Recognition of Japanese Historical Hand-Written Characters Based on Object Detection Methods

Tang, Yingping
Department of Informatics, Kyushu University

Hatano, Kohei
Faculty of Arts and Sciences, Kyushu University : Associate Professor

Takimoto, Eiji
Department of Informatics, Faculty of Information Science and Electrical Engineering, Kyushu University : Professor

https://hdl.handle.net/2324/2545009

出版情報：Proceedings of the 5th International Workshop on Historical Document Imaging and Processing, pp.72-77, 2019-09-20. ACM
バージョン：
権利関係：

Recognition of Japanese Historical Hand-Written Characters Based on Object Detection Methods

Yingping Tang
Department of Informatics
Kyushu Uniersity
tang.yiping.641@s.kyushu-u.ac.jp

Kohei Hatano
Faculty of Arts and Science
Kyushu Univerisity
hatano@inf.kyushu-u.ac.jp

Eiji Takimoto
Department of Informatics
Kyushu University
eiji@inf.kyushu-u.ac.jp

ABSTRACT

We consider the recognition problem of Japanese historical hand-written characters called "Kuzushiji". Unlike modern characters, Kuzushiji characters are harder to recognize partly because many of them are connected and not separated by spaces without any segmentation information. We propose two methods for segmentation and recognition of Kuzushiji characters. The first method learns segmentation rules and character classifiers simultaneously from data sets with character labels and segmentation information. Second method is for segmentation and can be used with any single character recognizer. Our methods outperform other baselines and achieve the state-of-the-art accuracy on both segmentation and recognition tasks on data sets of three consecutive Kuzushiji characters.

1 INTRODUCTION

The digital humanities are interdisciplinary fields studying humanities with the aids of information technology. In recent years, the digital humanities are getting more attentions from various fields , not only from humanities, but also natural language processing, data mining, image recognition, or machine learning. The technology for recognizing ancient or pre-modern hand-written characters is a key component to promote advances in the digital humanities.

The digital humanities are becoming popular in Japan as well, due to developments of various data sets by, e.g., the CODH(the Center for Open Data in the Humanities) [1]. Japanese historical literature is written with hand-written characters, called "Kuzushiji." There are increasingly more digitized images of Japanese pre-modern literature, however, recognizing texts from those images are not fully automated yet even with the state-of-the-art OCR technologies.

Compared to other languages for which recognition is successful (e.g., hand-written English[8] and Arabic [1]), recognition tasks of Kuzuji-ji characters are challenging in that (i) characters are often connected without explicit spaces, (ii) different symbols such as Chinese and Japanese ones are used to mean the same character, (iii) characters are often simplified or abbreviated.

More precisely, the difficulty of recognizing Kuzushiji stems from the hardness of segmenting sentences to characters. Once a sentence is segmented correctly to single characters, recognition of single charaters is rather tractable. For example, the CODH reported that the recognition rate of single characters by a deep neural network is 97.33% on 49 kinds of Japanese Hirakana of the data set published by the CODH [3]. The current best result for multiple characters recognition is achieved by Nguyen et al. [11], which is the winner of the PRMU contest[2] on Kuzushiji recognition. Their method achieves 12.88% error rate on three consecutive Kuzushiji characters. Their method is an end-to-end method consist of feature extractor and recognizer. Roughly speaking about their feature extractor, given an image of multiple characters (in one column), their method (i) makes multiple segmented images by running sliding windows of different sizes on the input, (ii) feeds each segmented image to a single character recognizer and get confidence rates of all possible characters, and (iii) the output of recognizer is a sequence of extracted features from top to bottom is processed by a Bidirectional Long Short-Term Memory network (BLSTM) to extract the relation information between sliding windows. (iiii) the output sequence of BLSTM is decoded by a Connectionist Temporal Classification (CTC)-based transcription layer. Their method, however, seems to have room for improvement. Their segmentation candidates are made with heuristics and are not optimized or learned from data. But this time, we want to try to use a completely image-oriented approach to achieve character recognition.

In this paper, we propose recognition methods for Kuzushiji characters based on object detection techniques. The first method learns segmentation rules and character classifiers simultaneously from data sets with character labels and segmentation information. To better utilize the side information of segmentation, we use a segmentation data set constructed from the CODH data set and the PRMU contest data set[17]. The data set is obtained by adding the infromation of bounding boxes segmenting characters to the PRMU contest data sets. Our technique is motivated by object recognition methods such as YOLOv3 [13], which, given images with labels and bounding boxes (as information of location) of objects as ground truth, learns a classifier predicting labels and bounding boxes of candidates of objects. We observed, however, that a naive application of such object recognition methods was not successful enough, since each character often consists of several parts and YOLOv3 tends to output only some of the parts as candidates. One of our technical contributions is to devise a better aggregation method of possible candidates of bounding boxes and associated labels suited for Kuzushiji character recognition.

Second method is a segmentation method and can be used with any single character recognizer. Our methods outperform other

[1] http://codh.rois.ac.jp/index.html.en

[2] https://sites.google.com/view/alcon2017prmu (in Japanese).

HIP '19, September 20–21, 2019, Sydney, NSW, Australia

ACM ISBN 978-1-4503-7668-6/19/09…$15.00
https://doi.org/10.1145/3352631.3352642

baselines and achieve the state-of-the-art accuracy on both segmentation and recognition tasks on data sets of three consecutive Kuzushiji characters.

In the experiments on recognition of three consecutive Kuzushiji characters, our first method outperforms baselines including a naive application of YOLOv3 (an illustration of our recognition results is shown in Figure 1). Our first method shows the advantage of learning detectors from data sets with labels and segmentation information. Our second method shows worse recognition performance compared to the first one, but achieved better segmentation accuracy.

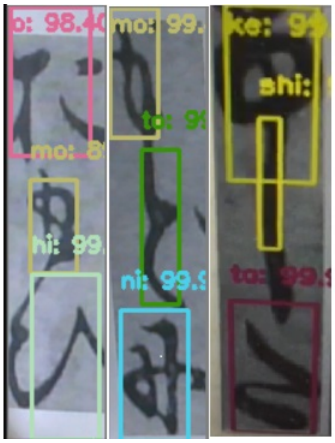

Figure 1: Illustration of recognizing and locating characters.

2 RELATED WORK

2.1 CODH data sets

In 2017, the CODH has published open data sets of Japanese historical characters and literature [3]. On January 2019, the data sets consist of images of pre-modern Japanese books and their transcriptions as well as data sets of images of 684,165 individual characters of 4,645 different types. Further, the CODH published two data set of Kuzushiji characters in 2018[4]. One of them is named Kuzushiji-49, consisting of 49 different kana characters. The other one is called Kuzushiji-Kanji, containing 3,832 Japanese kanji characters (Chinese characters). Clanuwat et al. also carried on the recognition job with these two data sets. They achieve the accuracy rate of 97.33% in Kuzushiji-49, and achieve the accuracy rate of 98.83% in Kuzushiji-MNIST for single character recognition.

2.2 PRMU algorithm contest data sets and their Kuzushiji recognition

The PRMU (Pattern Recognition and Media Understanding), a Japanese association for pattern recognition, organized a contest for recognition of Kuzushiji characters in 2017. In this contest, PRMU published learning data sets based on CODH's, which truncated the part of the original image with three or more consecutive Japanese Kana characters. The data sets contain 46 kinds of Japanese hiragana characters, and have three types of data sets depending on difficulty, from level 1 to 3 (see Figure 2 for an example). The level 1 images composed of pictures containing a single character. level 2 images composed of pictures containing three characters, and level 3 images composed of pictures containing more than three characters.

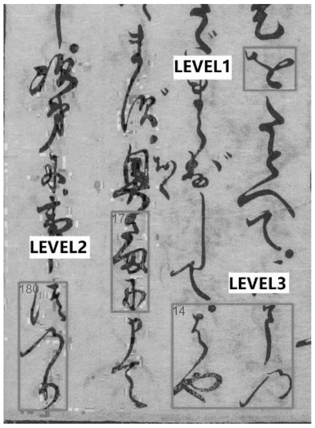

Figure 2: Examples of Kuzushiji separated data set.

Nguyen et al. won the contest [11]. They constructed a sliding windows to make proposal bounding boxes, and used CNNs to extract the character features from boxes. Their method achieved 12.88% label error rate (LER, defined in (4.2)) for three consecutive characters recognition tasks.

2.3 Object detection

The object detection techniques, such as Faster-RCNN[14], SSD[10], YOLO[12] and YOLOv3[13], can automatically retrieve and locate objects in video or images, such like automatic driving and pedestrian detection. Speaking of the object detection, there are two main types, the one-stage approach and the two-stages approach. The two-stages approach, such like RCNN[7] and Fast-RCNN[6], usually uses some techniques to find a large number of bounding boxes enclosing objects before recognition. Some of these techniques focus on color and texture changes between the target and the background in the image, or look up the selected box by using pixel distributions or heat maps. But in the case of Kuzushiji character recognition, it seems to be very hard for this approach to search bounding boxes since images are monochrome and contain few color and texture information. On the other hand, the one-stage approach, suck like YOLO[12], can learn to segment and recognize objects at the same time based on a neural network. Therefore, we consider to apply object detection techniques for recognizing and locatiing Kuzushiji characters.

3 PROPOSED METHODS

3.1 Frame Group Decision Method(FGDM)

Our first method, which we call FGDM (Frame Group Deicision Method) is based on YOLOv3[13], a modification of YOLO[12]. YOLOv3

[3] http://http://codh.rois.ac.jp/pmjt/(in Japanese)

is implemented by a deep neural network. YOLOv3 learns a predictor, which consists of pre-defined fixed grids on the input image. Each grid corresponds to a sub-predictor, which outputs a fixed number of bounding boxes and confidence rates of all possible objects (characters) for each box. YOLOv3 has two phases, training and prediction phases, respectively. In the training phase, given images with ground truth (pairs of bounding boxes and their associated labels), YOLOv3 tries to optimize a loss function which takes into account of classification errors, and location errors. Due to the space constraints, we omit the details of the loss function (see, [12] for details).

In the prediction phase, given a image, YOLOv3 applies the learned predictor to the input, and obtains bounding boxes with confidence rates from grids. Since there might be too many bounding boxes with low confidence rates, YOLOv3 remove such boxes with the following rule. YOLOv3 remove a bounding box if its maximum confidence rate for some object (character) is lower than some theshold. The rule is called Non-Maximal Suppression(NMS). The remaining boxes having maximum confidence rate for some character label are the output of YOLOv3. The NMS rule works well for general object detection. We observed, however, that in the case of Kuzushiji recognition, the NMS rule often eliminates correct bounding boxes with low confidence rate for the correct label. This might be due to the fact that in Kuzushiji recognition tasks characters are densely located while in general object detection tasks objects are sparsely located.

So, we propose a different aggregation method of bounding boxes. Compared to general object detection tasks, input images of Kuzushiji text are much simpler, since the images do not contain complicated background elements (just white background) . This motivates us to devise the following method:

Characters in the image are often arranged on lines from the top to thebottom. So, we can simply separate different columns of the text first. The specific implementation is shown below. Here, for simplicity, we assume the prior knowledge of number K of characters in the given image.

Step1 : Remove bounding boxes with lower confidence with some threshold.

Step2 : Perform clustering with K group for bounding boxes in terms of vertical positions of their centers (Figure 3 shows an illustration of the result) .

Step3 : For each group, choose the bounding box with the highest confidence and the associated character.

Note that our prior assumption can be relaxed. Even if the number of characters is unknown, if the distribution of the frames is balanced, the K-means++method[2] could be used to infer the number of groups. The key difference to the YOLOv3's original aggregation method is that we consider each different group as the attention area for the original image, and choose boxes of high confidence in a local sense. This change contributes higher accuracy in Kuzushiji recognition, which we will show in the experiments later.

3.2 Frame Decision Method Based on Selective Search(FDMBSS)

The selective search is a useful method in two-stage object detection research[15], such like RCNN[7] and Fast-RCNN[6], this technique can get lots of candidates of bounding boxes for characters quickly. The selective search is a method of making some bounding boxes of objects based on the nearby pixels. For example, a object in an image can be separated from the background, according to the color expression. After that, if the curve separating the object and the background is replaced with straight lines, from which rectangles are made. Then, we can use the frame combination method[16] to create boxes in the image. An illustration of bounding boxes constructed by the selective search (marked as A) and by FDMBSS (marked as B) are shown respectively in Figure 4.

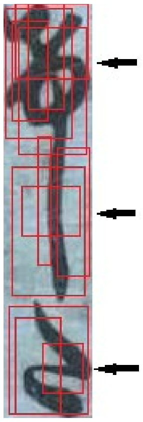

Figure 3: Illustration of small groups of bounding boxes formed by the FGDM. The details of the FGDM is shown on page 3 (Step 1 to 3).

When the selective search method is used for the Kuzushiji characters, we will find that the degree of freedom of the frames is very high. There are too large or too small boxes, and there are boxes that do not contain text pixels, which we automatically delete. At the same time, there will be extremely good frames among them, albeit in a limited number. And, we found that for hand-written characters like Kuzushiji, the histogram method can often find a nearly suitable frame. But it is hard to get a frame that fits perfectly with the character.

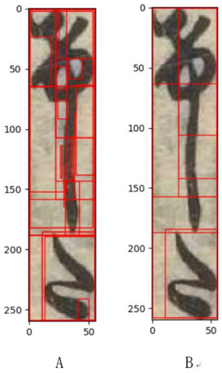

Figure 4: Illustration of the bounding boxes forKuzushiji characters constructed by the selective search.

We propose a special segmentation framework named FDMBSS to constrain the selective search by using the histogram method.

First, we use the selective search to obtain candidates of bounding boxes of characters. Then, we remove a box if it is too small or too large (w.r.t. some thereshod on the size) .As a result, only boxes with adequate size will remain.

Second, we use the histogram method. Given a number of characters K as a parameter, the histgram method counts the black pixels of each horizontal line of the image and constructs a histgram. Then it divides the image by $K-1$ horizontal lines, where the lines are determined based on the $K-1$ lowest points in the histgram.

Third, we compare bounding boxes obtained by the two methods, the selective search and the histgram method. For each bounding box (X, Y, W, H) of the histogram method ((X, Y) specifies the center and (W, H) specifies the widht and height), we calculate the sum of absolute differences of these 4 parameters between the outputs of the selective search and the histogram method. If the sum of the error is smaller than a threshold, we employ the bounding box of the selective search. Otherwise, the bounding box of the histgram method is employed.

Finally, after we segment characters by the FDMBSS method, we apply single-character recognizers, e.g., simple networks such as CNN, RNN, RNN+LSTM, and complex network such as RESNET-53[9], to obtain a label of each segmented character.

4 EXPERIMENTS

We conduct preliminary experiments for evaluating proposed methods and baselines. The cpu in our computer is Intel(R) Xeon(R) Gold 2.30GHz with 9 cores and gpu is NVIDIA Tesla P4000 with 1 node.

4.1 Data Sets and Evaluation Critetia

We use the data set of Tang et al. [17]. The data set is constructed by combining the CODH data sets and the PRMU contest data. The data set contains 77,000 images of three consecutive Kuzushiji characters (Level 2 data set of the PRMU contest) with information of bounding boxes of characters and labels. We split the data to 70,000 training images and 7,000 test images.

We define two evaluation criteria for measuring accuracy on predicting labels and bounding boxes.

Each bounding box i is specified with four parameters, position of the center (X_i, Y_i) and its width W_i and height H_i, respectively. Given the sequence of predicted bounding boxes $\hat{S} = ((\hat{X}_i, \hat{Y}_i, \hat{W}_i, \hat{H}_i)|i = 1, 2, \dots m)$ and ground truth bounding boxes $S = ((X_i, Y_i, W_i, H_i)|i = 1, 2, \dots m)$(as shown in Fig- ure 4), the consistency rate (CR) of the predicted sequcence of boxes is defined as follows:

$$\epsilon\text{-Y-CR} = \frac{\sum_{i=1}^{m} \mathrm{I}(|Y_i - \hat{Y}_i| \le \epsilon \times \tilde{H})}{m} \times 100\%, \tag{4.1}$$

where $\tilde{H}$ is the height of the input image, and I is the indicator function and outputs 1 if the argument is true and outputs 0, otherwise. The parameter ϵ is for the degree of tolerance. Here, CR only focuses on differences of bounding boxes in the vertical direction, which is sufficient for our purpose.

We also use an evaluation criterion for character recognition. The label error rate(LER) of a pattern classifier h w.r.t. a set Z of labeled instances (x, y) (x corresponds to an image and y corresponds to a sequence of labels) is defined as follows:

$$\mathrm{LER}(h, Z) = \frac{1}{\mathrm{T}} \sum_{(\mathrm{x,y}) \in Z} \mathrm{ED}(h(x), y), \tag{4.2}$$

where T is the total number of target labels in Z and $\mathrm{ED}(p, q)$ is the edit distance between two sequences p and q.

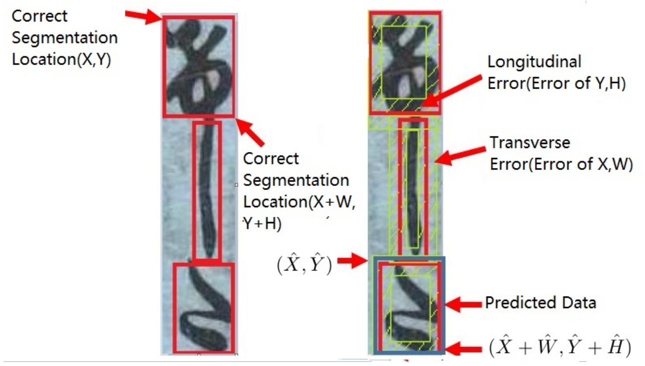
Correct Segmentation Location(X,Y)
Correct Segmentation Location(X+W, Y+H)
$(\hat{X}, \hat{Y})$
Longitudinal Error(Error of Y,H)
Transverse Error(Error of X,W)
Predicted Data
$(\hat{X}+\hat{W}, \hat{Y}+\hat{H})$

Figure 5: Examples of evaluation methods to tolerate coordinate errors, showed an illustration of ground-truth and predicted bounding boxes in the Kuzushiji character image.

4.2 Results

For the backclone network of YOLOv3, the darknet53[13] model is used and trained with the training set of 70, 000 images.

Regarding the recognizer of FDMBSS, we use the RESNET-53[9].

As comparator or baselines, we use YOLOv3-darknet53, which employs the YOLOv3's original aggregation method for bounding boxes. Also, for a naive baseline, we use a segmentation method which segments characters based on histograms along horizontal coordinates, combined with a single character recognizer implemented with darknet53. We resize all images in 64*288 and divide the original image into 2×9 grids to form some pre-predicted areas that one of areas must contain a object and let the anchor as (10,60, 15,20, 20,40, 35,45, 30,60, 40,55, 40,80, 50,90, 60,100) depend on the shape of Kuzushiji characters.

The results are summarized in Table 1.

Table 1: Results of character recognition tasks for three consecutive Kuzushiji characters.

	LER	0.07-Y-CR	0.05-Y-CR
Histogram+RESNET-53[9]	36.6%	63.46%	55.96%
YOLOv3	6.65%	75.8%	66.9%
FGDM-a	5.24%	81%	71.3%
FGDM-b	2.5%	83.18%	73.3%
FDMBSS	13.5%	86.5%	68.6%

Here, FGDM-a is denoted as the result of FGDM with the same learning rate of YOLOv3 and FGDM-b is the one with decreasing learning rate by multiplying 0.1 in every 40000 rounds.

As shown in Table 1. The proposed methods (FGDM-a,b) achieve low LER, but the prediction accuracy of bounding boxes is still insufficient. On the other hand, our second method, FDMBSS achieves best recognition results at 0.07-Y-CR for bounding boxes, while its LER is inferior to the first method FGDM. For the more detailed evaluation criteria, 0.05-Y-CR, the FGDM method is still more promising, but we still hope that the FDMBSS method can be an useful two-stage method for character recognition.

Also, we note that improvement of FGDM over YOLOv3 is due to the choice of bounding box aggregation method and the results show the effectiveness of our methods.

Finally, we compare our results with previouly reported result of Nguyen et al. [11]. Their method achieved LER 12.88%. Although the data sets and trainning and test splits are slightly different, our method seems to achieve better accuracy. It makes sense that use of the segmentation information makes recognition more accurate.

5 ANALYSIS OF THE RESULT ABOUT DIFFERENT CATEGORY OF CHARACTERS

There is an evaluation method called mAP[5] which is often used in the object detection framework. In this method, the map value of the YOLOv3 plus proposed method is 68%. In other words, not only the error rate of recognition but also the prediction of frame coordinates have reached high perfection.

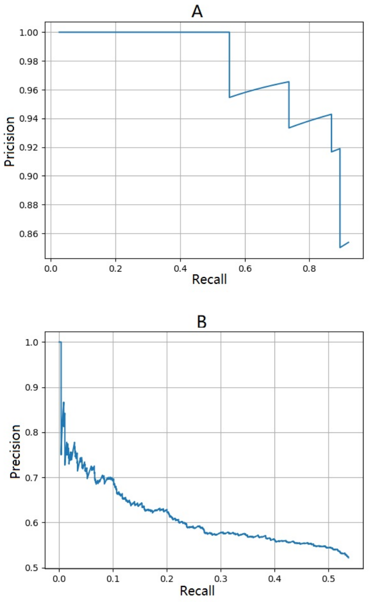

Figure 6: The precision rate and recall rate of Kuzushiji character

As we can see at Figure 6, the character "ho" of Figure 6-A has the best recognition result which have a high precision rate, although the line does not show a smooth curve due to the uneven sample. It would be because the characters are quite different from others. On the other hand, we observe that the recognition of character "shi" of Figure 6-B is hard to recognize which have an unsatisfactory precision rate. Since the shape of the character is like a straight line, it can be misclassified as a part of another character.

Note that pre-processing such as black-and-white binarization of character images will effectively improve the recognition rate and the applicability of the network, and random scrambling of data is also an important step.

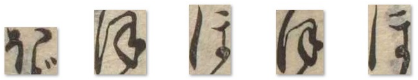

Figure 7: A:The character of Kuzushiji-"ho".

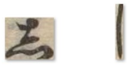

Figure 8: B:The character of Kuzushiji-"shi".

6 CONCLUSION

In this paper, we propose segmentation and recognition methods for Kuzushiji recognition, and achieve the state-of-the-art results on datasets of three consecutive Kuzushiji characters. It based on the object detection named YOLOv3 [13], and use of Kuzushiji segmentation character date set[17]. For future work, extension of our methods to recognition of multiple characters (more than 3) is quite important.

ACKNOWLEDGEMENT

We thank the anonymous reviewers for helpful comments. This work was supported by JSPS KAKENHI Grant Numbers JP18K18508, JP19H04174 and JP19H04067, respectively.

REFERENCES

[1] Adnan Amin. 1998. Off-line Arabic character recognition: the state of the art. *Pattern recognition* 31, 5 (1998), 517–530.

[2] David Arthur and Sergei Vassilvitskii. 2007. K-means++: The Advantages of Careful Seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '07)*. 1027–1035.

[3] Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. 2018. Deep Learning for Classical Japanese Literature. arXiv:1812.01718 http://arxiv.org/abs/1812.01718

[4] Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. 2018. Deep Learning for Classical Japanese Literature. arXiv:cs.CV/cs.CV/1812.01718

[5] Jesse Davis and Mark Goadrich. 2006. The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning*. ACM, 233–240.

[6] Ross Girshick. 2015. Fast R-CNN. *arXiv preprint arXiv:1504.08083* (2015). arXiv:cs.CV/http://arxiv.org/abs/1504.08083v2

[7] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2013. Rich feature hierarchies for accurate object detection and semantic segmentation. (2013). arXiv:cs.CV/http://arxiv.org/abs/1311.2524v5

[8] Klaus Greff, Rupesh Kumar Srivastava, Jan Koutnk, Bas R. Steunebrink, and Jrgen Schmidhuber. 2015. LSTM: A Search Space Odyssey. *IEEE Transactions on Neural Networks and Learning Systems* (2015). https://doi.org/10.1109/TNNLS.2016.2582924 arXiv:cs.NE/http://arxiv.org/abs/1503.04069v2

[9] He Kaiming, Zhang Xiangyu, Ren Shaoqing, and Sun Jian. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[10] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. 2016. Ssd: Single shot multibox detector. In *European conference on computer vision*. Springer, 21–37.

[11] Hung Tuan Nguyen, Nam Tuan Ly, Kha Cong Nguyen, Cuong Tuan Nguyen, and Masaki Nakagawa. 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.

[12] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 779–788.

[13] Joseph Redmon and Ali Farhadi. 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767* (2018).

[14] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*. 91–99.

[15] Christian Szegedy, Alexander Toshev, and Dumitru Erhan. 2013. Deep neural networks for object detection. In *Advances in neural information processing systems*. 2553–2561.

[16] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. 2013. Selective search for object recognition. *International journal of computer vision* 104, 2 (2013), 154–171.

[17] Tang Yiping, Kohei Hatano, Emi Ishita, Tetsuya Nakatoh, and Toshifumi Kawahira. 2018. Construction of Japanese Historical Hand-Written Characters Segmentation Data from the CODH Data Sets. In *Proceedings of the 8th Conference of Japanese Association for Digital Humanities (JADH'18)*. 183–185.