Construction of Japanese Historical Hand-Written Characters Segmentation Data from the CODH Data Sets

Tang, Yiping
Kyushu University

Hatano, Kohei
Kyushu University

Ishita, Emi
Kyushu University

Nakatoh, Tetsuya
Kyushu University

他

https://hdl.handle.net/2324/2320606

出版情報：Proceedings of the Conference of Japanese Association for Digital Humanities. 8, pp.183-185, 2018-09-11. 日本デジタル・ヒューマニティーズ学会
バージョン：
権利関係：

Construction of Japanese Historical Hand-Written Characters Segmentation Data from the CODH Data Sets

Tang Yiping, Kohei Hatano, Emi Ishita, Tetsuya Nakatoh and Toshifumi Kawahira

(Kyushu Univeristy)

Introduction

Techniques for character recognition are of key components in the digital humanities. These days, there are more digital images of historical documents made, due to the development of scanners and computers. Although huge amount of such digital images are available, typical OCR (optical character recognition) systems for modern characters cannot be directly applicable to pre-modern character recognition problems. The hardness depends on languages. In particular, for Japanese, the difficulties of recognizing pre-modern hand written characters with computers are that (i) such documents are written by brushes and many characters are often connected, not separated by spaces, (ii) several different symbols (e.g., Chinese and Japanese ones) are used for meaning the same character, (iii) some characters are simplified or abbreviated. Therefore, it is still a challenge to recognize Japanese pre-modern texts from their images.

The recognition task can be divided into two phases, segmentation of sentences to single characters and recognition of single characters. Given an image of a single character, it is now an easy task to recognize the character, say, by using machine learning techniques such as the deep neural networks. For example, Nguyen et al. reported that their system can recognize single characters with accuracy 97% (Nguyen et al., 2017). On the other hand, segmenting an image of sentence to those of single characters is a bottleneck. Nguyen et al. also reported that the accuracy of their system for three consecutive characters is about 88%. So, a good segmentation algorithm will further increase the accuracy of recognition systems. The goal of this work is to construct data sets of Japanese pre-modern text with the information of segmentation of sentences, for which researchers and developers could test their segmentation algorithms. Our data sets will be available through the QIR, the institutional repository of Kyushu university.

The CODH data sets and their variants

In 2017, the Center for Open Data in the Humanities (CODH) published open data sets of Japanese pre-modern characters and literatures (CODH, 2017a). The data sets consist of images of pre-modern Japanese books and their transcriptions as well as data sets of images of 403,242 individual characters of 3,999 different types. The data sets are released under the license of CC-BY-SA which can be freely used and modified if an appropriate citation is added.

Based on the data sets, the PRMU (PRMU, 2017), hosted the programming context of recognizing Japanese pre-modern hand-written texts (called the “Kuzujishi Challange") (CODH, 2017b) in 2017. In this context, they posted about 230,000 pages of Japanese kana characters data from the 15 Japanese ancient books from the CODH data sets, released it on the web site, so that it can be freely used by any potential participants. The contest data sets consist of tuples of an original image of particular text, the characters of interest (one to about six characters) which correspond to the “true" recognition result, and information of positions of characters expressed by the coordinates of the rectangle enclosing them.

The task of the contest is, given the image and the coodinates of the enclosing rectangle, to recognize the characters. In particular, the contest data sets has three types of data, the level 1, 2 and 3 depending on the hardness. The level 1 data set consists of 240,000 single characters. The level 2 data set consists of 80,000 sets of three consecutive characters. The level 3 data set consists of sets with multiple characters (more than 3). An illustration of these data sets is shown in Figure 1.

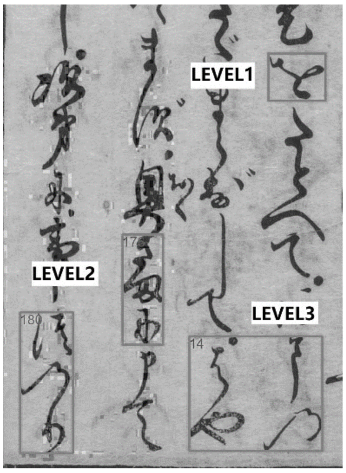

Figure 1: An example of contest data of the PRMU.

Construction of our data sets

Our technical work is to construct segmentation data sets by reforming the level 1,2, and 3 data sets from the PRMU contest. Our simple observation is that all the level 2 and 3 data sets contain single characters appearing in the level 1 data set. Therefore, by adding the information of enclosing rectangles of single characters in the corresponding, we can construct segmentation data sets of three or more characters. As a result, we construct

78,940 segmentation data sets of three consecutive characters and 12,583 multiple characters, respectively. More precisely, the each data is the tuple of the original image, information of a large rectangle enclosing three or more characters (the X and Y coodinates of the top-left corner, width and height), as well as small rectangles enclosing single characters within the large rectangle. Figure 2 represents examples of our data set.

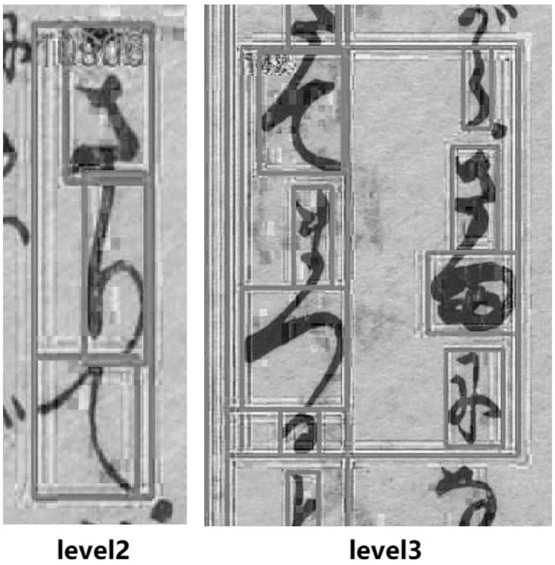

Figure 2: An example of our data set.

Furthermore, we also verified manually all of the constructed data. People checked the data set are 8 students specializing Japanese literatures, who can recognize Japanese historical characters. Through the manual check, we corrected 104 instances of the level 2 data set and 548 tuples of the data are excluded from our data set since students could not recognize them. Similarly, we corrected 95 instances of the level 3 data. Examples of difficult instances are shown in Figure 3.

Conclusions and future work

In this work, we constructed segmentation data sets of Japanese pre-modern characters from the CODH data sets and the PRMU contest. Our data set will be available under the license of CC-BY-SA at the web site. We will show some preliminary results of various segmentation methods over our data sets in the poster session.

Acknowledgement

We thank the support of the Tsubasa project at Kyushu University and JSPS KAKENHI Grant

Number JP18K18508.

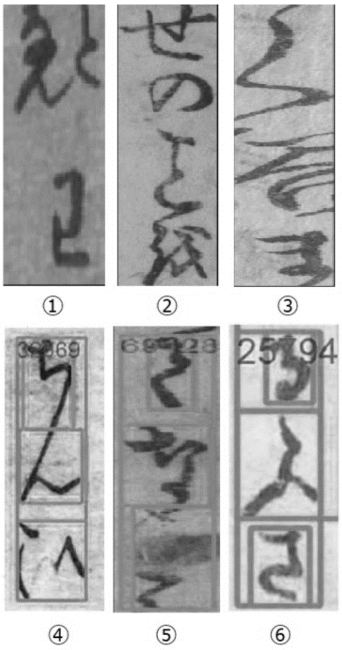

Figure 3: Examples of abnormal instances excluded from our data sets.

References

CODH. (2017a). http://codh.rois.ac.jp/char-shape/.

CODH. (2017b). http://codh.rois.ac.jp/old-char-challenge.

Nguyen, H. T., Ly, N. T., Nguyen, K. C., Nguyen, C. T., & Nakagawa, M. (2017). “Attempts to recognize anomalously deformed Kana in Japanese historical documents.” In *Proceedings of the 4th International Workshop on Historical Document Imaging and Processing (HIP2017)*, 31–36.

PRMU. (2017). https://sites.google.com/view/alcon2017prmu.