Improving OCR for Historical Documents by Modeling Image Distortion

Maekawa, Keiya
Kyushu University

Tomiura, Yoichi
Kyushu University

Fukuda, Satoshi
Kyushu University

Ishita, Emi
Kyushu University

他

出版情報：Digital Libraries at the Crossroads of Digital Information for the Future. 11853, pp.312-316, 2019-10-29. Springer Nature
バージョン：
権利関係：

Improving OCR for Historical Documents by Modeling Image Distortion

Keiya Maekawa, Yoichi Tomiura, Satoshi Fukuda, Emi Ishita and Hideaki Uchiyama

Kyushu University, 744 Motooka Nishi-ku Fukuoka, Japan
maekawa.keiya.946@s.kyushu-u.ac.jp
tom@inf.kyushu-u.ac.jp
fukuda.satoshi.528@m.kyushu-u.ac.jp
ishita.emi.982@m.kyushu-u.ac.jp
uchiyama.hideaki.667@m.kyushu-u.ac.jp

Abstract. Archives hold printed historical documents, many of which have deteriorated. It is difficult to extract text from such images without errors using optical character recognition (OCR). This problem reduces the accuracy of information retrieval. Therefore, it is necessary to improve the performance of OCR for images of deteriorated documents. One approach is to convert images of deteriorated documents to clear images, to make it easier for an OCR system to recognize text. To perform this conversion using a neural network, data is needed to train it. It is hard to prepare training data consisting of pairs of a deteriorated image and an image from which deterioration has been removed; however, it is easy to prepare training data consisting of pairs of a clear image and an image created by adding noise to it. In this study, PDFs of historical documents were collected and converted to text and JPEG images. Noise was added to the JPEG images to create a dataset in which the images had noise similar to that of the actual printed documents. U-Net, a type of neural network, was trained using this dataset. The performance of OCR for an image with noise in the test data was compared with the performance of OCR for an image generated from it by the trained U-Net. An improvement in the OCR recognition rate was confirmed.

Keywords: OCR Error, Information Retrieval, Historical Document Image

1 Introduction

In the modern age, information is often provided and circulated as digital data. In contrast, many archives and libraries still hold printed historical records and documents. Even in those institutions, some of the records have already been digitized: many of them are not in text form but only exist as image data. Our ultimate goal is to help researchers search these documents effectively and efficiently to find documents related to their information needs. However, image format is not suitable for searching for information. It is necessary to extract text information from the images to enable search for information related to researchers' needs. We focus on old records and documents, as shown in Fig. 1, most of which have deteriorated. In this paper, we call an image of

a deteriorated document a *distorted image*. It is difficult to automatically convert distorted images to correct text with optical character recognition (OCR). This causes low accuracy of information retrieval. In order to solve this problem, not only development of robust search method [1] for text containing OCR errors, but also improvement of the performance of OCR for a distorted image [2] is necessary.

In this research, we attempt to improve the recognition accuracy of OCR, by preprocessing a distorted image with a neural network to generate a clear (undistorted) image for input to the OCR. We need to construct a very large amount of training data to train the neural network that converts the image. It is difficult to manually create clear images not containing deterioration from distorted images. However, it is relatively easy to automatically add noise to clear images so that the created images mimic the deterioration in the actual historical documents.

We collected portable document format (PDF) files of official documents that did not exhibit deterioration, automatically added noise mimicking deterioration to the images of the collected documents and created pairs combining each distorted image with the original clear image. Then, using a part of the dataset, we then trained U-Net, a type of neural network, to generate a clear image from a distorted image. Finally, we investigated the effect of preprocessing by the trained neural network, using the rest of the dataset. The experimental results indicated a significant improvement in the recognition accuracy of OCR by preprocessing with the trained U-Net.

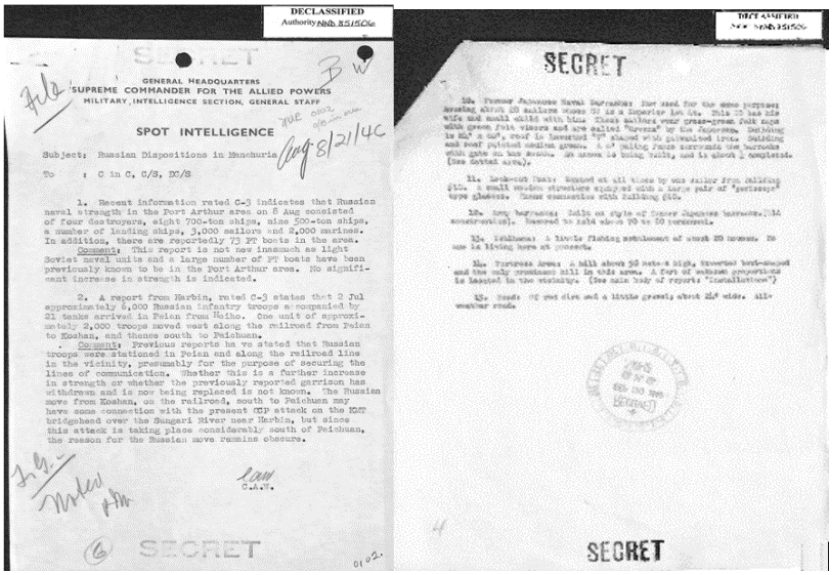

Fig. 1. Real historical printed material images

2 Generating Distorted Image

The target of this research is the collection of images of the historical documents of RG554 in the digital collection of the Japan National Diet Library (Fig. 1). The original

RG554 is preserved at the U.S. National Archives. By observing these distorted images, we confirmed that the distortions were due to partial paper discoloration, contrast reduction, feathering, ghosting, and noise caused by dust and blur, as shown in Fig. 2. We generated distorted images by adding noise imitating these deteriorations to clear images of documents having a format similar to the target documents. We added Gaussian noise (mean = 0, standard deviation = 0.2) to imitate paper discoloration; added salt-and-pepper noise (replace rate = 0–0.01) before smooth blurring (filter size = 3–5) to imitate feathering, blur of character, and dust on scanning; and reduced contrast (by converting the pixel value range from [0,255] to about [50,210]) randomly.

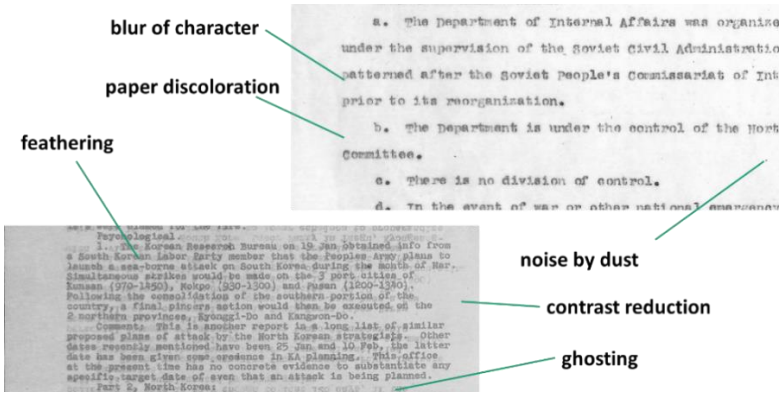

Fig. 2. Types of distortion in historical printed material images

3 Experiment

3.1 Experimental Setting

U-Net, a type of CNN composed of only convolution layers, was used for the architecture of the neural network in the experiment. The training data described in the previous section were used as input data and label data, because both input and output were images [3]. The output image was the binary conversion of the input image. The optimization algorithm for mini-batch learning was set to Adam, the base learning rate was set to 10^{-4}, and binary cross entropy was used as an error function.

First, we collected 27 PDFs of documents as clear images, through the United Nations official document retrieval system. The format of these documents was similar to the target documents. The total number of pages in the collected documents was 440. We then converted them to images with a size of 1024 × 1024 in the JPEG format, which we regarded as clear (undistorted) images. We generated distorted images from these clear images using the method described in Section 2. Finally, we converted the image files to text using a commercial OCR software, for evaluation purposes.

We trained U-Net using 404 pairs comprising a distorted image and its original clear image. We measured the error rate of OCR for each of the remaining 36 distorted images. We then measured the error rate of OCR for each of the images generated from

the 36 distorted images by the trained U-Net. To calculate the OCR error rate, we used the text generated by OCR from the original clear images as the correct data.

3.2 Result and Discussion

We compared the error rate of OCR for text generated from two categories of images: *Noise* and *Predict*. *Noise* means the images to which noise (distortion) has been added, as described in Section 2, and *Predict* means the images generated from the images in *Noise* by the trained U-Net. We used the evaluation tool *ocrevalUAtion* for the calculation of character error rate (CER) and word error rate (WER). Table 1 shows the results of the experiment. This indicates that removing the image noise by the trained U-Net increased the accuracy of OCR.

Table 1. CER and WER of generated text

Data image	CER	WER
Noise	**78.98**	87.49
Predict	**9.46**	30.20

4 Conclusion and Future Work

We proposed a method for converting images of deteriorated historical printed documents to text. In our method, images of deteriorated documents are converted to images not containing deterioration, and text is extracted from the converted images using a commercial OCR software. The experimental results indicated that our approach has great potential. An OCR system based on a CNN architecture could directly extract text from deteriorated document images. However, this approach would also require numerous pairs comprising a deteriorated image and the extracted text from it, for training. Our approach could be effectively adapted to this case. In the future, we will apply a method for creating distorted images that further mimic the actual deterioration, using Cycle-GAN [4].

References

1. Ghosh, K., Chakraborty A., Parui S.K., and Majumder, P.: Improving Information Retrieval Performance on OCRed Text in the Absence of Clean Text Ground Truth. Information Processing and Management, 52(5), 873–884 (2016).
2. Chen, Y. and Wang, L.: Broken and Degraded Document Images Binarization. Neurocomputing, 237, 272–280 (2017).
3. Ronneberger, O., Fischer, P., and Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: MICCAI, 9351, 234–241 (2015).
4. Zhu, J.-Y., Park, T., Isola, P., and Efros, A.A.: Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In: ICCV, pp. 2223–2232, (2017).