古文書IDが表示される。古文書IDは画像1枚ごとに生成される固有の番号である。右の「翻刻する」リンクをクリックすることで翻刻することができる。

③ 翻刻ページのリロードボタンまたはキーボードのF5キーを押すことで、他ユーザーの翻刻結果が適応される。

参考文献：

- 国文学研究資料館 古典籍共同研究事業センター：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。https://www.nijl.ac.jp/pages/cijproject/、2015
- 早坂太一、山本和明：くずし字を読むエッジA Iの開発、第7回日本語の歴史的典籍国際研究集会、2021
- テレビ愛知TV Aichiチャンネル：徳川家康の手紙をAIが解読！仕掛け人は信用金庫、https://youtu.be/fKxorilW3wo、2023

