Detecting Character in Historical Documents by Pixel-Level Labeling

Ryohei Nishimura, Tomo Miyazaki, Yoshihiro Sugaya, Shinichiro Omachi
Department of Communications Engineering, Graduate School of Engineering, Tohoku University
6-6-05, Aoba Aramaki, Aoba-ku, Sendai, 980-8579 Japan
{ryohei,tomo,sugaya}@iic.ecei.tohoku.ac.jp, machi@ecei.tohoku.ac.jp

Abstract—**Attempts have been made to detect and recognize characters in images of historical documents in order to utilize information written in historical documents. However, many of the conventional character detection methods are based on the premise that the individual characters are separated. It is difficult to detect characters with high accuracy by this approach. In order to improve the detection accuracy, we propose to learn the pixel-level labeling by reducing the character area, and to re-learn based on the output result. The experimental results show that the detection accuracy is improved by the proposed method.**

Keywords-component; Convolutional neural network, semantic segmentation, detecting character

I. INTRODUCTION

The historical document is a valuable source of information, and attempts have been made to utilize the information written in the historical document. However, because the font is different from the current characters and the amount of information is enormous, it is not practical to recognize the whole sentence in advance and convert it to text. To do so, it is first necessary to detect characters from the historical document images. Many of the conventional character detection methods are based on the premise that the individual characters are separated. However, since the characters in the historical documents are distorted and there is an overlap between the characters, it is difficult to detect characters with high accuracy using conventional method.

In this paper, we propose a character detection method using object detection technology for the purpose of realizing a keyword retrieval system for historical document images.

II. RELATED WORK

U-Net[1] is a network specialized in image semantic segmentation. Semantic segmentation is an image analysis task to assign labels to pixel and it outputs a mask image surrounding the object area for the input image. U-Net is a kind of Fully Convolutional Network(FCN). The difference from general FCN is that the information used in encode is passed when decoding the convolved image. It enables high-accuracy classification by combining features obtained in deep layers and positional information obtained in shallow layers with FCN.

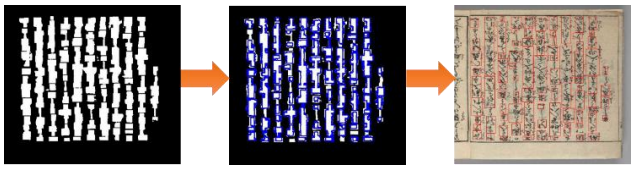


Figure 1. The procedure for character detection

III. METHOD

A. U-Net based detection method

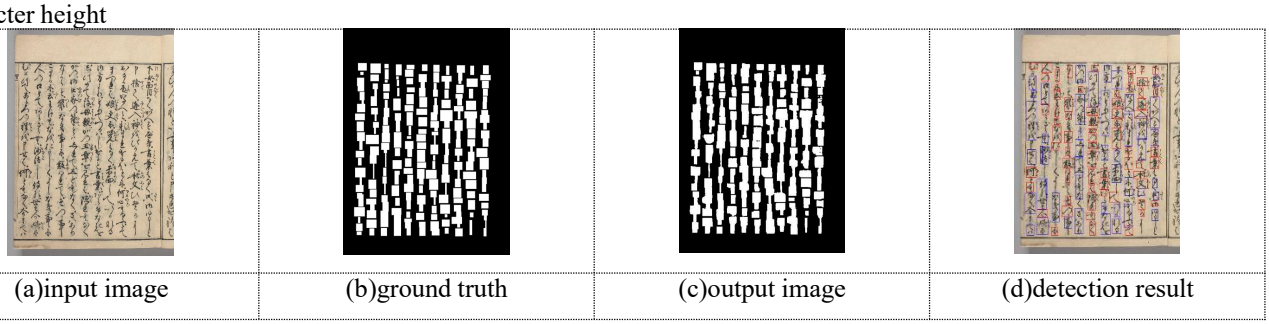
When using U-Net to detect characters from a historical document image, it is possible to learn by giving a mask image in which the rectangle surrounding the character are painted white and the historical document image. During detection, characters are detected by drawing a rectangle that encloses the outline of the character area in the output mask image. However, with this method, it is difficult to detect individual characters from character strings that overlap each other. In this paper, we propose to shrink the rectangle of the character area in the mask image so the inferred character area can be separated.

For learning and testing, we use a classic Japanese kanji character set [2]. This data set consists of many historical documents, and there is annotation data indicating the position and type of characters in the image. Based on this data, ground truth and training mask images are created. For learning, we use 1876 images consisting of 14 works. At this time, for the purpose of detection only, we study all characters as the same class. After resizing all images to 516 * 516, batch Set the size to 16 and learn 100 epochs. The test uses 346 images of one work that was not used for learning. The procedure for character detection is shown in Fig. 1. Output from U-Net Detection is performed by extracting the contour of the mask image and drawing a circumscribed rectangle surrounding the contour on the input image. In this paper, this method is used as an existing method and compared with the proposed method.

B. Proposal 1 : Learning the pixel-level labeling by shrinking the character area

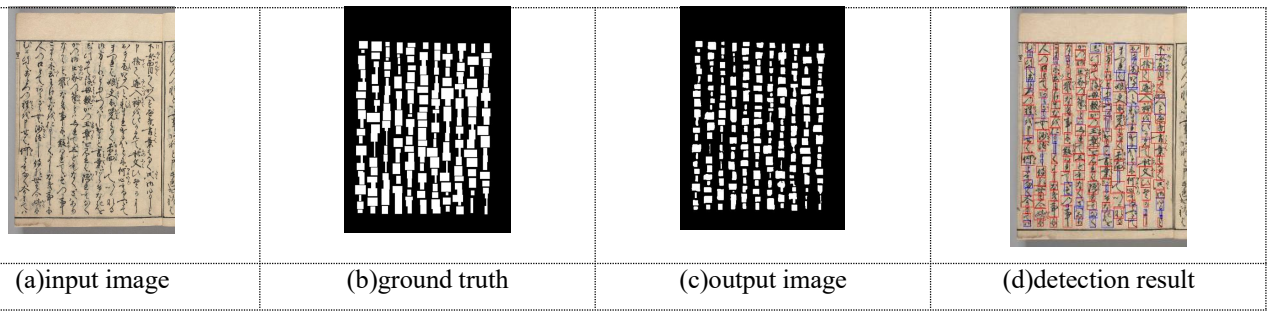
It is difficult to detect individual characters from character strings that overlap each other with exiting method. In this paper, we propose a method for learning by reducing the rectangle of the character area in the mask image given during learning. When creating a training mask image from annotation data, we will consider dividing the character area of the

training mask image one character at a time by multiplying the character height



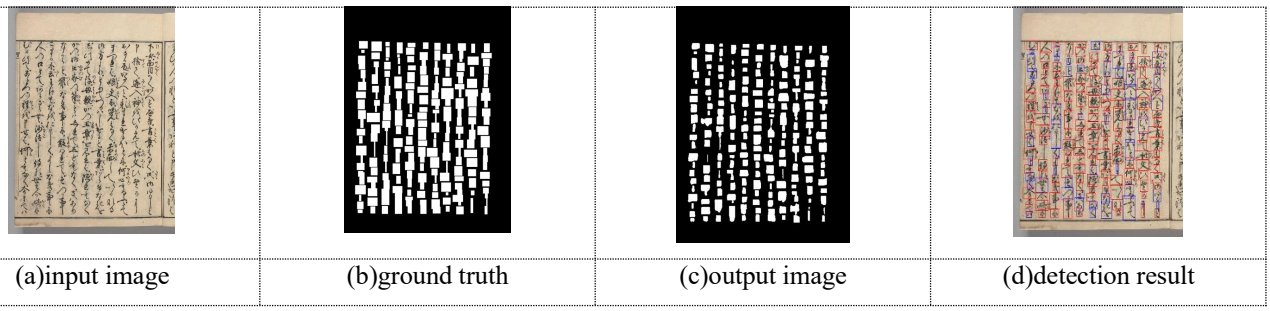
(a)input image (b)ground truth (c)output image (d)detection result

Figure 2. Result from exiting method



(a)input image (b)ground truth (c)output image (d)detection result

Figure 3. Result from proposed method 1



(a)input image (b)ground truth (c)output image (d)detection result

Figure 4. Result from proposed method 2

and width by 0.8. Further, when detecting a character, the height and width of the rectangle drawn in the input image are each multiplied by 1.25 so that the size is the same as the correct character size.

C. Proposal 2 : Re-learning based on output image

In the existing method and the proposed method 1, a mask was created from annotation data. In Proposed Method 2, the output image from Proposed Method 1 is used as a training mask image, and the network is learned again. This compares the artificially created mask image and the U-Net output image that are more suitable for learning.

IV. RESULT

A. Output image comparison

The results of using U-Net as an existing method are shown in Fig. 2. Fig. 2 (a) shows the input image, Fig. 2 (b) shows the ground truth, Fig. 2 (c) shows the output image, and Fig. 2 (d) shows the detection result. Although it seems that the output image is close to the ground truth, there are many parts where the character areas are connected, and since the connected area is detected as one character, false detection was seen.

Fig. 3 shows the results of using the proposed method 1 of learning by reducing the rectangle surrounding the outline of the character area. Fig. 3 (a) shows the input image, Fig. 3 (b) shows the ground truth, Fig. 3 (c) shows the output image, and Fig. 3 (d) shows the detection results. By learning by reducing the character area, the character area is output smaller, more parts can be detected one by one, and the detection accuracy is improved.

Fig. 4 shows the detection results of Proposed Method 2 using a network that has been relearned based on the output image. Fig. 4 (a) shows the input image, Fig. 4 (b) shows the ground truth, Fig. 4 (c) shows the output image, and Fig. 4 (d) shows the detection results. The appearance of the output

image and detection result seemed not much different from that of proposed method 1.

B. Comparison of evaluation indicators

Recall, Precision, and F-score by the existing method, the proposed method 1 and proposed method 2 are shown in Table 1. From these values, it can be seen that the detection accuracy is improved by reducing the character area and learning. Further, since the detection accuracy is improved by re-learning, it can be read that the output image by U-Net is easier to learn for the network than the artificially created mask image.

TABLE I. COMPARISON OF EVALUATION INDICATORS

Method	Recall	Precision	F-score
Exiting	0.1865	0.4597	0.2654
Proposal 1	0.6147	0.7114	0.6595
Proposal 2	0.6334	0.7335	0.6797

V. CONCLUSION

Using U-Net, we proposed a method to reducing the rectangle of the learning mask image and re-learning using the output image as the training image. These methods improved the accuracy compared to the conventional method.

In the future, we plan to improve the detection accuracy by examining learning methods and contour extraction methods for parts where character regions are not separated.

[1] Olaf Ronneberger , Philipp Fischer , and Thomas Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation", MICCAI 2015, 234-241.

[2] classic Japanese kanji character set, doi:10.20676/00000340, Center for Open Data in the Humantities, http://codh.rois.ac.jp/