ICIC Express Letters
Volume **18**, Number **2**, February **2024**
 ISSN 1881-803X
pp. **135–141**

A SEGMENTATION METHOD FOR KUZUSHIJI BASED ON K-MEANS CLUSTERING

Wenyi Cui and Kohei Inoue*

Faculty of Design
Kyushu University
4-9-1, Shiobaru, Minami-ku, Fukuoka 815-8540, Japan
cui.wenyi.045@s.kyushu-u.ac.jp; *Corresponding author: k-inoue@design.kyushu-u.ac.jp

Received March 2023; accepted June 2023

Abstract. *Ancient Japanese books record a great amount of information, which are valuable research materials in study of history and culture. Over the past few years, there was a large-scale research on digitization of ancient Japanese books and we are convenient to use them due to open access on the Internet. However, it is a challenging problem to recognize those ancient Japanese books due to the complex background and unsteady shape of Japanese characters, which is called Kuzushiji. In this paper, we proposed a method to segment Japanese characters by using image processing and clustering. Our method is based on the analysis of character characteristics, which could identify the segmentation points more accurately. The validity of the proposed method was confirmed by the evaluation experiment.*
Keywords: Handwritten character segmentation, Machine learning, K-means clustering, Japanese historical book, Kuzushiji

1. **Introduction.** Ancient Japanese books and documents contain a vast amount of information, making them significant research materials for studying the history and culture of ancient Japan. According to the General Catalog of National Books [1], the number of books written or published in Japan prior to 1867 exceeds 1.7 million.

In recent years, libraries and museums have built a huge amount of digital copies of the books and documents. For example, Center for Open Data in the Humanities (CODH) has released the dataset of Pre-Modern Japanese Text [2], including images and texts from over 3,000 ancient Japanese books, which are owned by the National Institute of Japanese Literature [3]. Ritsumeikan University Art Research Centre has also established The Early Japanese Books Portal Database [4], which contains a huge number of digital archives of ancient Japanese books.

There was a large-scale research on the recognition of ancient Japanese books in recent years. However, most books are written in Kuzushiji, a typeface which contains many cursive and connected characters. Kuzushiji had been extensively utilized for over a thousand years in Japan since the 8th century. In 1900, Kuzushiji was not included in the elementary education system due to the reform of the Japanese writing system. Nowadays, Kuzushiji is not commonly used by Japanese natives and could only be deciphered by very few specialists.

The traditional Optical Character Recognition (OCR) technology has been widely used since the early 1990s while digitizing historical newspapers [5]. However, it is not able to segment Kuzushiji since there are no clear boundaries between characters. To solve this problem, we proposed a method of automatic character segmentation from images of ancient Japanese books by using image processing and K-means clustering. Our method is based on the analysis of character characteristics, which enables us to identify the

DOI: 10.24507/icicel.18.02.135

segmentation points more accurately. The validity of the proposed method was confirmed by the evaluation experiment using the image of an ancient Japanese book. Our research has improved the segmentation accuracy to 82.7%, surpassing previous studies.

This paper is divided into 6 sections as follows.

Section 1 is introduction. In this section, we introduced the background of ancient Japanese books and the difficult points of Kuzushiji segmentation.

Section 2 is related work. In this section, we analyzed several related studies and identified their deficiencies.

Section 3 is character segmentation. In this section, we described the operation of character segmentation including preprocessing, column segmentation and character detection.

Section 4 is connected character separation. In this section, we described the operation principle of K-means clustering and the operation of connected character separation, including center section, corner detection and optimum point calculation.

Section 5 is the evaluation experiment. In this section, we conducted an evaluation experiment to demonstrate the feasibility of our method.

Section 6 is the conclusion. In this section, we analyzed the execution result of the system and listed future improvements.

2. **Related Work.**

The segmentation of Kuzushiji poses a significant challenge due to the absence of clear boundaries between Kuzushiji characters. In recent years, several studies have focused on the segmentation of connected parts between Kuzushiji characters.

Isshiki et al. [6] proposed a method in 2020 based on features of connection character, achieving a precision of 77.6%. Isshiki et al.'s method assumes that the connected character area is composed of two equally spaced characters. However, the connected area is often made up of multiple characters with varying heights in reality.

Gao et al. [7] proposed a method in 2021 based on projection analysis of characters with a precision of 75.9%. Gao et al.'s method uses the location with the fewest horizontally projected pixels as the segmentation point. However, since strokes of characters often overlap horizontally, it cannot be guaranteed that the position with the fewest pixels corresponds to the actual split point.

Lyu et al. [8] proposed a method in 2022 based on K-means clustering with a precision of 80.3%. Lyu et al.'s method uses K-means clustering to determine the number of connected characters and then applies evenly cutting of the characters in the vertical direction, which also leads to inaccuracies.

Handwriting recognition is an important issue in each country with respective languages. Rehman proposed a method for improving the performance of online Arabic handwriting recognition [9].

It can be observed that previous studies did not specifically investigate the characteristics of Kuzushiji characters, and their segmentation accuracy did not exceed 80.5%. There is still room for improvement in this regard.

Based on the deficiencies of existing studies, we propose a new method to segment connected Kuzushiji characters based on analysis of character features. Given that Kuzushiji characters are written continuously with a brush, the joints between different characters can be distinguished by pauses and changes in stroke direction, which would cause inflection points. In essence, by identifying these inflection points and conducting an analysis, we can accurately determine the optimal cutting points, resulting in improved segmentation accuracy.

3. **Character Segmentation.**

The proposed method is divided into four steps. First, we perform preprocessing by removing background noises, such as smudges on the paper, and convert the input RGB image to a binary image. Then, we segment columns through

a vertical projection process. Then, we detect character through connected component labeling. Finally, we cut the connected part between characters through K-means clustering and skeletonization.

First, we use a median filter to remove background noise whilst preserving edges of characters [10]. Then, we use adaptive histogram equalization to improve contrast between characters and background in the image [11]. Different from ordinary histogram equalization, adaptive histogram uses corresponding to a distinct section of the image to redistribute the lightness values of the image, which makes it suitable for improving the local contrast and enhancing the definitions of edges in each region of an image. Finally, we transfer the original RGB image into a grayscale image and then use threshold holding to obtain a binary image with Otsu's method [12]. The input image and output binary image are shown in Figure 1.

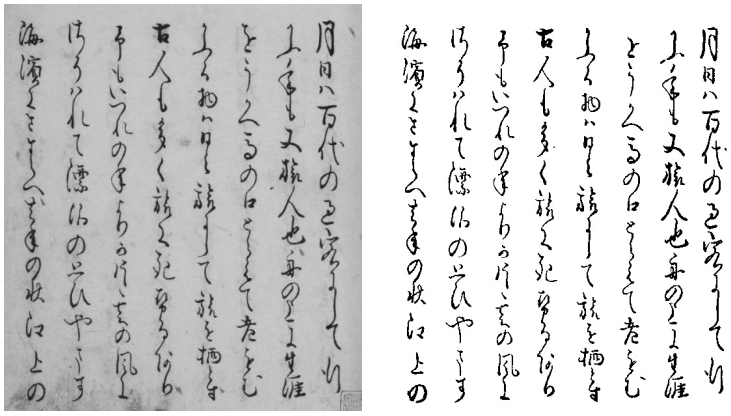

FIGURE 1. Input image and output binary image

Since most ancient Japanese books are written vertically, we cut the columns using vertical histogram projection. First, we count the total number of black pixels, which represent characters, along the columns of the image. Columns that represent the text have a high number of black pixels, which correspond to higher peaks in the histogram. Next step, we find the lower peaks in the histogram which represent the gaps in-between the columns. At last, we separate each combination of text lines by the lower peaks in the histogram. The vertical histogram of a binary image is shown in Figure 2.

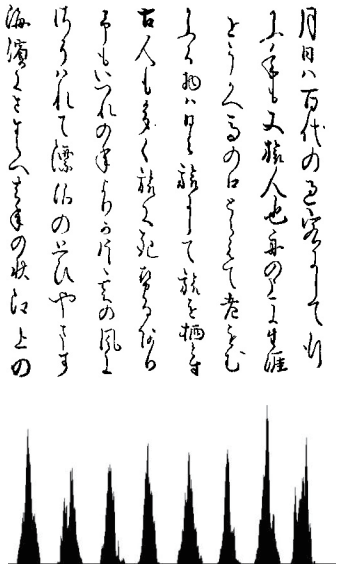

FIGURE 2. Vertical histogram

For each column, we use a morphology process to segment each character. First, we compute the dilation of the input image to enhance the width of strokes. Then, we segment characters through connected components detection [13]. The connected components and result of segmentation of the first column are shown in Figure 3 where different components are represented in different colors.

FIGURE 3. Connected components and result of segmentation

The result of segmentation of the whole image is shown in Figure 4 where character areas are surrounded by frames.

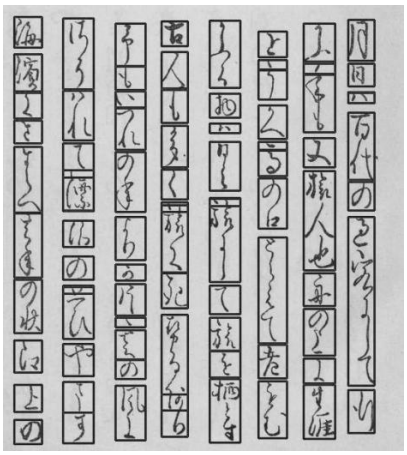

FIGURE 4. Character areas

4. **Separation of Connected Characters.** Figure 5 shows the process of connected character separation. Despite the rapid and continuous nature of Kuzushiji writing, there is typically a turning point between two characters. Therefore, we could separate characters by detecting all turning points and choosing the most suitable one.

First, we detect connected characters by comparing the height of the character area with the average height of character areas. We calculate the presume number of characters through function as follows:

$$n = \frac{h_i}{h},$$

where h_i denotes the height of character area, and h the average height of character areas. The character area will be decided as a connected character area if n is larger than 2.

Next, we skeletonize the connected characters, which reduces the foreground regions skeletal remnant that largely preserves the extent and connectivity of the original region

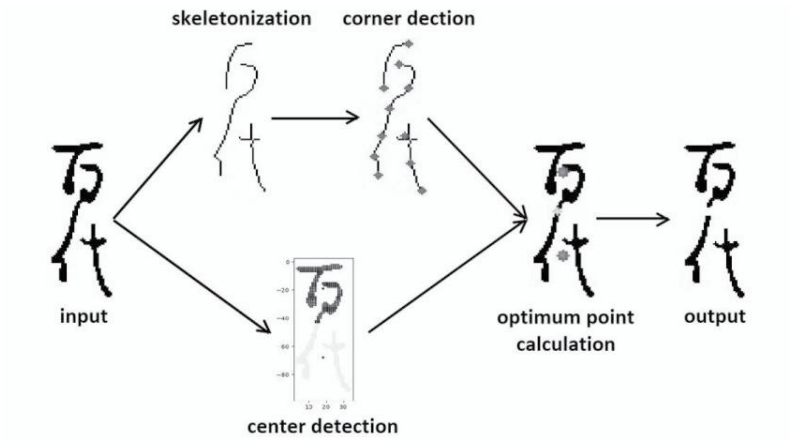

FIGURE 5. Process of connected character separation

FIGURE 6. Result of corner detection

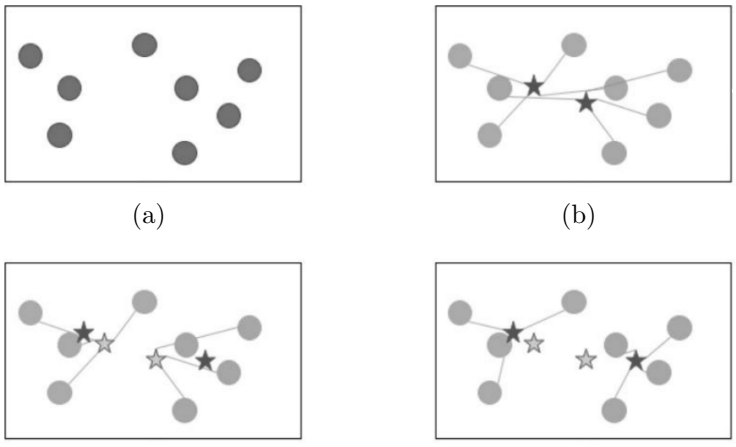

FIGURE 7. Algorithm of K-means clustering

while throwing away most of the original foreground pixels. Then, we calculate all the turning points through corner detection [14]. The sample of results of corner detection is shown in Figure 6.

Then, we use the K-means clustering analysis to classify characters, which can effectively cluster characters and reduce segmentation errors caused by complex backgrounds or variations in character shapes [15]. Given a set of observations $(x_1, x_2, \ldots, x_n)$, where each observation is a d-dimensional real vector, K-means clustering aims to partition the n observations into k ($\leq n$) sets $S = \{S_1, S_2, \ldots, S_k\}$ so as to minimize the within-cluster sum of squares. The implementation steps of the algorithm are shown as follows in Figure 7.

Step 1, k initial "means" (in this case $k = 2$) are randomly generated within the data domain. Step 2, k clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means. Step 3, the centroid of each of the k clusters becomes the new mean. Step 4, Steps 2 and 3 are repeated until convergence has been reached.

We use the number of characters n calculated before as the numbers of clusters k and record the centroid of each character. The result is shown in Figure 8.

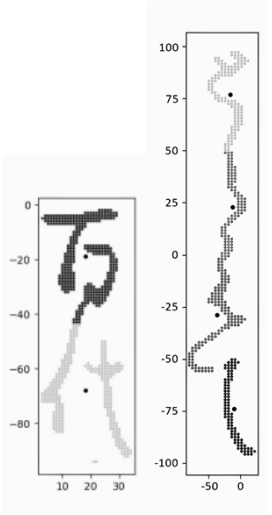

FIGURE 8. Result of K-means clustering

Then, we calculate the Euclidean distances s_1, s_2 between each turning point (x_i, y_i) with two neighboring centroids (x_1, y_1) and (x_2, y_2). The distances are calculated through the function below:

$$s_1 = \sqrt{(x_i - x_1)^2 + (y_i - y_1)^2}$$
$$s_2 = \sqrt{(x_i - x_2)^2 + (y_i - y_2)^2}$$

The turning point with a minimum difference between two neighboring centroids (x_1, y_1) and (x_2, y_2) will be chosen as the optimum cut point.

5. **Evaluation Experiment.** We use the Izutsuya version of Oku no Hosomichi, which was written by the famous Japanese poet Basho in the late 17th century as the input image to run the evaluation experiment. The final segment result is shown in Figure 9.

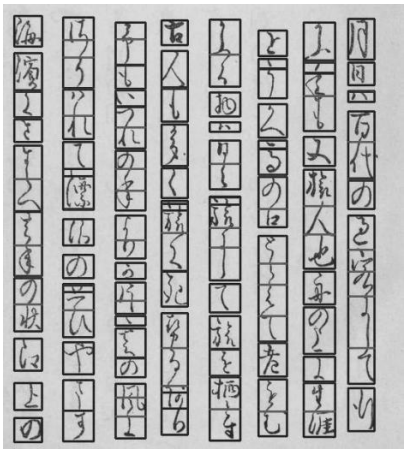

FIGURE 9. Result of segmentation

91 characters are detected from a total of 110 characters. This indicates that the segmentation has achieved a precision of 82.7%, demonstrating the feasibility of our method and surpassing previous studies with the highest precision of 80.3%.

6. **Conclusions.** In this paper, we proposed a method of automatic character segmentation from images of ancient Japanese books by using image processing and K-means clustering and confirmed the validity of the proposed method by the evaluation experiment. There are still problems to be improved. For example, when the height of character is far smaller than the average, the presume number of connected characters n will be miscalculated and the connected characters would not be separated correctly. The sample is shown in Figure 10.

FIGURE 10. A character not separated correctly

For the future, the system will be improved by calculating the presume number of connected characters more precisely.

Acknowledgment. This work was supported by JSPS KAKENHI Grant Number JP21K 11964.

REFERENCES

[1] I. Shoten, *General Catalog of National Books*, Iwanami Shoten, 2002 (in Japanese).
[2] CODH, *Dataset of Pre-Modern Japanses Text*, http://codh.rois.ac.jp/pmjt/20, 2019.
[3] *CODH: Center for Open Data in the Humanitie*, National Institute of Japanese Literature, https://www.nijl.ac.jp/21, 2019.
[4] ART Research Center Ritsumeikan University, *The Early Japanese Books Portal Database*, https://www.dh-jac.net/db1/books/search_portal.php, 2005.
[5] H. F. Schantz, *The History of OCR, Optical Character Recognition*, https://api.semanticscholar.org/CorpusID:59810624, 1982.
[6] K. Isshiki, Y. Muraki and K. Kobori, A method of Kuzushiji character recognition using connections feature, *FIT2020*, 2020.
[7] H. Gao, B. Lyu, Z. Wang and M. Lin, Kuzushiji segmentation with image processing and cluster analysis, *FIT2021*, 2021.
[8] B. Lyu, H. Li, A. Tanaka and L. Meng, The early Japanese books reorganization by combining image processing and deep learning, *CAAI Transactions on Intelligence Technology*, vol.7, no.4, pp.627-643, DOI: 10.1049/cit2.12104, 2022.
[9] A. Rehman, Neural computing for online Arabic handwriting recognition using hard stroke features mining, *International Journal of Innovative Computing, Information and Control*, vol.17, no.1, pp.177-191, 2021.
[10] E. Arias-Castro and D. L. Donoho, Does median filtering truly preserve edges better than linear filtering?, *Annals of Statistics*, vol.37, no.3, pp.1172-1206, 2009.
[11] S. M. Pizer, E. P. Amburn, J. D. Austin et al., Adaptive histogram equalization and its variations, *Computer Vision, Graphics, and Image Processing*, vol.39, no.3, pp.355-368, 1987.
[12] N. Otsu, A thresholding selection method from grey-level histogram, *IEEE Transactions on Systems, Man, and Cybernetics*, vol.9, no.1, pp.62-66, 1979.
[13] M. B. Dillencourt, H. Samet and M. Tamminen, A general approach to connected-component labeling for arbitrary image representations, *Journal of the ACM*, vol.39, no.2, pp.253-280, DOI: 10.1145/128749.128750, 1992.
[14] A. Willis and Y. Sui, An algebraic model for fast corner detection, *2009 IEEE 12th International Conference on Computer Vision*, pp.2296-2302, DOI: 10.1109/ICCV.2009.5459443, 2009.
[15] J. B. MacQueen, Some methods for classification and analysis of multivariate observations, *Proc. of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, pp.281-297, 1967.