Character Segmentation for Japanese Woodblock Printed Historical Books

Chulapong Panichkriangkrai
Graduate School of Science and Engineering
Ritsumeikan University

Liang Li
Ritsumeikan Global Innovation Research Organization
Ritsumeikan University

Kozaburo Hachimura
College of Information Science and Engineering
Ritsumeikan University

Abstract— **This paper describes methods of character segmentation for digitized Japanese historical woodblock-printed books. The segmentation method includes stain and smear removal, binarization, character lines extraction, and characters extraction by region labeling with integration and separation technique. The experimental results show that the proposed method can segment all text lines correctly and extracted more than 80% of the complicated characters.**

Keywords-Historical documents; layout analysis; text line extraction; digital archives

I. Introduction

The *Edo* period (1603-1867) was the turning point of printing culture in Japan, and publishing industry was developed very rapidly [1]. In that period, over 10,000 titles of books were published with more than 10 million copies on the markets. Wide varieties of books were published in response to the demands of an expanding reading population. However, currently only small number of people can recognize the characters in Edo book, and only small numbers of value book titles printed in Edo period were transcribed into modern book productions.

In that period, while in Europe moveable type printing process had been used, Japanese developed and used wood block printing process. In Japanese Edo book, every two pages were printed by a single engraved woodblock. With woodblock production, each character does not have the same size but may vary in shape for one similar character even in the same book.

In this paper, we propose a character segmentation system for character shape comparison, character images retrieval, and to make a statistical analysis of use of characters in a single or multiple books.

II. Research Problems

In this study, we focus on the books that have *Kanji* characters as main texts with *Furigana* which indicates pronunciation of some difficult *Kanji* characters to read. *Furigana* is written in *Hiragana* character, or Japanese syllabic character.

We use data of "*Chinsetsu Yumiharizuki*" (椿説弓張月), a novel written by *Kyokutei Bakin* (曲亭馬琴) from the document image database at Art Research Center, Ritsumeikan University [2], for this experiment. This book was woodblock printed in 1807 of Japanese Edo period. Fig. 1 shows a sample page image of the book and a part of a text line with a *Furigana* line.

The characters in this historical book have unique characteristics that make them difficult to segment. They had linkage line between couple characters in some part. In some case, they have conjunction area of two characters.

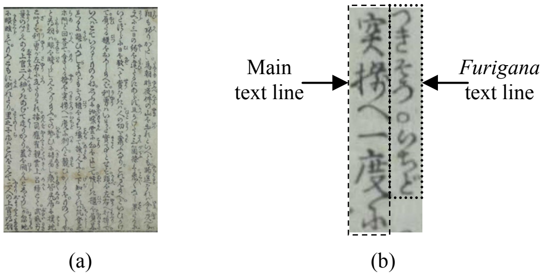

Figure 1. *Chinsetsu Yumiharizuki* (a) example of page image (b) example of a part of text line with *Furigana.*

III. Our Approach

In this study, we use process as shown in Fig. 2. The whole processing step consists of two main sub processes.

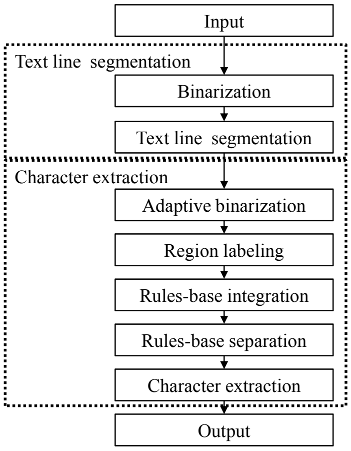

Figure 2. Flow of the process.

A. *Text Line Segmentation*

We use vertical projection data of binarized image obtained with the Otsu's method [3] for separating each text line. First, we find main text line using vertical projection (Fig.3). We found that there had a difference in data height between main text line and the *Furigana* text line, because *Furigana* are not attached to every *Kanji* words in the main text line. We set threshold at 50% of maximum vertical projection for segmenting the main text lines. Then, we find the *Furigana* text lines by finding maximum peak between main text lines. The width of *Furigana* line section is at 50% of maximum projection data found in each peak. At last, we find partitioning lines, which are used for separating each combination of main and *Furigana* text lines, by finding minimum data between the *Furigana* text line and next main text lines.

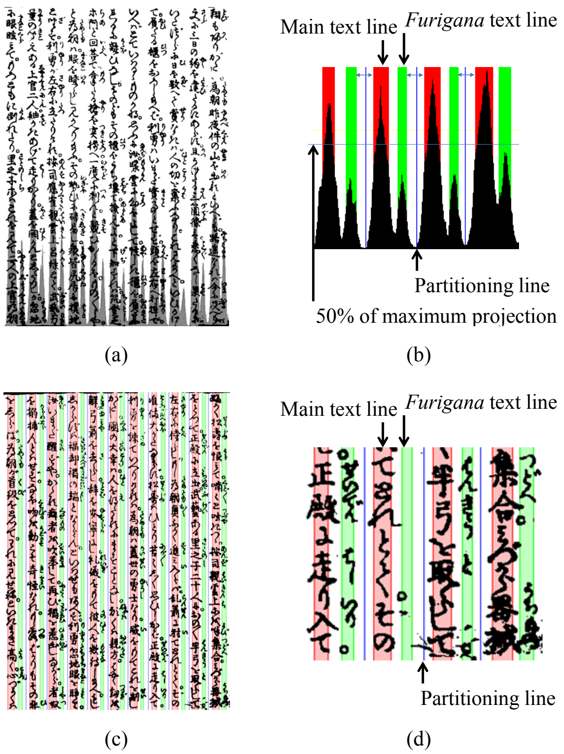

Figure 3. Text line segmentation: (a) projection data, (b) text lines separation, (c) partitioning line overlaid on original image, and (d) local magnification of (c).

B. *Character Extraction*

After line separation, we carry out adaptive binarization [4] in each rectangular area enclosed by consecutive partitioning lines. We then applied a region labeling to segmented main text lines to extract *Kanji* characters. Since the labeling may separates one character into several parts, we apply integration rules to merge several labeling areas into one character (Fig. 4). First we made a bounding box around labeling area. Then we applied rules: if two boxes have overlapped area with other box more than a threshold (Fig. 4 (a)), or if the space between two boxes is under threshold, they would be merged into one.

On the other hand, some labeling areas contain two or more characters. We separate them into individual characters, if the height of the bounding box is greater than the threshold. We separate them by finding a local minimum on horizontal projection data (Fig. 4 (b)). Fig. 4 (c) shows the comparison between original data and character extraction results.

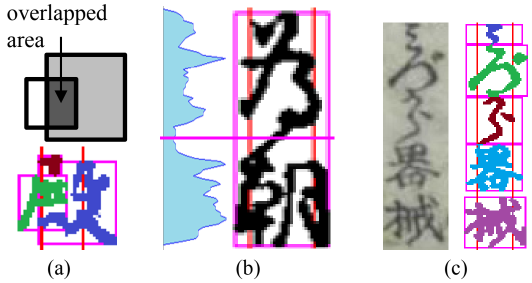

Figure 4. Character extraction: (a) example of character integration, (b) example of character separation, and (c) example of character extraction result.

IV. EXPERIMENTAL RESULTS

We applied the proposed method to 12 pages of *Chinsetsu Yumiharizuki*, which contain 132 text lines with 3885 characters in the main text lines. As a result, all lines were correctly segmented in both main *Kanji* text lines and sub *Furigana* lines. More than 3100 characters (80%) were correctly extracted.

V. DISCUSSION AND FUTUREWORK

From the result, we could get successful result in both text line segmentation and character extraction result. However, some characters still cannot be correctly extracted. To improve the extraction performance, we will investigate optimum parameter determination for integration and separation by using adaptive methods.

In addition, we will make character shape comparison and character images retrieval. Then we would like to apply character spotting to search the similar character images and make a concordance or index of character image.

REFERENCES

[1] K. Hioki, “Japanese printed books of the Edo period (1603-1867): History and Characteristics of Block-Printed Books”, Journal of the Institute of Conservation, vol. 32, No. 1, pp. 79-101, 2009.

[2] Art Research Center Ritsumeikan University Database, Website : http://www.arc.ritsumei.ac.jp/dbroot/top.htm

[3] N. Otsu, “A thresholding selection method from grey-level histogram”, IEEE Trans. Systems Man Cybernet, SMC-8, pp. 62-66, 1979.

[4] B. Gatos, I. Pratikakis, S. J. Perantoni, “An adaptive binarization technique for low quality historical documents”, Proc. IAPR International Workshop on Document Analysis Systems (DAS 2004) LNCS 3163, pp. 102–113, 2004.