An Effective Approach for Noise Reduction from Shirakawa's Hand-Notated Documents on OBIs Research

Ziming Wang[1], Xuebin Yue[1], Lin Meng[1]

Dept.of Electronic and Computer Engineering, Ritsumeikan University,
Kusatsu, Shiga, Japan.
{gr0518eh@ed,gr0468xp@ed,menglin@fc}.ritsumei.ac.jp

Abstract

As one of the most influential professors in the field of Chinese language research in the last century, Professor Shirakawa had left many precious hand-notated research documents that have not been organized and made public. This paper aims to automatically organize hand-notated Oracle Bone Inscriptions research documents by image processing. In detail, Area-Mutation-Segmentation is applied for separating Oracle Bone Fragments from Oracle Bone Inscriptions document. Three-Stage-Denoise and Spider-Web-Extinction are proposed to reduce outside noise and border noise respectively. The experimental results show that the accuracy of Noise Reduction achieves 97.8%. The border noises of Oracle Bone Fragment images have been reduced perfectly without errors, and the number of valid Oracle Bone Inscriptions extracted has increased by 32.5%, comparing with previous research. In summary, the experimental results demonstrate the effectiveness of our proposed method.

1 Introduction

With the changes of the times and the rapid development of science and technology, people have never stopped searching for history while looking forward to future potential. Archaeologists and researchers tend to explore ancient culture by analyzing ancient characters carved on relics. The main component of relics is oracle bone. Although a large number of oracle bone inscriptions(OBI) have been unearthed, there are still many OBI features that have not been extracted. It is a serious problem that researchers have to face.

As discussed in [6], and [5], OBI is a form of character used in ancient China that was buried in ruins for about 3000 years and remained undiscovered until about 120 years age [12]. At present, more than 150,000 pieces of bone and turtle fragments have been excavated throughout China [2]. However, due to natural weathering, water erosion, and some carving habits, a large number of OBI characters connected with the border of Oracle Bone Fragments (OBFs) were not extracted in the previous paper. Professor Shirakawa was a prominent researcher on Chinese culture in the second half of the 20th century. He had left behind a large number of hand-notated Oracle Bone Inscriptions research documents that provide us with a very rich OBIs images dataset. During the organization of these documents, the same cases of OBI features related to the border of OBFs have been found. Attempting to extract these OBI features, we treat the border of OBFs as noises and reduced them by image processing based on algorithms. They include segmentation of OBFs by Area-Mutation-Segmentation, Noise Reduction by Three-Stage-Denoise and Spider-Web-Extinction.

In terms of segmentation, OBFs are segmented from OBIs images for Noise Reduction. As shown in Fig.1, the input image is an OBIs image from Shirakawa's documents, and the enlarged image cut from the OBIs image is one of the OBF images. In Noise Reduction, some noises such as OBF's number and features of adjacent OBF's border in the enlarged image

© 2022 Copyright for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3131/paper4.pdf

are treated as outside noise. Three-Stage-Denoise is proposed to reduce the outside noise. Spider-Web-Extinction is raised to reduce the border noise that refers to the border of each OBF.

This method not only provides us a new Noise Reduction method but also generates a relatively sufficient OBIs dataset for enhancing the robustness of OBIs recognition. Therefore, organizing the hand-notated OBIs research documents by Professor Shirakawa is of great significance for the research and protection of Chinese culture.

In section 2, some methods of Noise Reduction of images, recognition and extraction of OBIs are introduced. In section 3, the processing of the segmentation of OBFs, Noise Reduction of outside noise and border noise are depicted in detail. We report and discuss the evaluation results in section 4, and Section 5 proposes a conclusion and mentions the future work.

2 Related work and OBI analysis

Thanks to their importance in potentially unlocking the history of ancient China and helping with the evaluation of early Chinese characters, OBIs have recently been attracting more research attention. Anyang Normal University has created an OBI database that stores photographs and rubbings of OBI [10]. Many researchers have extracted features for recognition from OBIs images by various methods. For example, [6] proposed recognition of the OBIs line features using Hough Transform, [4] provided an Oracle image segmentation method by using fully convolutional networks. In addition, other researchers utilize various methods for Noise Reduction, such as that [11] had presented the results of applying different noise types to an image model and investigates the results of applying various Noise Reduction techniques. Especially, for solving the problem of a few dataset, [8] used SSD for detection and recognition of OBIs, and [7] proposed a new Deep Learning method for OBIs recognition by data augmentation.

3 Organization flow

Figure 1 shows the organizational flow of the proposed methods. The first step is OBFs segmentation, which aims to segment OBFs from original OBIs images. In the second step, Three-Stage-Denoise is employed to reduce the tiny outside noise. Ultimately, ensuring that OBI features that are connected with the OBF's border can be extracted completely, we have to separate the features of the border noise to facilitate the subsequent extraction of OBI character features. Hence, the third step is to propose Spider-Web-Extinction for Noise Reduction of border noise.

3.1 Segmentation of OBFs by Area-Mutation-Segmentation

During the preprocessing of segmentation, Fig.2 shows intermediate results of segmentation processing. Figure 2(a) shows the original image, and as Fig.2(b) shows, the original image is converted into grayscale for reducing the amount of calculation. In binarization, the OTSU method [9] is applied for obtaining a binarized image that is shown in Fig.2(c).

In OBFs segmentation, Area-Mutation-Segmentation is proposed to identify and locate OBFs on OBIs images. In detail, there is a large gap between the sizes of a single OBI character feature and a single OBF feature. In this algorithm, we get the threshold by catching the mutation point where the gap is located to decide whether it is OBF or not, and take a suitable

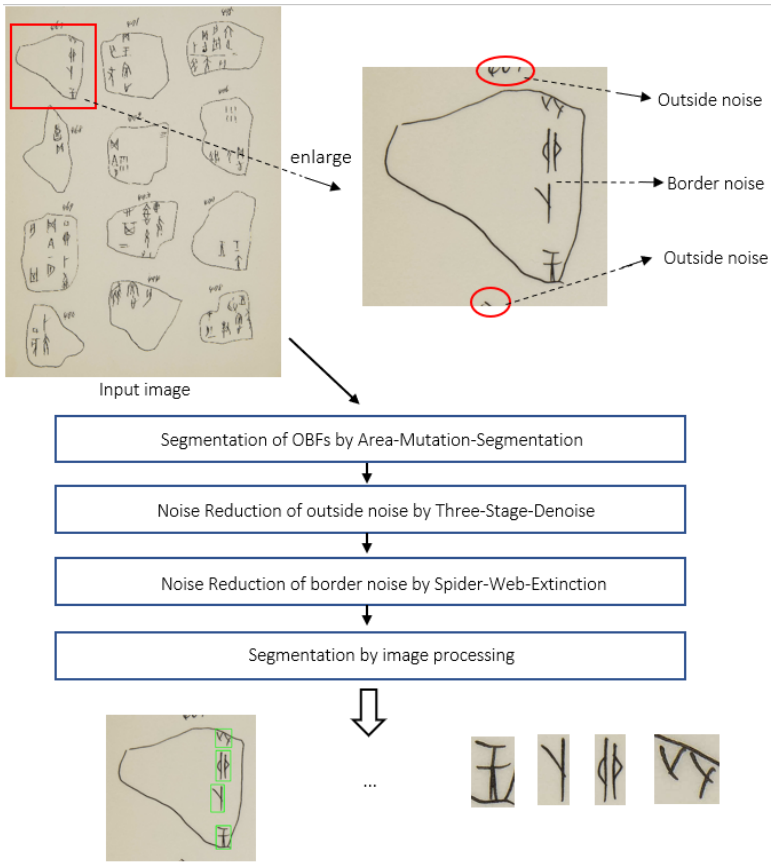
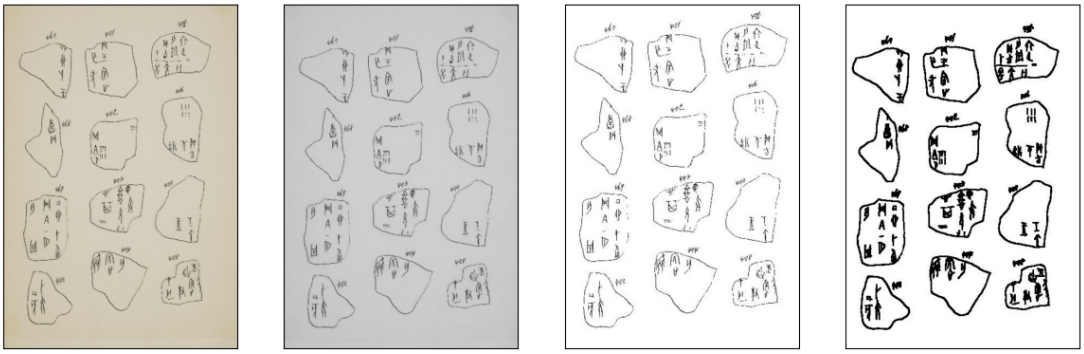


Figure 1: Overview of the proposed method

range to filter out OBF features. The processing is divided into organizing the size of bounding boxes, catching mutation points, and filtering OBFs.

1. ***Organizing size of bounding boxes***. Perform feature detection on OBIs images by cv2.Findcontours in OpenCV and its return value can be used to draw a bounding box. The sizes of the bounding boxes are sorted in order by quick sort and saved in the list. The sudden change between the sizes of a single OBI and a single OBF is detected by getting the max rate of change(gap). Avoiding being disturbed by the sudden changes in the size between the OBFs and the entire OBIs image, we delete the max data, which is the size of the OBIs image, from the list.

2. ***Catching mutation point***. Starting with the second element in the list, we calculated the difference between every element and the previous element, as shown in Equation 1. $S(i)$ means the size of the ith bounding box, and $D(i)$ refers to difference between $S(i)$ and $S(i+1)$. The mutation point of size can be caught by finding the max $D(i)$. Figure 3 shows a part of the size of bounding boxes ordered including mutation point. Assuming that $D(k)$ is the max rate of change, the larger size $S(k+1)$ of the two bounding boxes means the smallest single OBF's size, and the smaller size $S(k)$ refers to the size of the



(a) original image (b) grayscale image (c) binarized image (d) eroded image

Figure 2: Image during processing

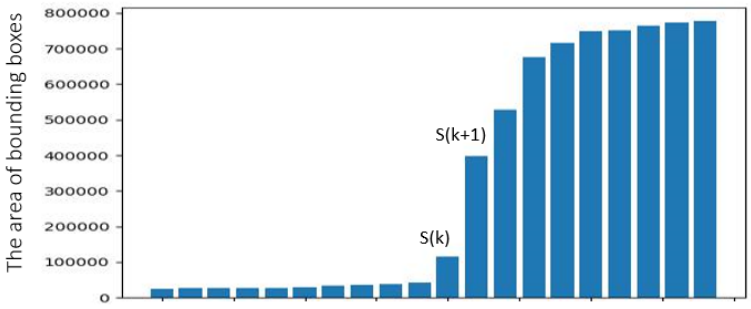


Figure 3: A part of list that show bounding box size ordered

biggest OBI character.

$$D(i) = S(i+1) - S(i)(i = 1, 2, ...) \tag{1}$$

3. ***Filtering OBFs***. All the bounding boxes with a size between D(k) and the OBIs image size border the OBF feature. Ensuring that every bounding box includes an OBF feature completely, we expand the range of 100 around the bounding boxes before cutting them out. Although the expanded bounding box may border outside noise by mistakes, these can be reduced clearly by the next processing.

3.2 Noise Reduction of outside noise by Three-Stage-Denoise

Three-Stage-Denoise is proposed to perform Noise Reduction. We take Fig.4(a) as an example to show the specific processing of Noise Reduction. Originally, the image is converted into grayscale, as shown in Fig.4(b). Meanwhile, in the median filtering [1] method, after sorting the surrounding pixels and center pixels, the median value is taken. The median filter can remove not only isolated clutters but also slightly dense clusters, and the result is shown in Fig.4(c). Further, utilize the OTSU method to obtain the binarized image which is shown in

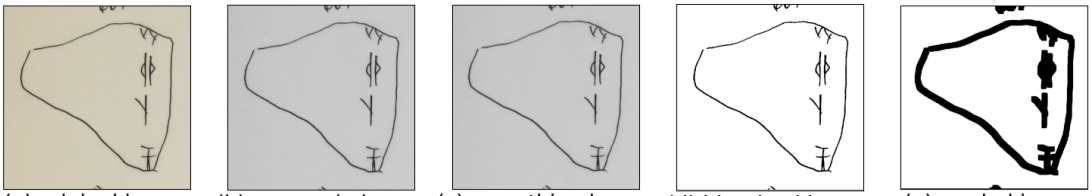


Figure 4: OBF image during processing

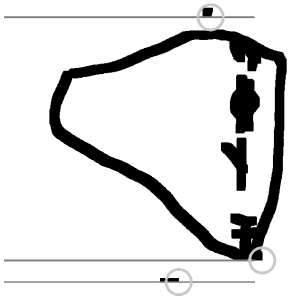


Figure 5: Determine the position where noise may appear

Fig.4(d). Eventually, we apply morphological processing [3] to ensure that the border of the OBF becomes a closed figure like Fig.4(e).

Before Noise Reduction, cv2.copyMakeborder in OpenCV is used to expand the white border around the picture for extracting the features that are close to the picture border. Three-Stage-Denoise is proposed to reduce noise by three stages, and processing on every stage includes eroding processing, organizing each OBF image by list, getting the point where noise may appear, and filtering noise out.

1. ***Eroding processing***. Some noise like the OBF number that is generally composed of three digits, multiple iterations eroding operation needs to be performed to make the three digital features connected. However, too many iterations will make the noise feature connected with the border of the OBF by mistake, so that the outside noise feature can not be identified as one object. In order to deal with such problems, at the first stage, multiple iterations are proposed to make some noise that is far away from the border connected as much as possible. In the second stage, we slightly reduce the iterations of the eroding operation so that the noise near the OBF border can be connected without contacting the OBF border. In the third stage, a light iteration of the eroding operation is utilized to target some noises that have not yet been reduced clearly during the first and second Noise Reduction.

2. ***Organizing each OBF image by list***. Scan each row of the OBF image from right to left and record how many pixels we have traversed until encounter the first black pixel. If there are no black pixels in a row, then record the distance we traverse at the row W_i, which is the width of the picture. In this way, from the first row to the last row, we save H_i (H_i refers to the height of the picture) data in the list. The same processing should be carried out again from right to left for organizing the image by another list.

3. ***Getting the point where noise may be appear***. After morphological processing, the black pixels of the OBF's border should be continuous relatively. Hence, the rate of change between two adjacent data in the two lists should be smooth besides some corner points and noise. We get this point to capture where outside noise may appear, and try to set a threshold as a criterion to decide whether the rate of change in the list is smooth or not. If the rate of change exceeds this threshold, the positions are marked as places where the noises may appear by a circle, as shown in Fig.5.

4. ***Filtering noise out***. Drawing bounding boxes and judging whether the positions where noise may appear are inside a small bounding or not. The gap between the sizes of OBF and the outside noise is large, and the threshold can be set easily. Further, we treat the marked positions bordered by a small bounding box as outside noise and reduce it by turning it into white.

3.3 Noise Reduction of OBFs's border by Spider-Web-Extinction

Spider-Web-Extinction is used to reduce border noise. During this algorithm, the OBF image was scanned from top to bottom, bottom to top, left to right, and right to left respectively. The case of left to right is chosen as an example for a detailed explanation. Firstly, start scanning from left to right every Tth row (T determines how big the border noise can be decomposed into pieces). Secondly, stop scanning when black pixels are encountered. Thirdly, turn the N pixels into white (N determines the degree of decomposition). Because all of the outside noise has been reduced in the second step, the black pixel encountered by scanning is judged as a part of border noise. The conversion of border pixels to white is a decomposition of the border noise.

To ensure that the border noise can be completely decomposed, the maximum value of the border width is set as N during Spider-Web-Extinction. Although this will affect part of the OBI features, the final erosion process can retain most of the character features. When the four aspects of processing are over, we can find that the border of the OBF is broken down into small squares and strips, which look like a spider web. The border noise is decomposed into many tiny features smaller than $T * N$. By turning all the features with a size below $T * N$ into white, the Noise Reduction is completed. It is noteworthy that avoid reducing the small OBI features inside the border due to mistakes, we try not to choose too big T.

4 Evaluation

In the experiment, we choose 79 OBIs images from Shirakawa's hand-notated document and process these OBF images for the organization in CPU(i7). OBF segmentation, Noise Reduction of outside noise and border noise are the three key components of the organization. In this section, we show and evaluate the experimental results of these three parts.

4.1 Evaluation result

In segmentation, we chose to use the Area-Mutation-Segmentation to filter the threshold to help us cut each OBF out. The result is that we cut 364 OBF images from 79 OBIs images from Shirakawa's hand-notated documents, such as Fig.6.

During Three-Stage-Denoise, outside noise has been reduced perfectly. As shown in Fig.7, it is the control group where we processed the two OBF images respectively. Fig.7(a) and

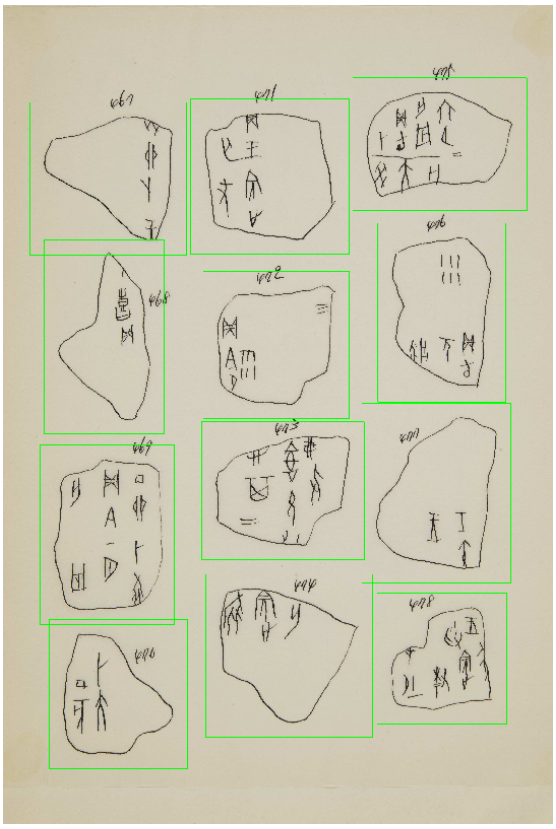


Figure 6: Segmentation of OBFs

Fig.7(d) are the original image of OBF a and OBF b respectively. And so on, Fig.7(b) and Fig.7(e) are the a and b that after the first stage of Noise Reduction. Similarly, Fig.7(c) and Fig.7(f) are the results of the second Noise Reduction. The outside noise in b is far away from the OBF's border, so it was reduced in the first stage. Relatively, the position of noise is close to the border of the OBF so that outside noise can be connected with the border after the first stage Noise Reduction. As a result, the noise has to be reduced in the second stage. The third time result is not shown here, because all the outside noise had been cleared mostly. The third stage is utilized to target some noises that have not been reduced clearly during the first and second stages. We performed Noise Reduction on the 364 OBF images totally, and on 356 pieces of them, outside noises have been reduced perfectly, and the accuracy of Noise Reduction is 97.8%.

In Noise Reduction of border noise, Spider-Web-Extinction has been used to decompose the border into strips and block features like spider webs, as shown in Fig.8(a). Fig.8(b) is a partially enlarged image of Fig.8(a). These features are reduced by filtering the size of the bounding box, and the result is shown in Fig.8(c). In a summary, the border noise of 356 OBF images has been reduced perfectly without errors.

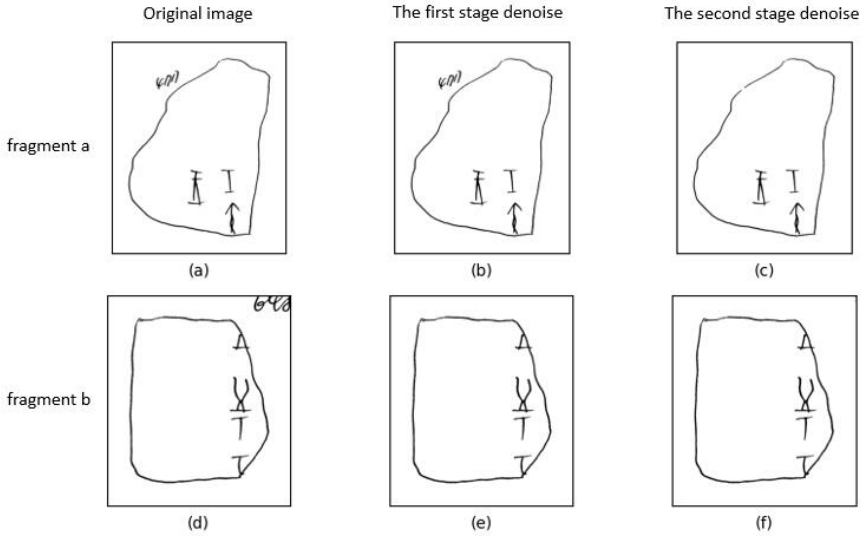


Figure 7: Process of Three-Stage-Denoise

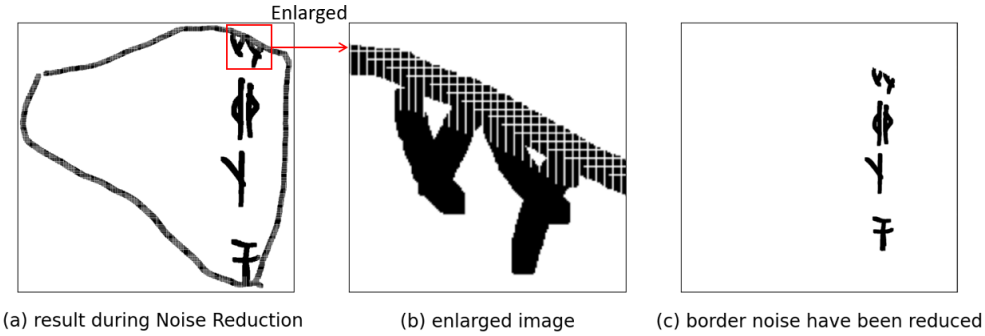


Figure 8: Image during Noise Reduction of border noise

4.2 Discussion

In terms of segmentation, the OBF features differentiated from OBI features by catching a large rate of difference between two bounding boxes' size at the first time. It is intuitive to help us distinguish OBF features accurately. However, problems still exist in using Area-Mutation-Segmentation to cut OBFs out. If the rate of change between the size of OBFs is particularly large, the position we locate by Area-Mutation-Segmentation may not be the dividing line between the OBIs and OBFs, but a dividing line between large OBFs and small OBFs. Fortunately, Professor Shirakawa sorted out and put OBFs with similar sizes in the same OBIs image as much as possible. However, facing these special cases, the algorithm should be optimized in the future.

During Noise Reduction of outside noises, we masterly utilize the characteristic that the

border of OBFs and the outside noise are almost not smoothly connected together to reduce the noises by three stages. About the reduction of border noise, we propose a creative method to effectively eliminate the OBF border, and it can almost deal with the reduction of any border noise during the organization perfectly.

5 Conclusion

In light of the recent discovery of Professor Shirakawa's hand-notated OBIs research documents, this paper aims to organize hand-notated Oracle Bone Inscriptions research documents by image processing. It is helpful for us to extract more OBI features from the limited OBIs images [5], [8]. In detail, Area-Mutation-Segmentation, Three-Stage-Denoise, and Spider-Web-Extinction are applied to perform segmentation and Noise Reduction respectively. Specifically, we found that the accuracy of Noise Reduction on outside noise was 97.8%, and the border noise was reduced perfectly without errors. During an organization, the number of OBI features extracted from OBIs image was increased by 32.5%. In the future, these methods based on image processing should be optimized for a higher accuracy of Noise Reduction. Additionally, after these Noise Reductions, we should be committed to segmenting every OBI feature clearly by some ingenious method. In this way, we can use Professor Shirakawa's document more efficiently and try to make contributions to the research of OBIs recognition.

References

[1] Ginu George, Rinoy Mathew Oommen, Shani Shelly, Stephie Sara Philipose, and Ann Mary Varghese. A survey on various median filtering techniques for removal of impulse noise from digital image. In *2018 Conference on Emerging Devices and Smart Systems (ICEDSS)*, pages 235–238. IEEE, 2018.

[2] Shuangping Huang, Haobin Wang, Yongge Liu, Xiaosong Shi, and Lianwen Jin. Obc306: A large-scale oracle bone character recognition dataset. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 681–688. IEEE, 2019.

[3] Nursuriati Jamil, Tengku Mohd Tengku Sembok, and Zainab Abu Bakar. Noise removal and enhancement of binary images using morphological operations. In *2008 International Symposium on Information Technology*, volume 4, pages 1–6. IEEE, 2008.

[4] Guoying Liu, Xu Song, Wenying Ge, Hongyu Zhou, and Jing Lv. Oracle-bone-inscription image segmentation based on simple fully convolutional networks. In *MIPPR 2019: Pattern Recognition and Computer Vision*, volume 11430.

[5] Guoying Liu, Jici Xing, and Jing Xiong. Spatial pyramid block for oracle bone inscription detection. In *Proceedings of the 2020 9th International Conference on Software and Computer Applications*, pages 133–140, 2020.

[6] Lin Meng. Two-stage recognition for oracle bone inscriptions. In *International Conference on Image Analysis and Processing*, pages 672–682. Springer, 2017.

[7] Lin Meng, Naoki Kamitoku, and Katsuhiro Yamazaki. Recognition of oracle bone inscriptions using deep learning based on data augmentation. In *2018 Metrology for Archaeology and Cultural Heritage (MetroArchaeo)*, pages 33–38. IEEE, 2018.

[8] Lin Meng, Bing Lyu, Zhiyu Zhang, CV Aravinda, Naoto Kamitoku, and Katsuhiro Yamazaki. Oracle bone inscription detector based on ssd. In *International Conference on Image Analysis and Processing*, pages 126–136. Springer, 2019.

[9] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66, 1979.

[10] Anyang Normal University. OBI database of anyang normal university, Last accessed 24 Feb. 2021. http://jgw.aynu.edu.cn/.

[11] Rohit Verma and Jahid Ali. A comparative study of various types of image noise and efficient noise removal techniques. *International Journal of advanced research in computer science and software engineering*, 3(10), 2013.

[12] Yikang Zhang, Heng Zhang, Yongge Liu, Qing Yang, and Chenglin Liu. Oracle character recognition by nearest neighbor classification with deep metric learning. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 309–314. IEEE, 2019.