Deriving Orthographic Data from Classical Japanese Texts with Machine-Learning Methods

Herman Chau / Michael R. Zeng / Paul S. Atkins (University of Washington, Seattle)

Abstract: This project applies advanced machine-learning techniques to extract orthographic data—specifically *jibo* 字母, the Chinese character matrices underlying cursive Japanese hiragana—from classical Japanese manuscripts. Inspired by the National Diet Library's NDLkotenOCR and the Center for Open Data in the Humanities' (CODH) KuroNet, our aim is to automate the generation of *jibo* data from manuscript images. This automation enables large-scale orthographic analysis and scribal attribution, which has traditionally required extensive manual effort. By integrating modern computer vision techniques, we seek to create a robust pipeline that identifies *jibo* to facilitate deeper linguistic and historical insights into classical Japanese texts.
Keywords: orthography, calligraphy, hiragana

概要：このプロジェクトでは、高度な機械学習技術を用いて、古典日本語の写本から表記データ（すなわち、ひらがなの基盤となる漢字「字母」)を抽出する．国立国会図書館の NDLkotenOCR や、人文学オープンデータ共同利用センター（CODH）の KuroNet に触発され、写本画像から字母データを自動生成する仕組みを構築する．この自動化により、従来は膨大な手作業を要していた大規模な表記分析や筆者の特定が可能になる．最新のコンピュータビジョン技術を統合することで、字母を識別する堅牢なパイプラインを実現し、古典日本語テキストに対する言語的・歴史的理解を深める．
キーワード：表記、字母、ひらがな

1. Introduction

This project applies advanced machine-learning techniques to extract orthographic data—specifically *jibo* 字母, the Chinese character matrices underlying cursive Japanese hiragana—from classical Japanese manuscripts. Inspired by the National Diet Library's NDLkotenOCR and the Center for Open Data in the Humanities' (CODH) KuroNet, our aim is to automate the generation of *jibo* data from manuscript images [1,2]. This automation enables large-scale orthographic analysis and scribal attribution, which have traditionally required extensive manual effort. By integrating modern computer vision techniques, we seek to create a robust pipeline that identifies *jibo* to facilitate deeper linguistic and historical insights into classical Japanese texts.

2. Background and Previous Work

The writing system known as *hiragana* represents the sounds of the Japanese language using phonetic symbols that are highly cursivized forms of Chinese characters with similar pronunciations (e.g., the phonetic symbol あ *A* is derived from the character 安 *AN* meaning 'safety'). Modern *hiragana* represents each sound with one and only one character (e.g., the sound *A* is always represented by あ) but, until the twentieth century, *hiragana* was polygraphic: each sound could be represented by multiple characters (e.g., the sound *A* could be represented also by cursivized versions of the characters 阿, 亜, 悪, or 愛) and the choice of which to use appears to have been largely a question of personal preference. The historian of Japanese calligraphy Komatsu Shigemi (1925-2010) was the first scholar to demonstrate that statistical analysis of *jibo* 'character matrices,' the Chinese characters from which premodern scribes derived their hiragana symbols, could be used to identify the scribe of a manuscript [3].

Although Komatsu's method was innovative and his findings convincing, he did not apply modern statistical techniques of his analysis and therefore did not provide basic information, such as the probability that the results could have been obtained by chance. However, Komatsu's method has been refined by and is being used to great effect by Professor Saitō Tetsuya of Shukutoku University in Japan. He has published a suite of articles categorizing texts by time period or scribe using statistical analysis of *jibo* orthographic data [e.g., 4,5].

Recent work by Atkins and Zeng (under review) applies contemporary statistical techniques to the Ogura *shikishi* 小倉色紙, a famous corpus of some fifty poems attributed to the hand of the medieval Japanese poet Fujiwara no Teika 藤原定家 (1162-1241), revealing that none were likely inscribed by him. These findings underscore the potential of *jibo* analysis for historical scholarship.

Statistical analysis of *jibo* holds tremendous potential for revising the history of the production of

premodern Japanese manuscripts. The major obstacle, however, is the acquisition and preparation of data. In order to undertake large-scale studies of the vast corpus of classical Japanese literary texts, it is necessary to find some way of generating *jibo* data in Unicode from images of manuscripts in .pdf or other formats. This would entail using machine-learning techniques related to computer vision. Indeed, machine learning has already been used to decipher cursivized classical Japanese texts (called *kuzushiji*, or "collapsed characters") with promising results, the most prominent application being a delightful smartphone app called Miwo みを that allows users to decipher texts onsite in libraries or museums.

Current OCR tools such as Miwo and NDLkotenOCR, however, bypass *jibo* identification, limiting their utility for orthographic research [6, 1]. Our project addresses this gap by developing a machine-learning pipeline capable of extracting *jibo* data directly from manuscript images, thereby enabling new forms of analysis and discovery.

3. Objective

Our first goal is to enable automatic batch recognition of entire classical texts that includes *jibo* data. Subsequent goals include the automatic tabulation of *jibo* frequencies in a text and visualizing the relative frequency of a given text compared to known frequencies from various scribes. We plan for our interface to be public-facing so that other scholars are able to leverage *jibo* frequency analysis to help attribute scribeship of a given text. This tool will empower scholars to conduct statistical analyses of orthographic patterns, support or question scribal attributions, and contribute to broader studies in historical linguistics and manuscript studies. By making this tool publicly accessible, we aim to democratize access to advanced orthographic analysis and foster interdisciplinary collaboration.

4. Methodology

Dataset Preparation

Our main dataset is the *Kuzushiji Dataset* offered by the Center of Open Data in the Digital Humanities (CODH), which contains 44 labeled classical Japanese manuscripts, totaling over 5,000 pages and over a million single-*kuzushiji* cropped images [7]. The dataset is formatted as a series of manuscript images together with annotation files in the COCO format that describe bounding boxes of individual characters in the image as well as their character class labels. Altogether, there were over 4,000 character classes when including both *kuzushiji* and *kanji* characters. There were two primary limitations to the dataset that we had to address. First, the annotations in the CODH dataset only use modern *kana* instead of individual *kuzushiji* annotations. Second, there are severe class imbalance issues where the most frequent character classes have more than 30,000 samples whereas about 1,000 classes have fewer than 10 samples. Altogether, the 1,000 most common character classes account for over 95% of the dataset. See Figure 1 for the distribution of character class samples.

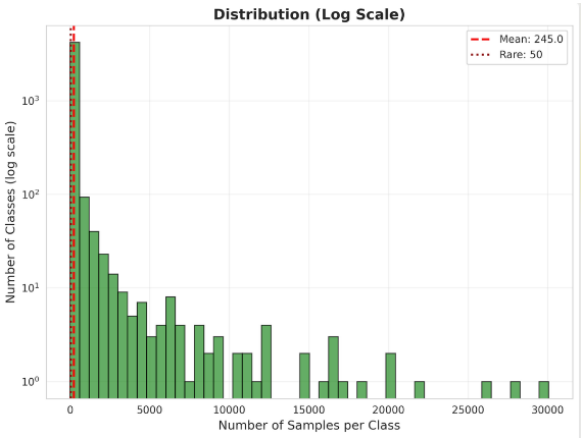

Figure 1. Histogram of sample sizes across character classes.

To refine the CODH dataset to include *kuzushiji* annotations, we leveraged a smaller *kuzushiji*-labeled dataset prepared by the National Institute for Japanese Language and Linguistics (NINJAL) [8]. We used the NINJAL dataset to bootstrap training of a vision transformer classifier for single-*kuzushiji* cropped images. This classifier was applied to the CODH dataset to generate an initial refinement with *kuzushiji* annotations. Manual visual inspection was then performed to correct annotation errors and ensure labeling accuracy. The whole process took about 3 months to complete, with significant speedup due to the initial bootstrapped classifier.

To address the class imbalance issue in the CODH dataset, we applied three data augmentation techniques. First, in the training pipeline, standard distortions such as affine transformations and blurring are applied to all character classes. Second, for rare classes, we trained a Generative Adversarial Network (GAN) to generate synthetic *kuzushiji* images. For classes with 50-300 samples, our GAN models generate high quality images and we augment these classes with the GAN output to a total of 500 samples. Third, for classes with fewer than 50 samples, we experimented with generating synthetic images through the GPT-4o API, which achieves higher quality synthetic images at an elevated cost. See Figure 2 for a comparison of the

GAN generated images and the LLM generated images for the rare character *kanadzu/sō* 奏.

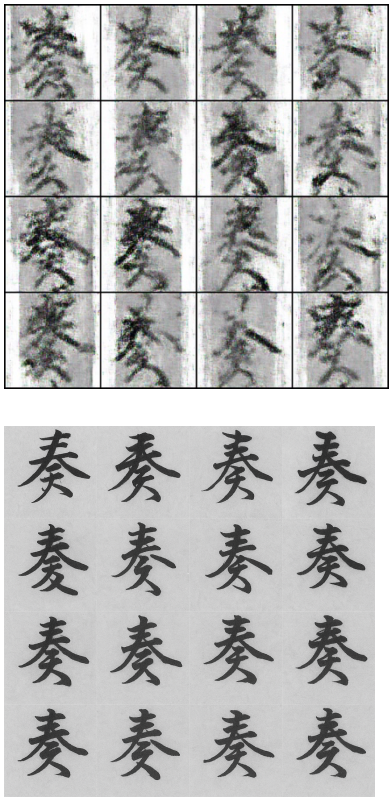

Figure 2. Comparison of the GAN generated images and the LLM generated images for the relatively rare character *kanadzu/sō* 奏.

Model Training and Architecture

We propose a three-stage pipeline for processing manuscript images. In the first stage, we apply an image detection model to generate bounding boxes on an entire page of manuscript. In the second stage, individual characters are cropped according to the outputted bounding boxes and each individual crop is classified according to a vision transformer model. Because cursive writing can often be ambiguous and even human experts struggle to identify certain *kuzushiji* without context, we concatenate the output of the second stage to be fed into our third stage. The third stage ensembles the visual predictions with a lexical model that leverages context to catch vocabulary and grammatical errors. Figure 3 contains a schematic of the proposed pipeline. All our model training is performed on a Tesla T4 GPU with 16 GB of RAM through Amazon Web Services (AWS) and all models are trained with a 90/10 training-test split of our augmented CODH dataset.

The image detection models evaluated for the first stage in our pipeline are Cascade R-CNN with a ResNet-101 backbone, YOLOv11, and a DETR transformer-based detection model. We observed that both our Cascade R-CNN and YOLOv11 models are performant whereas the DETR was unable to provide bounding boxes for a vast majority of the characters on our test manuscript pages. A summary of the training time and mAP@50 score for our three models is shown in Table 1.

Table 1. Comparison of detection model architectures

Model	**Epochs**	**Time Taken**	**mAP@50**
Cascade R-CNN	36	~72 hours	0.848
YOLOv11	~200	~24 hours	0.973
DETR	60	~72 hours	N/A

We fine-tune a vision transformer for use in the second stage of our pipeline. We start with a model that is pre-trained on ImageNet-21k and fine-tuned on ImageNet 2012 and further fine-tuned it on our individual *kuzushiji* cropped images. The vision transformer is trained for 48 epochs and achieved an overall accuracy of 96.93% on our hold-out test set.

For the third stage of our pipeline, we started with a RoBERTa model pretrained on texts from Aozora Bunko. The model is then further fine-tuned on the text present in the CODH manuscript data. We did not observe a noticeable improvement in accuracy when ensembling our vision model predictions with this fine-tuned RoBERTa model and are experimenting with integrating an LLM-based step.

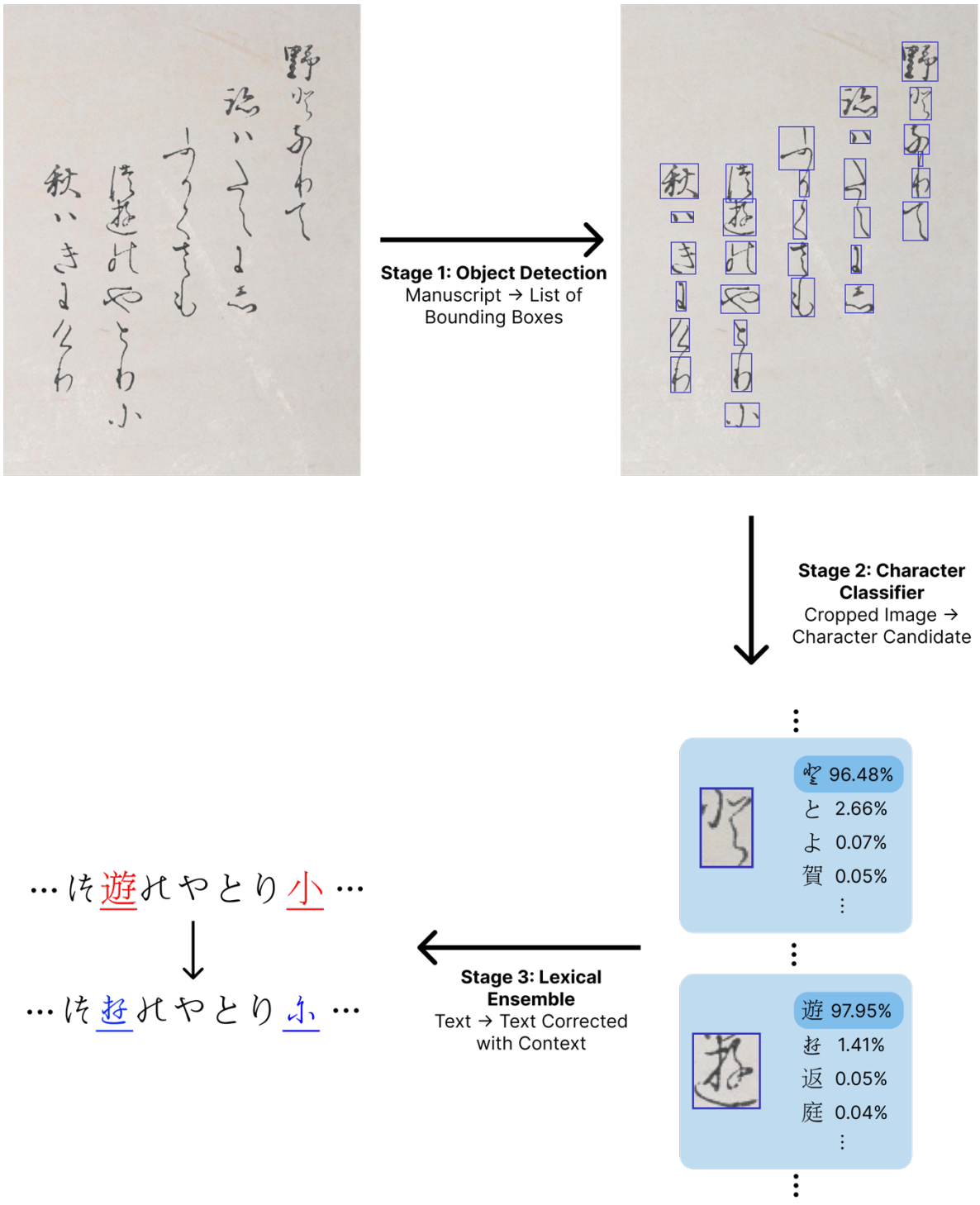

Figure 3. Illustration of our 3-stage pipeline from manuscript image to final character recognition.

Results and Limitations

We evaluated our Cascade R-CNN based model and YOLOv11 model on our hold-out test set and report on the overall F1 score. While not directly comparable due to the use of different evaluation datasets, we summarize in Table 2 our model scores together with reported scores from NDLkotenOCR v2 and KuroNET as points of comparison. We note that the models are also not directly comparable insofar as neither NDLkotenOCR nor KuroNET supports *jibo* recognition.

Table 2. Comparison of F1 scores across our pipeline and other classical Japanese OCR software

Model	Average F1 Score
Our pipeline with Cascade R-CNN	0.85
Our pipeline with YOLOv11	0.96
NDLkotenOCRv2 [1]	0.91 (median)
KuroNET [2]	0.79 (median)

For further analysis, we computed the F1 scores across our test set for each character type. The F1 score drop-off for both the Cascade R-CNN model and YOLOv11 model is shown in Figure 4. We note that the YOLOv11 model is significantly more robust with over 1000 classes achieving a perfect F1 score. However, not all 4,444 character classes present due to some characters missing in the randomized 90/10 train-test split. This also explains the difference in the cut-off points for the YOLOv11 versus Cascade R-CNN model.

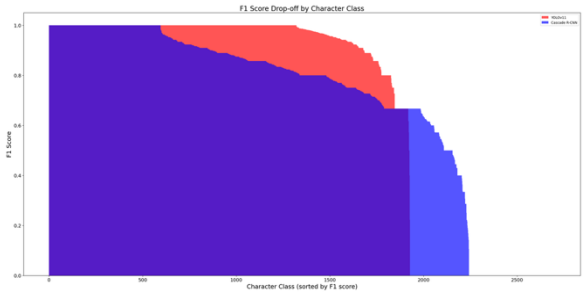

Figure 4. Comparison of F1 score drop-offs sorted by character classes between Cascade R-CNN and YOLOv11.

5. Future Development

Future training phases will continue to focus on data augmentation to address the data imbalance issue. We are experimenting with generating synthetic Japanese manuscripts with synthetic single-character data. Another focus is contextual correction through a suitable lexical model. Our focus is on maximizing accuracy, even at the expense of speed, in order to provide reliable results for scholarly use.

While we have addressed some shortcomings of our dataset, we acknowledge that there are other limitations that we have not been able to account for. For example, the dataset is heavily skewed towards xylographs, as opposed to manuscripts, and the early modern era (1600 – 1868). This bias in our training data would lead to a dip in accuracy of our models on images outside of the kind represented in the training data.

We will build a public-facing web interface that allows users to upload manuscript images, run model inference, and receive electronic transcriptions. Additional features will include *jibo* frequency analysis and scribal attribution tools. This interface will be designed for ease of use and will support both academic and casual users interested in classical Japanese texts. Additionally, give the consent of users, we will build a database of digitized manuscript data.

6. Significance

The *jibo* harvester is currently hosted on Amazon Web Services (AWS) and will eventually be publicly released. Scholars will be able to use the tool to derive *jibo* data for their texts and compare them with previously uploaded texts. These uploaded texts will also contribute to refining our model. The labeled dataset and web interface will be open access, promoting transparency and collaboration.

By bridging the gap between traditional manuscript studies and modern machine learning, this project aims to transform the field of classical Japanese orthography, giving its users a powerful tool with which to rewrite the history of classical Japanese language, literature, and calligraphy.

Acknowledgements

We are grateful to Professor Kitamoto Asanobu of CODH for valuable advice. We thank Amazon Web Services (AWS) and the eScience Institute of the University of Washington, Seattle, for providing cloud computing credits. We thank the Simpson Center for the Humanities, University of Washington, Seattle for Digital Humanities Summer Fellowships.

References (all accessed 2025.10.10)

[1] National Diet Library (NDL). *Reiwa 4-nendo NDLOCR tsuika kaihatsu jigyō narabi dōjigyō seika ni taisuru kaizen sagyō* 令和 4 年度 NDLOCR 追加開発事業及び同事業成果に対する改善作業: https://lab.ndl.go.jp/data_set/r4ocr/r4_software/

[2] Tarin Clanuwat, Alex Lamb, Asanobu Kitamoto. "KuroNet: Pre-Modern Japanese Kuzushiji Character Recognition with Deep Learning." 2019: https://arxiv.org/abs/1910.09433

[3] Komatsu Shigemi 小松茂美 (1961), *Gosen wakashū: kōhon to kenkyū* 後撰和歌集：校本と研究, 2 vols, Tokyo: Seishin Shobō.

[4] Saitō Tetsuya 齊藤鉄也 (2016), 'Kana *jibo* no shutsugen hindoritsu ni motozuku Fujiwara no Teika shosha shiryō no nendai suitei' 仮名字母の出現頻度率に基づく藤原定家書写資料の年代推定, *Jinmonkon 2016 ronbunshū* じんもんこん 2016 論文集, 197-202.

[5] Saitō Tetsuya 齊藤鉄也 (2018), 'Kana *jibo* no shutsugen keikō o mochiita Fujiwara no Teika shosha shiryō no chōsa' 仮名字母の出現傾向を用いた藤原定家書写資料の調査, *Jōhō Shori Gakkai Ronbunshi* 情報処理学会論文誌 59:2, 315-22.

[6] Clanuwat, T., & Kitamoto, A. (2021). 'miwo'AI Kuzushiji Recognition Application for Document Examination. In *Proceeding of IPSJ Humanities and Computer Symposium*.

[7] Center for Open Data in the Humanities. *Nihon kotenseki Kuzushiji deetasetto* 日本古典籍くずし字データセット:

https://codh.rois.ac.jp/char-shape/

[8] National Institute for Japanese Language and Linguistics (NINJAL). *Kokugoken hentaigana jikei deetabeesu* 国語研変体仮名字形データベース: https://cid.ninjal.ac.jp/hentaiganaDB/index.html