Japanese Cursive Character Recognition for Efficient Transcription

Kazuya Ueki and Tomoka Kojima
School of Information Science, Meisei University, Tokyo, Japan

Keywords: Character Recognition, Japanese Cursive Character, *Kuzushiji*, Convolutional Neural Network.

Abstract: We conducted detailed experiments of Japanese cursive character recognition to promote Japanese historical document transcription and digitization by using a publicly available *kuzushiji* dataset released by the Center for Open Data in the Humanities (CODH). Using deep learning, we analyzed the causes of recognition difficulties through a recognition experiment of over 1,500-class of *kuzushiji* characters. Furthermore, assuming actual transcription conditions, we introduced a method to automatically determine which characters should be held for judgment by identifying difficult-to-recognize characters or characters that were not used during training. As a result, we confirmed that a classification rate of more than 90% could be achieved by narrowing down the characters to be classified even when a recognition model with a classification rate of 73.10% was used. This function could improve transcribers' ability to judge correctness from context in the post-process—namely, the previous and subsequent characters.

1 INTRODUCTION

As Japanese characters have changed considerably over time, it has become difficult for non-experts to read classical Japanese literature. By transcribing historical documents, we are able to understand the events of past eras. Therefore, research on the transcription and digitization of historical documents has been conducted. However, a vast number of literary works have not yet been digitized. One of the considerable barriers to this digitization is that Japanese classical literature was written in a cursive style using *kuzushiji* characters that are very difficult to read compared with the contemporary style.

The primary difficulty that arises in *kuzushiji* character recognition is rooted in the fact that most *kuzushiji* characters are very difficult to distinguish clearly from other characters because the writing style varied among different eras and authors. Therefore, in this study, assuming the actual transcription process, we considered a method that could automatically detect which characters were difficult to classify using deep learning and mark them as unknown characters without making the final decision.

2 RELATED WORK

Currently, handwritten characters can be recognized with high accuracy by using convolutional neural networks (CNNs). This trend has been highly motivated by a CNN called LeNet with a convolution and pooling structure that was proposed by (LeCun et al., 1989)(Lecun et al., 1998); this network successfully recognized handwritten digits. CNNs have also been widely used for recognizing *kuzushiji* characters in classical literature and achieved relatively high accuracy (Hayasaka et al., 2016)(Ueda et al., 2018). These studies on *kuzushiji* recognition have only focused on classifications of less than 50 *hiragana* character classes. However, classical documents are not limited to *hiragana*, as they also include *katakana* and *kanji* characters. To create electronic texts of classical literature or documents, it is necessary to recognize a wide range of characters and improve the accuracy with which they are transcribed. Based on this, we focused on recognizing not only *hiragana* characters but also *katakana* and *kanji* characters using the publicly available *kuzushiji* dataset (Clanuwat et al., 2018).

When a text includes many character classes, we need to address data imbalance problems. Yang et al. conducted experiments of hand-written characters from documents recorded by the government-general of Taiwan from 1895 to 1945 (Yang et al., 2019). The data imbalance problem was solved by applying data augmentation when training deep learning models.

Reducing the labor load on transcribers was also considered an important issue. Thus, a new type of OCR technology was proposed by (Yamamoto and Osawa, 2016) to reduce labor in high-load reprint work. It was deemed important to divide reprinting work among experts, non-experts, and an automated

Ueki, K. and Kojima, T.
Japanese Cursive Character Recognition for Efficient Transcription.
DOI: 10.5220/0008913204020406
In *Proceedings of the 9th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2020)*, pages 402-406
ISBN: 978-989-758-397-1; ISSN: 2184-4313

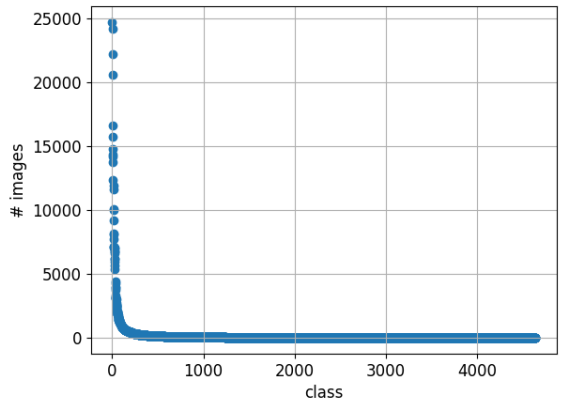

Figura 1: The number of images in each class in the *kuzushiji* dataset

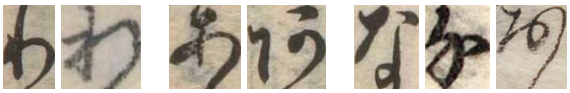

Figura 2: Left: Examples of different characters with similar shapes. Center: Examples of identical characters with separate deformations caused by different *jibo*. Right: Examples of differences in deformation despite being linked to the same *jibo* depending on the work or writer.

process rather than relying on automating processing entirely. The objective was not to achieve a decoding accuracy of 100% with OCR automatic processing alone; rather, ambiguous letters were marked as “ 〓 (*geta*)” and passed on to the expert in charge of post-processing to make a decision. As a result, it was possible to achieve rapid and high-precision reprinting. We conducted an experiment to confirm whether it was possible to automatically eliminate erroneously classified characters or characters that were not used in the training.

3 KUZUSHIJI RECOGNITION EXPERIMENTS

3.1 Kuzushiji Image Dataset

We used the Japanese cursive *kuzushiji* character image dataset (Clanuwat et al., 2018) released by the Center for Open Data in the Humanities (CODH). This *kuzushiji* database includes cropped images of three different sets of characters—namely, *hiragana*, *katakana*, and *kanji*. In total, 684,165 images with 4,645 character classes were included. These images were cropped from 28 literary works written in the Edo period (i.e., 1603–1868 AD). The number of images in each class is heavily unbalanced. For example, the classes with the largest and second largest number of images are “ に ” (Unicode: U+306B) and

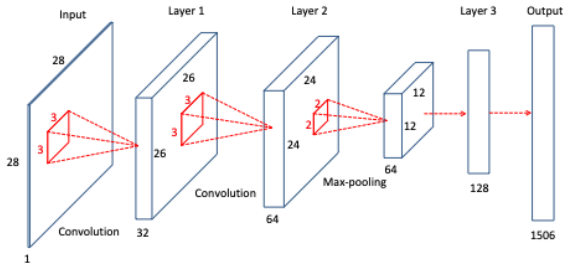

Figura 3: Neural Network Structure.

“ の ” (Unicode: U+306E), which have 24,710 and 24,171 images, respectively. In contrast, many classes only have a few images, and 882 classes only have one image. This data bias complicates recognition.

As we checked images in the dataset, many characters were difficult to distinguish clearly from others due to the variation in their shape, as indicated on the left in Fig. 2. There were also identical characters whose appearances differ greatly, as shown in the center of Fig. 2. This is because *hiragana* and *katakana* characters were originally derived from *kanji* characters called *jibo* (maternal glyph), and identical *hiragana* and *katakana* characters were derived from different *jibo*. Moreover, identical characters may have the same *jibo* but be represented differently depending on the literary work or writer, as shown on the right of Fig. 2.

3.2 Dividing the Data Used in the Experiment

In this study, *kuzushiji* character images from 28 literary works were divided into training, validation, and test datasets. The training data was used to train a recognition model, and the validation data was used to select the optimal model by evaluating the data for each epoch. Test data was evaluated using the trained model, and the model’s performance was measured. Specifically, character images from two books (“Oraga Haru” and “Ugetsu Monogatari”) were used as test data, and the remaining 26 literary works were used as training or validation data. The number of images in the 26 literary works was 628,136 with 4,493 character classes.

3.3 Neural Network Structure

In our experiments, we used the CNN architecture of LeNet as shown in Fig. 3, which consists of two convolution layers—including pooling layers and one fully-connected layer with a softmax function—to output probabilities for each character class. With regard to the other layers, a rectified linear unit (ReLU) activation function was used. Models were trained with the AdaDelta algorithm.

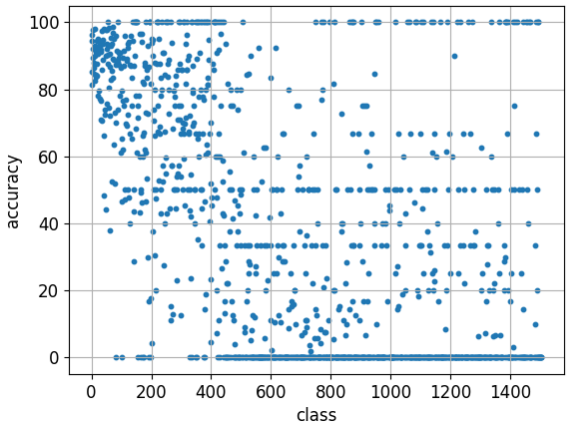

Figura 4: Classification rate for each character class.

4 KUZUSHIJI RECOGNITION EXPERIMENTS

In this section, we show how well various types of characters can be recognized using deep learning and analyze the causes of character misclassification.

4.1 Data Used in the Experiments

As the number of images in each class was unbalanced, we only used character classes with more than 20 images in the training set. The number of classes used in our experiments was 1,506. Because the classification rates tend to deteriorate for classes with a limited number of training samples, training data were augmented to ensure that the number of images in each class was over one a hundred using the following transformations: rotation ($\leq$ 3 degrees), width shift ($\leq 0.03 \times image\_width$), height shift ($\leq 0.03 \times image\_height$), RGB channel shift ($\leq$ 50), shear transformation ($\leq \pi/4$), and zoom ($[1 - 0.03 \times image\_size, 1 + 0.03 \times image\_size]$). After the data augmentation, the number of images was 664,840. We divided these images into training data (80%: 531,872 images) and validation data (20%: 132,968 images).

The total number of test data is 56,029 with 2,249 character classes. In the test data, many character classes were not included in the 26 literary works and character classes that were excluded during training because of the limited number of samples (less than 20 images in the class). By eliminating these character classes, the number of images that could be recognized was changed to 53,026.

Figura 5: Example images of the character “ 夢 ” (Unicode: U+5922) in the test data. Two images on the left were misclassified as “ 藤 ”. Two images on the right were misclassified as “ 義 ”.

Figura 6: Top: Example images of the character “ 夢 ” (Unicode: U+5922) in the training data. Middle: Example images of the character “ 藤 ” (Unicode: U+85E4) in the training data. Bottom: Example images of the character “ 義 ” (Unicode: U+7FA9) in the training data.

4.2 Experimental Results

The classification rate for the test data was 73.10% (= 40,952 / 56,029). When considering the character classes in the training data only, the classification accuracy was 77.23% (= 40,952 / 53,026). However, the average accuracy of each class was as low as 37.22%. The classification rate for each class is shown in Fig. 4. It is clear that many character classes could not be correctly classified at all.

4.3 Error Analysis

By analyzing instances in which performance was poor for specific characters, we were able to identify certain root causes. Here, we show some typical misclassified examples.

The classification rate of the character “ 夢 ” (Unicode: U+5922), which is displayed in Fig. 5, was very low at 8.8% (=3/34). As shown in the top of Fig. 6, this character trained many variations in shape. When this character was tested, most images were misclassified as the character “ 藤 ” (Unicode: U+85E4), which is shown in the middle of Fig. 6, or the character “ 義 ” (Unicode: U+7FA9), which is shown in the bottom of Fig. 6. The left two images of Fig. 5 were misclassified as the character “ 藤 ”, and the right two images of Fig. 5 were misclassified as the character “ 義 ”.

Other examples are shown in Fig. 7. The classification rate for the character “ 茶 ” (Unicode:

Figura 7: Example images of the character “ 茶 ” (Unicode: U+8336) in the test data.

Figura 8: Example images of the character “ 茶 ” (Unicode: U+8336) in the training data.

U+8336) was 4.2% (= 2 / 48). Corresponding training data, which are shown in Fig. 8, include many types of literary works and characters that were styled differently by authors. Furthermore, as *kuzushiji* characters were handwritten by different authors, the character shape greatly differed in between training and testing. Thus, *kuzushiji* character classification was notably difficult.

4.4 Elimination of Difficult-to-Recognize Characters

In the test data derived from “Oraga Haru” and “Ugetsu Monogatari,” there were characters that were not included in the training data (26 literary works). Furthermore, character classes with less than 20 images in the training data were not trained. In the testing process, it is difficult to know whether there are unknown characters that have not been trained. Thus, a function that automatically eliminates such characters is necessary. In this experiment, 3,003 of 56,029 images in the test data were character classes that were not trained.

The character recognition rate using the experimental training model was 73.10%, which will lead to a reduction in work efficiency when actually performing transcription work. Therefore, it is important to eliminate false recognition as much as possible and leave only the correct recognition results.

Based on the error analysis described in the previous section, we found that the maximum output probability values from the softmax function tended to be low if character images that were not correctly classified were input to CNN. Thus, we confirmed whether it was possible to eliminate erroneously classified characters with low maximum output probability values (hereinafter referred to as confidence values) by using characters with high confidence values. Specifically, confidence values were used as a threshold, and characters below the threshold were reserved as unknown characters. As shown in the paper (Yamamoto and Osawa, 2016), efficient transcription will be achieved by marking low-confidence characters as

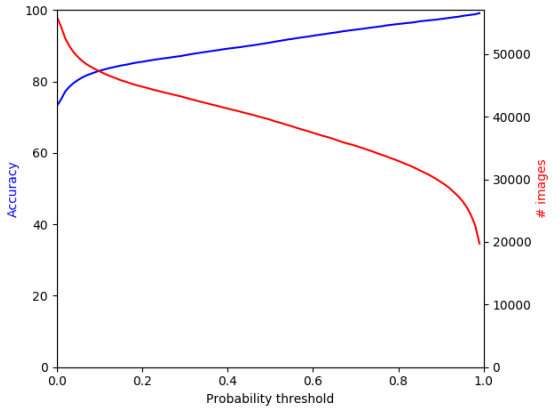

Figura 9: Relationship between the number of images that can be classified and the classification accuracy according to the confidence values. Horizontal axis: the threshold of confidence values. Vertical axis: the number of images that can be classified (red), and the classification rates (blue)

“ 〓 (*geta*),” and asking experts in the post-process to make a final judgment.

Figure 9 illustrates the relationship between the classification rates and the number of images that can be classified when the threshold of confidence values is changed. For example, if tested images were limited to the confidence value of 50% or more, the number of characters that could be classified was reduced to 40,209. Among them, the number of correctly classified images was 36,544, and the recognition rate improved to 90.89% (= 36,544 / 40,209). When we confirmed the result of 3,003 images of the character class not used for training, we found that 2,917 images (97.14%) could be eliminated, which was expected.

Figure 10 provides an example of the recognition results of one page from “Ugetsu Monogatari.” When the recognition results of the original book are displayed in typographical form, it is necessary to tell transcribers in the post-process which characters have low confidence values. In this experiment, we displayed characters with low confidence values in the red rectangles as shown in the center and on the right of Fig. 10. If the confidence value threshold is less than 50%, too many characters remained to be judged by humans, so it was necessary to reduce the number of characters to be eliminated in order to improve the efficiency of transcription. However, if the threshold of confidence values was set to less than 10%, the number of characters to be eliminated would be reduced. Thus, it would be easier to predict characters with low confidence values from the previous and subsequent characters with high confidence values.

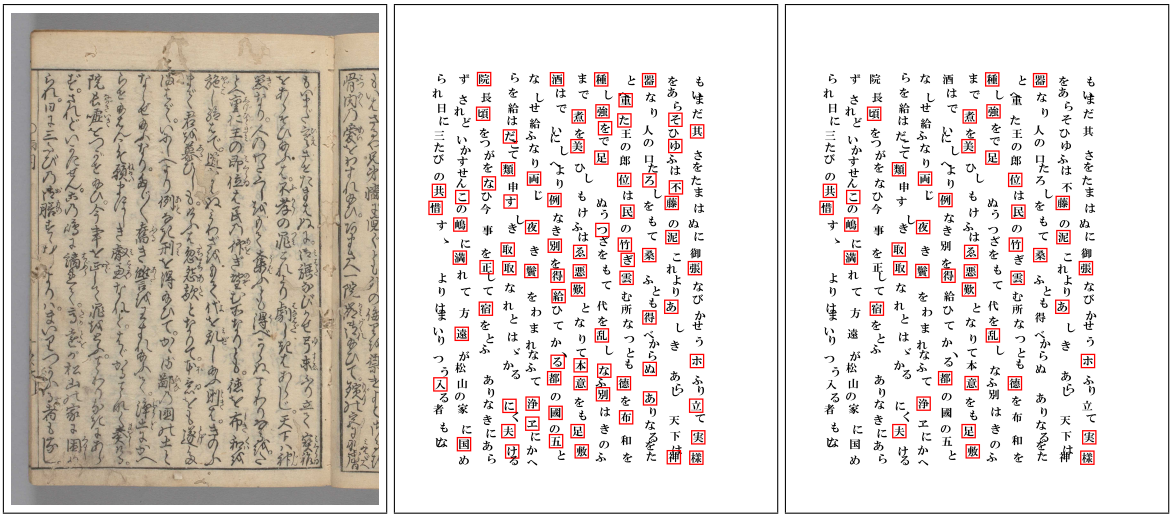

Figura 10: Example of the recognition result of one page from "Ugetsu Monogatari." Left: Original image of one page from "Ugetsu Monogatari." Center and right: Images showing the recognition results in print. Red rectangles indicate characters with confidence values of less than 50% and 10%, respectively.

5 SUMMARY

This study aims to improve the efficiency of transcribing Japanese classical literature. We conducted the experiments using deep learning to recognize Japanese cursive *kuzushiji* characters with over 1,500 classes from a large-scale dataset. By augmenting the data for classes with a small number of images, we achieved a classification rate of 73.10% for 1,506-class character classification, but the average classification rate over all classes was very low at 37.22%. As this was not sufficient to contribute to efficient transcription, we considered a method for automatically eliminating images of characters that may be erroneous or those that were not used for training. We attempted to automatically eliminate such characters by focusing on the confidence values from CNNs and confirmed that a recognition rate of 90% or more could be achieved by changing the confidence value threshold.

However, future studies must consider characters with low confidence values. We also plan to implement the following measures to improve the transcription process:

- We will implement a transcription system for transcribers in the post-process and obtain the feedback to improve the system. The system will likely have to provide some candidate characters for images with low confidence values to improve their ability to identify the correct character.
- We will incorporate rules based on expertise in *kuzushiji* characters in addition to improving recognition technologies such as deep learning.
- We will incorporate character recognition beyond the character-level, focusing on deriving information from context—namely, word-, phrase-, and sentence-level.

REFERENCES

Clanuwat, T., Bober-Irizar, M., Kitamoto, A., Lamb, A., Yamamoto, K., and Ha, D. (2018). Deep Learning for Classical Japanese Literature. *CoRR*, abs/1812.01718.

Hayasaka, T., Ohno, W., Kato, Y., and Yamamoto, K. (2016). Recognition of Hentaigana by Deep Learning and Trial Production of WWW Application. In *Proc. of Information Processing Society of Japan (IPSJ) Symposium*, pages 2121–2129. (in Japanese).

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Comput.*, 1(4):541–551.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, pages 2278–2324.

Ueda, K., Sonogashira, M., and Iiyama, M. (2018). Old Japanese Character Recognition by Convolutional Neural Net and Character Aspect Ratio. In *ELCAS Journal*, volume 3, pages 88–90. (in Japanese).

Yamamoto, S. and Osawa, T. (2016). Labor saving for reprinting Japanese rare classical books. In *Journal of Information Processing and Management*, volume 58, pages pp.819–827. (in Japanese).

Yang, Z., Doman, K., Yamada, M., and Mekada, Y. (2019). Character recognition of modern Japanese official documents using CNN for imblanced learning data. In *Proc. of 2019 Int. Workshop on Advanced Image Technology (IWAIT)*, number 74. (in Japanese).