Journal of Physics: Conference Series

PAPER • OPEN ACCESS

HWOBC-A Handwriting Oracle Bone Character Recognition Database

To cite this article: Bang Li *et al* 2020 *J. Phys.: Conf. Ser.* **1651** 012050

View the article online for updates and enhancements.

HWOBC- A Handwriting Oracle Bone Character Recognition Database

Bang Li[1,3,4]*, Qianwen Dai[1,2,3,4], Feng Gao[1,3,4], Weiye Zhu[1,3,4], Qiang Li[1,3,4] and Yongge Liu[1,3,4]

[1]School of Computer & Information Engineering, Anyang Normal University, Anyang Henan 455000, China

[2]School of Information Engineering, Zhengzhou University, Zhengzhou Henan 450000, China

[3]Key Laboratory of Oracle Information Processing in Henan Province, Anyang, Henan 455000, China

[4]Key Laboratory of Oracle Bone Inscriptions Information Processing, Ministry of Education of China, Anyang Henan 455000, China

*Corresponding author's e-mail: 01782@aynu.edu.cn

Abstract. The oracle bone character (OBC) from ancient China is the most famous ancient writing systems around the world. Identifying and deciphering OBCs is one of the most important topics in oracle bone study. In research, one of the challenges is that the literature review usually leads to a huge cost of time and manpower. Therefore, the digitazation of OBC literature through the automatic recognition is the inevitable trend of future development. However, the OBCs in the literature are usually writing characters while the database of handwriting OBC has not yet been presented. In this paper, we establish a handwriting oracle bone character database called HWOBC, containing 83,245 character-level samples which are grouped into 3881-character categories. We also present the performance of several baseline DCNN-based methods, in which Melnyk-Net exhibits the best accuracy of 97.64%. It is anticipated that the publication of this database will facilitate the development of OBC research.

1. Introduction

Oracle bone characters (OBCs) is identified as an early form of the written language used in the Chinese Bronze Age and is the earliest systematic character known in China [1]. Abundant information of the Shang dynasty of ancient China was recorded by OBC on animal bones or turtle shells. The study of OBC is of great significance to the research of Chinese etymologies and the culture and history of the Shang dynasty, ancient China. Thus far, more than 150,000 pieces of bone and turtle fragments have been excavated throughout China, and more than 4,000 different oracle bone characters (OBC) have been discovered on them [2]. Among which, only about 2,200 characters have been deciphered [3], indicating that there is still a long way to go in the study of OBCs.

One of the difficulties lies in the judgment of glyph. From the perspective of research object, it is generally known that the glyphs of OBC are various and complex, an OBC usually has several variants [4], causing a huge cost of time and manpower. From the perspective of research method, the research of OBC's glyphs requests a great deal of literature review, however, each paper book of OBCs is usually rich in content and hard to retrieve [2-5]. In this way, the literature review of OBC is called "semi-manual labour". The digitization of OBC literature is urgently required for the existing and future OBC research. In response, all the literature of OBC we can collected are well digitized and stored on our website "yin qi wen yuan" (OBC library) [6]. 150,302 Oracle bone fragments in different versions and 32,117 articles related with OBCs can be viewed on our website. However, for various reasons, it is difficult to input OBCs by computer directly, thus most of OBCs in the published literatures are printed by screenshot or photocopying. The recognition of these printed OBCs in literatures still remains a challenge.

Recently, the field of computer vision has significantly advanced thanks to the deep convolutional neural network (DCNN) methods [7-11]. The great achievements of DCNNs in the computer vision community inspired us to adopt them to the identification of OBCs. However, these DCNN-based methods rely on a large number of labeled training data, which are usually unavailable to obtain, especially in this specific field. Moreover, the recognition of OBCs in the literature has a precondition, which requires to converse the character pictures into character codes. In this paper, we collected a handwriting oracle bone character database named HWOBC, containing 83,245 character-level samples which are grouped into 3881 character categories. The sample collection is based on the OBC font published on our website [6], which is establishment by encoding the standard glyph of OBC. In addition, HWOBC will be published on our website either [6].

Through statistical and visual analyses, we reveal several difficulties of identifying OBC and analyze the potential future extensions or challenges. We also present some evaluation results on the database using several DCNN-based methods for benchmarking. We hope that this work could be a step to breaking down the technical barriers in the process of digitizing Oracle literature. In addition, this work has also been applied to the handwriting input of OBC. We hope that this technology could bridge the gap between information experts and archaeologists or paleontologists, meanwhile, facilitate the writing and publishing of OBC literature in the future.

2. Related work

2.1. Glyph of Oracle Bones Character

OBCs have been investigated since the last century and many related reference books have already been published, including collections of rubbings, dictionaries, literature integration and general explanations of oracle bones etc. [2-5, 12, 13]. However, the total amount of OBCs in different literatures is not consistent. As shown in Table I, Jiaguwenzi Bian(Compilation of OBC) contains 4378 OBCs but the quantity of OBCs reduced to 4078 in Jiaguwenzi Xinbian(New Compilation of OBC). The inconsistent count of OBC indicates that the glyphs of OBCs are still controversial, resulting in a difficulty in the establishment of standard character set (like GB2312) for the computer input of OBC. Therefore, the OBCs in published oracle bones literature are usually screenshots of oracle bone facsimile copies or photocopying with OBCs written on them. Furthermore, whether it is a handwritten OBC (printed by photocopying) or a copy of OBC (printed by screenshot), the recognition of oracle bone characters in the literature is carried by a special handwritten character recognition. Handwriting recognition has always been an active and challenging research area. The impendent of handwriting OBC recognition hinders the progress toward automatically identifying and deciphering OBCs. Thus, we design a handwriting oracle bone character database to address this issue.

Table 1. The published date and the total number of OBCs in each literature.

Literature	OBC number	published date
Jiaguwen Wenzi Gulin(Collected commentaries on the oracle-bone characters)[12]	3556	1996
Xinbian Jiaguwen Zixingzongbiao(New General table of OBC glyphs)[3]	4071	2001

Xinbian Jiaguwen Zixingzongbiao (Update)	4024	2008
Xinbian Jiaguwen Zixingzongbiao (Update)	4159	2018
Jiaguwenzi Bian(Compilation of OBC)[13]	4378	2012
Jiaguwenzi Xinbian(New Compilation of OBC)[4]	4078	2017

2.2. Oracle-Bone Inscription Recognition

The research on the recognition of OBC focus on the recognition of rubbing image of inscriptions. The methods of early research works are mainly based on graph theory and topology, such as graph isomorphism, topological features and even a Fourier descriptor based on curvature histograms [14-16]. Recently, the research of the recognition of oracle-bone inscriptions (OBIs) based on machine learning are also published, Guo J. et al. regarded this research as a sketch recognition task and constructed a hierarchical representation [17]. Liu J. et al. adopted support vector machines and block histogram-based features to recognise OBI [18], Zhang H. proposed the OBI recognition by nearest neighbour classification with deep metric learning [19]. The database of these researches is collected from the publications of the oracle bone rubbings, aiming to assist the deciphering of OBI [20]. However, there are obvious distinctions between the OBIs and the handwritten OBCs. The OBI image database is not suitable for the recognition of handwriting OBC.

2.3. Handwritten Chinese Characters Recognition (HCCR)

Recently, DCNNs have made great achievement in computer vision tasks such as large-scale image classification. After the application of DCNNs in HCCR problem, the progress in the past 5 years has greatly surpassed traditional approaches. The first model that outperformed human-level performance was introduced by Zhong et al. (2015) [7], their single HCCR-Gabor-GoogLeNet and ensemble HCCR-Ensemble-GoogLeNet-10 models achieved a recognition accuracy of 96.35% and 96.74%, respectively. Zhang et al. (2017) used the traditional gradient maps as network input and obtained an accuracy of 96.95%. With an additional adaptation layer, the recognition accuracy was further improved to 97.37%, which set new benchmarks for offline HCCR [8]. Xiao et al. (2017) employed the low-rank expansion and pruning technique to solve the problems of speed and storage capacity [9]. Another well-balanced network in terms of the speed, size, and performance was recently introduced by Li et al (2018) [10]. By utilizing fire modules and the proposed global weighted average pooling (GWAP) concept along with quantization, the model achieved an accuracy of 97.11% requiring only 3.3 MB for storage. Melnyk et al. (2019) modified GWAP to global weighted output average pooling (GWOAP) and improved the accuracy to 97.61% [11]. As the source of Chinese characters, OBC and modern handwritten fonts have several similar properties (e.g., strokes), using DCNN methods to identify OBC is a feasible approach. We hence evaluate several DCNN methods to provide a baseline for the constructed database.

3. Construction of the Database

HWOBC was built for the purpose of providing a refined offline HWOBC database for researchers in the field of offline handwritten OBC recognition, aiming at the digitization of OBC literatures and the handwriting input of standard OBC characters during the editing of related literatures. Compared with the handwritten Chinese character database, the proposed HWOBC database has the following special characteristics. Frist, the characters in the historical literatures are handwritten by tracing the character in the rubbings. Unlike handwritten Chinese characters, OBC is not the native language of the writer, which means that its font is not strongly influenced by personal writing style of writers. Therefore, the samples in the database should possess a high intra class similarity. Moreover, the purpose of font tracing in the historical literatures is to restore the original meaning of oracle bone characters as much as possible, requiring the writers of sample collection to be knowledgeable about the meaning of OBC. Second, there is always a specific strategy in sampling objects of a handwriting database. For example, HWDB1.1 took the 3755 frequently used Chinese characters of level 1 in the GB2312-80 standard as sampling objects, while HIT-MW adopted various kinds of paragraphs in China Daily. However, the amount of OBCs has increasing gradually with the extension and expansion of OBIs, and there are also some disputes over the glyph of OBCs to resolve. Therefore, it is very necessary to establish an authoritative OBC glyph reference standard.

In the choice of sampling glyphs, we established an OBC font library as reference standard. With the efforts of authoritative OBI experts, the glyph of every character in the font library is confirmed by its inherent meaning. Besides, the font models in the font library are all written by the calligraphy expert. Then, these OBCs are saved as a special font named AYJGW encoding by a six character/number code and announced on our website [6]. AYJGW will be updated with the expansion research of OBCs. In AYJGW, we selected 3,881 characters with uncontroversial font shapes as targets, and 22 Oracle researchers of different specialties (including computer, history, calligraphy, archaeology, etc.) as writers. 83,245 OBC samples are written by PC or smart mobile phones and collected through the local area network. As illustrated in Fig. 2, the input area in our program is designed to be a box of static size (400*400 pixel), large enough to accommodate a complex character. In addition, considering that the writer maybe unfamiliar with a particular OBC, we place the “next” and “last” button to skip or rewrite this character.

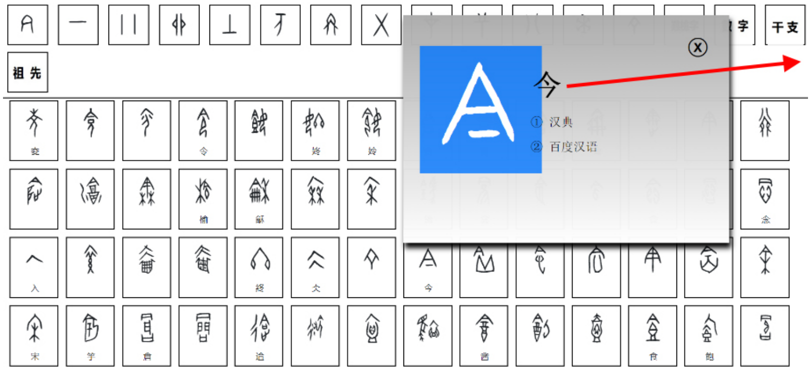

Figure 1. The encode OBC on our website

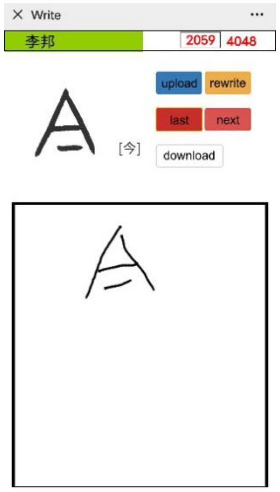

Figure 2. An illustration of layout of our data collection software

4. Baseline DCNN Method Evaluation and Future Challenges

We evaluated five popular DCNN models on our database: AlexNet [21], ResNet [22], VGG11 [23], Cascaded model [10] and Melnyk-Net [11]. To trade-off between accuracy and required storage, the traditional fully connected (FC) layers of these models has been modified by global weighted average pooling (GAP), global weighted average pooling (GWAP) and global weighted output average pooling (GWOAP), respectively [11]. We randomly selected 75% of the samples from each class for training and reserved the rest for testing. There is a series of pre-processing before training the data. First, the white background without characters in the picture is cut off. Then, the character image is normalized to the size of 96 × 96 pixel, and the pixel intensity is inversed. Considering the invariance of OBC flipping (horizontal or vertical), we induced data augmentation through horizontal and vertical flipping. We conducted the training process on Mxnet [24] using a GeForce RTX 2080 Ti GPU.

Table 2. The average accuracy of the different methods

Method	FC	GAP	GWAP	GWOAP
AlexNet	96.62%	94.47%	96.56%	96.44%
Vgg11	97.57%	90.36%	97.46%	97.26%
ResNet	97.44%	96.85%	96.96 %	97.08%
Cascaded model	97.57%	96.50%	96.84%	96.57%
Melnyk-Net	97.67%	97.36%	97.41%	97.11%

As shown in table 2, Melnyk-Net, which yields state-of-the-art accuracy for single-network methods trained only on handwritten Chinese Character data, performs the best accuracy (97.67%) of our database. Replacing the FC layer by GAP leads to a decline in accuracy while reduces the number of parameters. By optimizing GAP to GWAP, the accuracy has been improved, but there is still a gap between the accuracy with FC layer. The effect of GWOAP varies among different models, which is believed to be related to the size of the feature map in its previous layer. In summary, DCNNs demonstrate promising ability to recognize OBCs, which shows similar accuracy compare with handwritten Chinese character recognition.

In this paper, we emphasize handwritten OBC recognition of standard glyphs, however, our purpose is much more than that. Developing a recognition system that can handle all glyphs of OBC will be a significant challenge in future. OBCs with multiply glyphs usually have similar components but different font structures. It should be a zero-shot learning task for the reason that the characters can be classified by its structure and components. To our knowledge, the research of resolving disputed OBC classes has attracted no attention, yet we have already started the splitting process of the oracle bone structure. Our future research will focus on the zero-shot learning of disputed OBC recognition. If success, we can not only recognize the new variants of existing glyph classes and even recognize new glyph classes, and thus automatically expand our database until it contains the full glyphs of all distinct OBCs.

5. Conclusion

We presented a database of hand writing oracle bone Character to support the digitization of OBI literature and the advancement of oracle-bone script analysis. This database consists of 83,254 handwriting OBC images covering 3881 classes, which are written by 22 Oracle researchers of different specialties. The samples in the database have highly intra class similarity because OBCs are not the writers' native language. We also provide results for several state-of-the-art DCNN algorithms for HCCR as baseline results and modified the bottleneck layer of these models by weighted average pooling and global weighted output average pooling. We hope that our database will accelerate the digitization of Oracle literature and assist future studies on deciphering oracle-bone characters**.**

Acknowledgments

This work was supported by the Oracle Bone Inscriptions Research and Application Special Project of Ministry of Education and National Language Committee of China (Grant number: YWZ-J010, YWZ-J023), the National Natural Science Foundation of China (Grant number: 61806007), the National Social Science Fund Major Entrusted Project of China (Grant number: 16@ZH017A3), Program for Changjiang Scholars and Innovative Research Team in University (Grant number: 2017PT35), Key scientific and technological projects in Henan Province (Grant number: 202102310562), the Fund of Key Laboratory of Oracle Bone Inscriptions Information Processing, Ministry of Education of China (Grant number: OIP2019M006), and the Research Foundation of Anyang Normal University(Grant number: AYNUKPY-2019-04).

References

[1] Chinasage. (2016) Early Chinese writing - Shang oracle bones. http://www.chinasage.info/oraclebones.htm

[2] Peng, B., Xie, J., and Ma, J. (1999) Jiaguwen heji bubian (A Supplement to the Comprehensive Dictionary of Oracle Characters). Language Publishing House, Beijing, China,.

[3] Chen, N. (2017) Jiaguwenzi Xinbian(New Compilation of OBC). Wire-bound bookstore, Beijing,China.

[4] Shen, J. and Cao, J. (2001) Xinbian Jiaguwen Zixingzongbiao(New General table of OBC glyphs), Shanghai Lexicographical Publishing House, Shanghai, China.

[5] Guo, M. and Hu, H. (1978) Jiaguwen heji (The Comprehensive Dictionary of Oracle Characters). Zhonghua Book Company, Beijing,China.

[6] Key Laboratory of Oracle Bone Inscriptions Information Processing, Ministry of Education of China. (2019) http://jgw.aynu.edu.cn/

[7] Zhong, Z, Jin, L,Xie, Z (2015) High performance offline handwritten Chinese character recognition using googlenet and directional feature maps. In: 2015 13th International conference on literature analysis and recognition (ICDAR). IEEE, Tunisia, pp 846–850

[8] Zhang, X., Bengio, Y., Liu, C. (2017) Online and offline handwritten Chinese character recognition: a comprehensive study and new benchmark. Pattern Recognit. 61:348–360

[9] Xiao, X., Jin, L., Yang, Y., Yang, W., Sun, J., Chang, T. (2017) Building fast and compact convolutional neural networks for offline handwritten Chinese character recognition. Pattern Recognit. 72:72–81

[10] Li, Z, Teng, N, Jin, M, et al. (2018) Building efficient CNN architecture for offline handwritten Chinese character recognition. International Journal on Literature Analysis and Recognition, 21(4): 233-240

[11] Melnyk, P., You, Z., and Li, K., (2019) A high-performance CNN method for offline handwritten Chinese character recognition and visualization. soft computing, 24: 7977-7987.

[12] Yu, S. (1996) Jiaguwen Wenzi Gulin(Collected commentaries on the oracle-bone characters) ,Zhonghua Book Company, Beijing,China.

[13] Li, Z. (2012) Jiaguwenzi Bian(Compilation of OBC) Zhonghua Book Company, Beijing,China.

[14] Li, Q., Yang, Y., and Wang, A. (2011) Recognition of inscriptions on bones or tortoise shells based on graph isomorphism," Jisuanji Gongcheng yu Yingyong(Computer Engineering and Applications), 47(8): 112–114.

[15] Gu, S. (2016) Identification of oracle-bone script fonts based on topological registration, Computer & Digital Engineering, 10:029.

[16] Lv, X., Li, M., Cai, K., et al. (2010) a graphicbased method for chinese oracle-bone classification," Journal of Beijing Information Science and Technology University, 25:92–96.

[17] Guo, J., Wang, C., Roman-Rangel E.,et al. (2016) Building hierarchical representations for oracle character and sketch recognition," IEEE Transactions on Image Processing, 25(1) :104–118.

[18] Liu, Y. and Liu, G. (2017) Oracle bone inscription recognition based on svm, Journal of Anyang Normal University, 2:54-56.

[19] Zhang, Y., Zhang, H., Liu, Y., et al. (2019) Oracle Character Recognition by Nearest Neighbor Classification with Deep Metric Learning. In: international conference on literature analysis and recognition, Sydney, Australia, pp. 309-314.

[20] Huang, S., Wang, H., Liu, Y., et al. (2019) OBC306: A Large-Scale Oracle Bone Character Recognition Database. In: international conference on literature analysis and recognition Sydney, Australia, pp. 681-688.

[21] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012) Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, 2012, pp. 1097–1105.

[22] Simonyan K. and Zisserman A. (2014) Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556.

[23] He K., Zhang X., Ren S., and Sun J. (2016) Deep residual learning for image recognition," in: computer vision and pattern recognition, Las vegas, pp. 770–778.

[24] Chen T.，Li M.，Li Y. (2015) MXNet: A Flexible and Efficient Machine Learning Library for Heterogeneous Distributed Systems, arXiv preprint arXiv: 1512.01274v1.