



OPEN

DATA DESCRIPTOR

An open dataset for oracle bone character recognition and decipherment

Pengjie Wang^{1,5}, Kaile Zhang^{1,5}, Xinyu Wang², Shengwei Han³, Yongge Liu³, Jinpeng Wan¹, Haisu Guan¹, Zhebin Kuang¹, Lianwen Jin⁴, Xiang Bai¹ & Yuliang Liu¹✉

Oracle bone script, one of the earliest known forms of ancient Chinese writing, presents invaluable research materials for scholars studying the humanities and geography of the Shang Dynasty, dating back 3,000 years. The immense historical and cultural significance of these writings cannot be overstated. However, the passage of time has obscured much of their meaning, presenting a significant challenge in deciphering these ancient texts. With the advent of Artificial Intelligence (AI), employing AI to assist in deciphering Oracle Bone Characters (OBCs) has become a feasible option. Yet, progress in this area has been hindered by a lack of high-quality datasets. To address this issue, this paper details the creation of the HUST-OBC dataset. This dataset encompasses 77,064 images of 1,588 individual deciphered characters and 62,989 images of 9,411 undeciphered characters, with a total of 140,053 images, compiled from diverse sources. The hope is that this dataset could inspire and assist future research in deciphering those unknown OBCs. All the codes and datasets are available at <https://github.com/Pengjie-W/HUST-OBC>.

Background & Summary

Oracle bone script, etched onto turtle shells and animal bones, stands as one of the earliest forms of writing discovered in China (see Fig. 1). These inscriptions, dating back about 3,000 years, offer a window into the human geography of the Shang Dynasty (1600 BCE - 1046 BCE), an ancient feudal dynasty ruled in the Yellow River valley. The content encompasses a range of topics including astrology, meteorology, animal husbandry, religion, and ritual practices^{1,2}. Similar to other ancient scripts, the meanings of many Oracle Bone Characters (OBCs) have been lost over time. Of the 160,000 pieces unearthed, they reveal more than 4,600 distinct OBCs, yet only about a thousand of these have been deciphered, with their meanings and corresponding modern Chinese characters confirmed³.

The target of deciphering these ancient inscriptions is to translate OBCs into modern Chinese, with characters corresponding one-to-one with the same meaning. However, the deciphering task at the character level is complicated by several factors. Historically, the methods of preservation and excavation were not always ideal, leading to many of the oracle bones being damaged. This damage often results in partial, unclear, or illegible inscriptions, making interpreting them arduous. Therefore, most of the images used in current Oracle Bone Character (OBC) research are scanned images that have undergone denoising and other processing or artificially transcribed images. In addition, the nature of oracle bone script as an early writing system means that it underwent significant evolution. There is a considerable variation in the form of characters, with many characters appearing in multiple, sometimes radically different forms⁴ but corresponding to the same Chinese character. This variability adds another layer of complexity to the deciphering process. All these factors contribute to making the full understanding of OBC not only challenging but also a rare feat, attracting the keen interest of scholars and historians alike in the field of ancient Chinese studies.

The past decades have witnessed the widespread applications of Artificial Intelligence (AI) in various fields. Notably, the immense success of handwritten text recognition (HTR)⁵⁻⁷ technology in processing modern texts has sparked interest in the potential use of AI to aid in deciphering OBCs. Modern AI algorithms, particularly those centered around deep learning models like artificial neural networks, typically require an extensive volume

¹Huazhong University of Science and Technology, Wuhan, 430074, China. ²The University of Adelaide, SA, Adelaide, 5005, Australia. ³Anyang Normal University, Anyang, 455000, China. ⁴South China University of Technology, Guangzhou, 510641, China. ⁵These authors contributed equally: Pengjie Wang, Kaile Zhang. ✉e-mail: yliu@hust.edu.cn



Fig. 1 3000-year-old Shang Dynasty oracle bones. They were all unearthed at the Huayuanzhuang East in Yinxu, China, and are currently housed at the Anyang workstation of the Institute of Archaeology, Chinese Academy of Social Sciences. These oracle bones date back to the reign of King Wu Ding of the Shang Dynasty.

Dataset	Image Source	Deciphered		Undeciphered	
		#Characters	#Images	#Characters	#Images
HWOBC	Book	1,566	33,555	2,315	49,690
OBI-100	Book	100	4,748	0	0
OBI-125	Book	125	4,257	0	0
Oracle-20k	Web	261	20,039	0	0
HUST-OBC	Book, Web, Database	1,588	77,064	9,411	62,989

Table 1. Comparison of the scale of HUST-OBC with other oracle bone character datasets.

Modern Character Category	Oracle Bone Character Images			Variant Characters Category	Oracle Bone Character Images		
暴				𧇧			
並				竝			
叉				叉			
春				𡗗			
登				𡗗			

Table 2. Comparison of oracle bone character annotations using Modern and Variant characters as category labels.

of data for training. This approach enables them like AlphaGo who defeated the world champion in a Go game⁸ to achieve, and sometimes surpass, human-level performance in specific tasks. A fundamental step towards employing these models to decipher OBCs involves the creation and annotation of a comprehensive high-quality dataset of OBCs. In the dataset, each OBC is labeled by its modern counterpart, while different OBCs with the same label are referred to as a category in the dataset. There have been some pioneering efforts in this area. For instance, Li *et al.*⁹ built the HWOBC dataset by engaging experts from diverse academic backgrounds to hand-write OBCs. Fu *et al.*¹⁰ and Yue *et al.*¹¹ subsequently proposed the OBI-100 and OBI-125 datasets, with the OBC images collected from books related to OBC research. Additionally, Guo *et al.*¹² collected more than 20k OBC images from various websites to build the Oracle-20k dataset. These efforts lay a solid foundation for digitalization and recognition research of OBC. In addition, Li *et al.*¹³ created the Oracle Bone Inscriptions Multi-modal Dataset (OBIMD), but this dataset focuses on entire rubbings and lacks rich data on individual OBCs. However, the fact that these datasets have certain limitations also hinders their use in AI-assisted OBC decipherment:

- They often have limited categories and samples of OBCs due to data collection from a single source.
- The annotation of the categories might not be deduplicated. As shown in Table 2, the same OBCs are categorized into different classes.
- The lack of cross-validation from multiple sources casts doubt on the accuracy of some data.

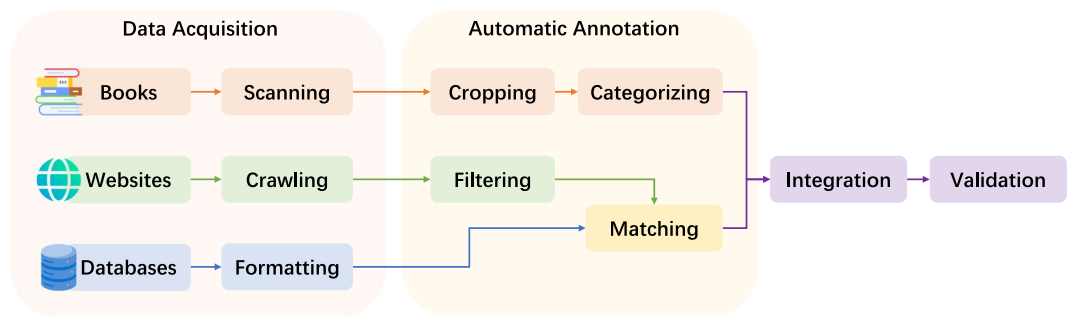


Fig. 2 Flowchart of building the HUST-OBC dataset.

Source	Type	#Categories		#Samples	
		Deciphered	Undeciphered	Deciphered	Undeciphered
New Compilation of Oracle Bone Scripts	Book	1,203	2,288	17,609	6,494
Oracle Bone Script: Six Digit Numerical Code	Book	1,148	4,444	9,609	4,444
YinQiWenYuan	Website	1,164	2,363	1,697	2,363
GuoXueDaShi	Website	1,366	0	16,259	0
HWOBC	Database	1,122	2,315	31,890	49,688
Total		1,588	9,411	77,064	62,989

Table 3. Statistic of oracle bone character data collected from various sources.

- The datasets comprise only deciphered oracle bone images, making them unsuitable for deciphering tasks.
- Some datasets contain unprocessed images, filled with noise or blur.

To address these issues, we propose the high-quality HUST-OBC¹⁴ dataset. The HUST-OBC dataset was collected from three different sources, including books, websites, and existing databases. HUST-OBC includes two types of OBC sample images: a) OBC images obtained from processed scans of rubbings of the original oracle bones; and b) handwritten OBC images based on the original oracle bones, further subdivided into traced images based on rubbings and manually drawn images based on the glyphs. To build HUST-OBC, we designed a semi-automatic pipeline that collects and annotates data from various sources and had OBC experts review the dataset. As shown in Table 1, HUST-OBC contains over 10k deciphered and undeciphered OBC categories and more than 140k images, making it one of the largest datasets for OBCs recognition and deciphering to date. We hope HUST-OBC will aid and inspire future AI-assisted OBC research.

Methods

To construct a diverse dataset, we gathered images of OBCs from three distinct sources: *book*, *website*, and *database*. To organize and merge data from these varied origins, as shown in Fig. 2, we designed a semi-automated pipeline comprising four key steps: *Data Acquisition*, *Automatic Annotation*, *Data Integration*, and *Data Validation*. In this section, we will delve into the details of each step.

Data acquisition. OBCs were inscribed on turtle shells and animal bones and buried underground for over 3,000 years. These precious artifacts are dispersed in museums and private collections worldwide, where they are meticulously preserved, making direct access to the text inscribed on the original oracle bones quite challenging. Thankfully, most of the publicly available oracle bones have been transcribed by experts, making them accessible in various forms for scholarly research. Specifically, for most authoritative books or websites, the images are processed or traced based on rubbings of oracle bones by experts. Building on this, HWOBC hired experts to manually draw each oracle bone character glyph, thus expanding the dataset with handwritten OBCs. As illustrated in Table 3, the HUST-OBC dataset is constructed by gathering data from these diverse sources.

To ensure the diversity of the dataset, the HUST-OBC was built using data collected from various sources, including books, websites, and databases. As shown in Fig. 2, we designed specific pipelines for each data source to process and extract OBC images and their corresponding labels, detailed as follows.

Books. Books remain the predominant form in documenting OBCs, with most discovered characters to date collected and interpreted in volumes like the *New Compilation of Oracle Bone Scripts*, which ensures accuracy by incorporating the latest research in this field. Specifically, we utilized the following books as data sources while constructing the HUST-OBC dataset.

- New Compilation of Oracle Bone Scripts** (新甲骨文编 <https://books.google.com/books?id=S0RergEA-CAAJ>)¹⁵ encompasses samples of OBCs found since its initial discovery, as presented in all public materials.

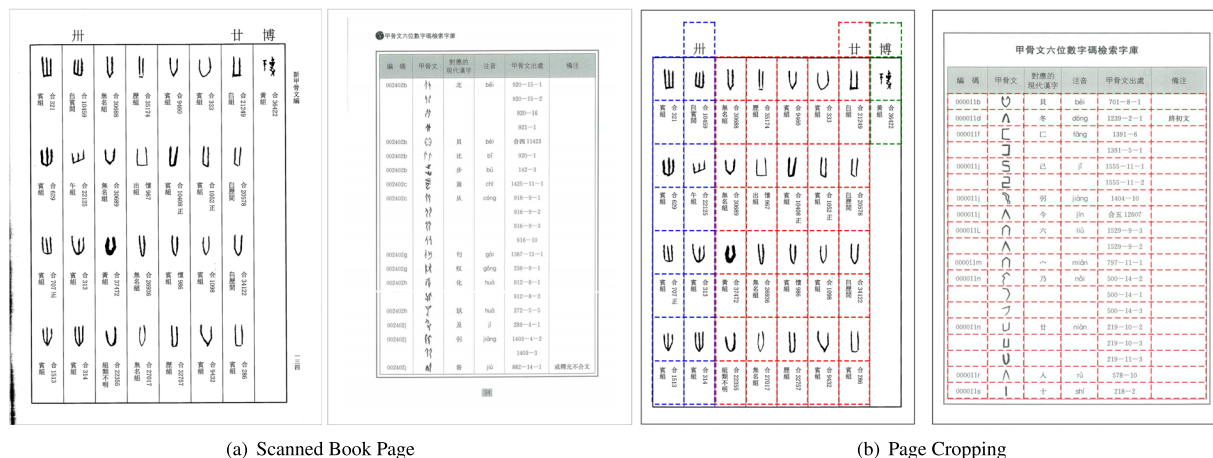


Fig. 3 Extraction of OBC images from books. *New Compilation of Oracle Bone Scripts* (left)¹⁵ and *Oracle Bone Script: Six Digit Numerical Code* (right)¹⁶.



Fig. 4 Screenshots of example website pages from YinQiWenYuan and GuoXueDaShi.

B. Oracle Bone Script: Six Digit Numerical Code (甲骨文六位数字码检索字库 <https://books.google.com/books?id=pgvaxQEACAAJ>)¹⁶ assigns digital codes to OBCs, annotating each code with its corresponding oracle bone character, modern Chinese character form, provenance, and other relevant details.

Since these books do not provide electronic databases or original image data, we manually scanned the pages of these books, obtaining 1,054 and 700 pages from books A and B, respectively. An example of scanned pages is presented in Fig. 3(a).

Websites. With the widespread adoption of the Internet, websites have emerged as an alternative for hosting oracle bone data, offering more convenient retrieval capabilities. We have designed a web crawler program to collect data from the following websites:

- C. **GuoXueDaShi** (国学大师 <https://www.guoxuedashi.net/jgwhj/>), initiated and maintained by enthusiasts of Chinese classical studies, which includes various historical texts including dictionaries, histories, etc. A screenshot of the website is shown in Fig. 4(b).
- D. **YinQiWenYuan** (殷契文渊 <https://jgw.aynu.edu.cn>) is a data platform maintained by the Key Laboratory of Oracle Bone Inscriptions Information Processing, archives various types of data, including photos of the original oracle bones, transcribed characters, and related research articles. A screenshot of the website is shown in Fig. 4(a).

These websites feature well-organized collections of OBC images, which have been meticulously scanned, cropped, and aligned. They are systematically categorized across various web pages, facilitating the use of web crawler technology to download these images in a categorized format efficiently.

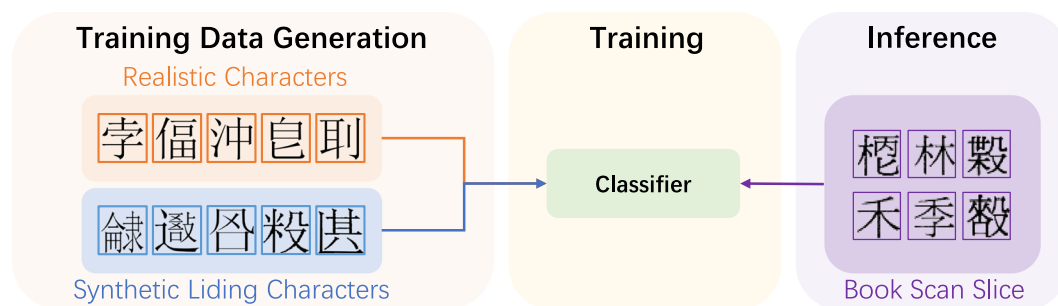


Fig. 5 Schematic of the proposed category assigner.

Databases. In recent years, the digitalization of ancient manuscripts and advancements in handwritten text recognition technology have opened new avenues in the study of OBC, which has led to the proposal of relevant datasets. We have included the following databases in HUST-OBC.

- E. **HWOBC** (<https://jgw.aynu.edu.cn/home/down/detail/index.html?sysid=2>) is a database specifically designed for the study and recognition of handwritten OBCs⁹. Compared to other books and websites that process or trace rubbings, to obtain more extensive handwritten samples of OBCs, the HWOBC dataset hired experts to manually draw each character glyph on a 400×400 pixel white background using a PC or smartphone, and then upload them to create a richer set of 83,245 handwritten OBC images.

Automatic annotation. Through data acquisition, we have gathered raw data from diverse sources. However, this data, in its current format, is not immediately usable. Hence, it necessitates further processing, including tasks such as cropping, annotating, and filtering.

Books. The raw data for the books consists of scanned images of pages, each displaying several OBCs along with their corresponding annotations in modern standard Chinese. As shown in Fig. 3(a), despite differences in the layout of the *New Compilation of Oracle Bone Scripts* (left) and *Oracle Bone Script: Six Digit Numerical Code* (right), they both employ a table-like vertical format. This arrangement facilitates the use of computer vision algorithms, such as edge detection, to automatically extract content from these pages. Specifically, as shown in Fig. 3(b), we employed edge detection and other techniques from the OpenCV toolkit¹⁷ to crop the original scanned images by oracle bone characters, thereby obtaining individual slices of these characters. These slices are then categorized according to the layout rules, with each assigned a corresponding category ID. For example, as illustrated on the left side of Fig. 3(b), the top of each column in the book is marked with the modern Chinese character equivalent to the OBC. If a column lacks such a marking, it implies that it belongs to the same category as the adjacent column on its right. In the figure, we used different colors of dashed lines to distinguish between categories. Using this method, we extracted 24,558 and 14,053 OBC images from source A and B books, respectively.

Although the slices are grouped during the cropping process, the corresponding modern Chinese character of each category remains unknown. A straightforward solution to determine the specific characters for each category is to use OCR techniques to recognize the marks in the books. However, most of the off-the-shelf OCR engines were trained only on commonly used modern Chinese characters and struggled to recognize the uncoded *Liding* and unknown characters that may be presented in these books. To address this issue, we trained a category assigner (see Fig. 5) to automatically identify these labels. The specific training procedures are detailed as follows:

1. **Training Data Generation:** The Chinese character labels we need to identify are all in standard print typeface (as seen in the Chinese characters on the outside of the table at the top of Fig. 3(b), left side), and each cropped image contains only a single, individual character. Thus, we can conveniently generate corresponding training samples using a similar SimSun font. As shown in the block on the left side of Fig. 5, we generated font images for all realistic Chinese characters according to the Ideographic Description Sequence (IDS) and assigned each a unique category ID. Additionally, to address the recognition of uncommon characters that may appear in books, we randomly synthesized *Liding* text using components like radicals of Chinese characters to serve as the *Liding* category for training purposes. Practically, we generated one image for each of the total 88,899 Chinese characters included in the IDS and randomly synthesized α *Liding* character images in each training epoch.
2. **Training:** Since each sample image contains only a single character, it is sufficient to train a simple classifier for recognition. For this purpose, we employed ResNet-50¹⁸ as the backbone network to train the classifier. Additionally, we utilized a weighted balanced cross-entropy loss L to address the issue of the imbalance in the number of training samples across different categories:

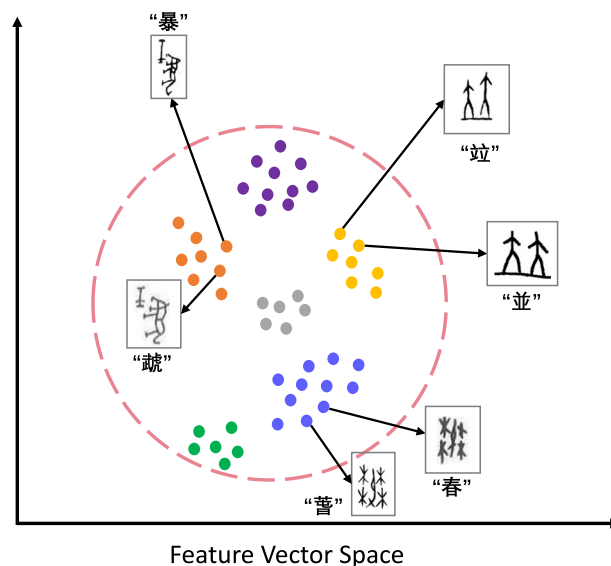


Fig. 6 Oracle bone character images classified into the redundant category exhibit higher similarity in the feature vector space.

$$L = -\frac{1}{N+1} \left[\sum_{i=1}^N y_i \log(p_i) + \frac{1}{\alpha} y_{N+1} \log(p_{N+1}) \right] \quad (1)$$

where N is the number of categories, P_i and Y_i respectively represent the probabilities of the predicted and true labels being the i^{th} category, taking values of 1 or 0, and α is the number of synthesized *Liding* samples in each training epoch.

3. **Inference:** During the inference phase, we input the Chinese character label images, which are cropped from the books, into the classification model trained in the second step. This process helps us determine the corresponding Chinese character for each category ID or whether it is an uncoded *Liding* character.

After completing the aforementioned procedures, all OBC images contained within the scanned pages of sourced books acquired during the data acquisition phase have been automatically extracted and accurately categorized according to their respective classes.

Websites & Databases. The images of the OBCs collected from websites and databases have already been pre-processed by scanning, cropping, and alignment. Therefore, there is no need to design automatic annotation algorithms for this data, unlike the approach required for data from book sources. However, the following essential processes are still required:

1. **Filtering:** It is important to note that a portion of the data on the GuoXueDaShi website, contributed by enthusiasts of ancient Chinese culture, cannot be fully guaranteed for reliability. Specifically, these OBC images are of higher resolution and quality compared to other sources, but their unreliability stems from their labels. Currently, only about 1,500 categories of OBC have been deciphered, whereas the GuoXueDaShi website has 2,756 categories. This indicates that some undeciphered OBCs have been labeled by enthusiasts without expert verification, making them unreliable. Consequently, in our filtration process, we cross-referenced these with other sources. This allowed us to identify 1,390 categories of OBC images that were unique to GuoXueDaShi and could not be verified. As a result, we retained only 1,366 out of the initial 2,756 categories after excluding these unverifiable samples. The samples of these 1390 categories, due to their lack of reliability, have not been classified as deciphered or undeciphered samples and are stored separately in the dataset.
2. **Code Matching:** The OBC images from online and database sources are marked with specific codes, which we further mapped into modern Chinese characters. For the oracle bone inscriptions of YinQiWenYuan and HWOBC, the HUST-OBC dataset only includes individual oracle bone characters, not compound characters. The term 'compound characters' refers to oracle bone characters corresponding to two or more words. Moreover, HWOBC is classified based on the character forms, leading to multiple character forms corresponding to the same Chinese character. Here, we merge them into the same category based on the corresponding Chinese character.

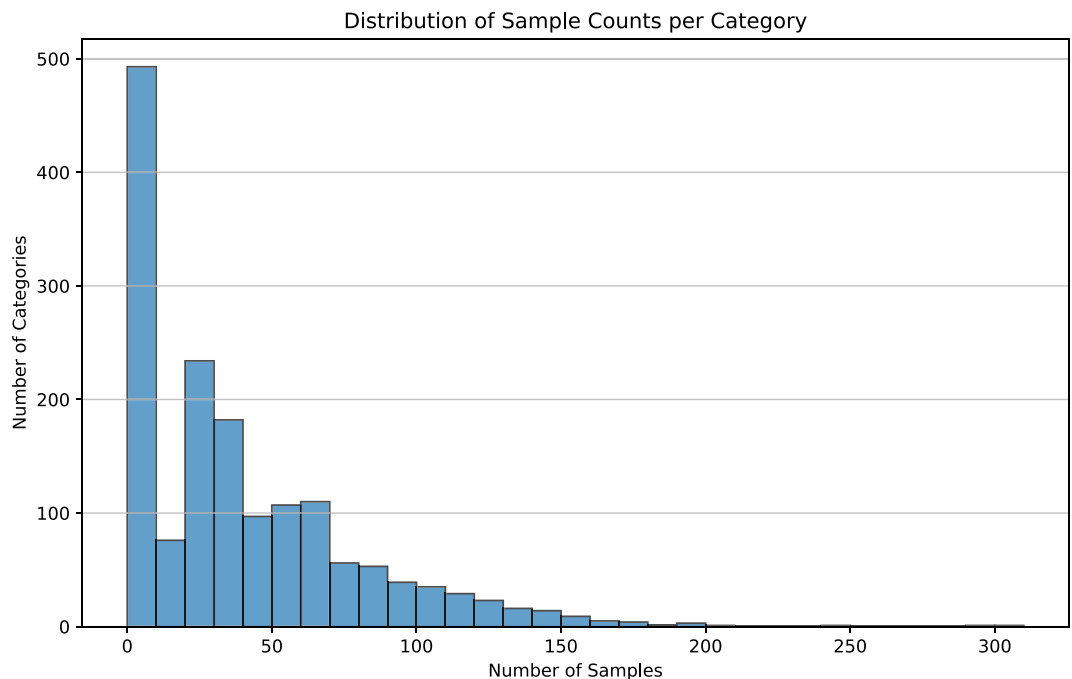


Fig. 7 Histogram illustrating the distribution of sample counts across various categories in the HUST-OBC deciphered split.

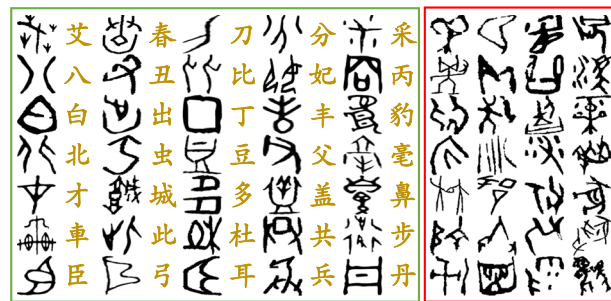


Fig. 8 Samples of **deciphered** and **undeciphered** Oracle Bone Characters from the HUST-OBC dataset.

Integration. In the stages of *Data Acquisition and Automatic Annotation*, images of OBCs from distinct sources were collected and annotated. However, it is important to note that the annotation conventions for one OBC may vary depending on the source. For instance, as shown in Table 2, some sources might use standard modern Chinese characters for annotation, while other sources may prefer using corresponding Variant Chinese characters¹⁹ (https://en.wikipedia.org/wiki/Variant_Chinese_characters) for annotation. This leads to a scenario where images of OBCs that should belong to the same category are classified into different categories, creating redundant categories. Table 2 illustrates this with examples of duplicate annotations, where each row shows how the same OBC image is categorized differently under the *Modern Character Category* and the *Variant Character Category*. To eliminate these redundancies, we integrate the data from different sources. For this purpose, we trained a widely-used unsupervised visual representation learning model MoCo²⁰, with OBC images from all sources. Subsequently, all the oracle bone images were encoded into a feature vector by the model. As illustrated in Fig. 6, by calculating the similarity of these feature vectors, we merged similar samples into the same categories. In this way, we were able to reduce the original 1,781 categories obtained from different sources to 1,588, eliminating redundant categories.

Validation. After undergoing all the procedures, we obtained a preliminary dataset. However, due to potential errors that might occur in the automated data acquisition and annotation process, we enlisted the expertise of OBC scholars from Anyang Normal University to meticulously review our dataset. Using authoritative books and the HWOBC database fonts as reference standards, they compared and evaluated OBC data in the HUST-OBC, discarding samples with errors and retaining the relatively accurate ones. This review produced the HUST-OBC dataset.

Ground-truth Category	#Samples	Sample Input from Different Sources					Test Accuracy
		A	B	C	D	E	
子	307						90.32%
祝	295						93.10%
牢	209						100.00%
羌	199						100.00%
伐	190						100.00%
逐	180						94.44%
河	179						100.00%
不	106						80.00%
白	93						100.00%
電	63						100.00%
安	55						100.00%

Table 4. Validation accuracy of ResNet-50 on selected categories of the HUST-OBC dataset.

Data Records

The HUST-OBC¹⁴ comprises a total of 140,053 images sourced from five different origins, divided into deciphered and undeciphered sections. The deciphered section contains 77,064 images spanning 1,588 categories of individual characters, and the undeciphered section features 62,989 images across 9,411 categories of characters. Due to the lack of annotations for undeciphered categories, there may be duplicates among these 9,411 categories of undeciphered OBCs, which can only be merged once they are deciphered. Table 3 provides detailed statistics of the OBC images obtained from these sources. Additionally, Fig. 7 presents a distribution histogram showing the number of sample images per category in our dataset. It reveals that most categories have fewer than 10 sample images, with the largest category boasting over 300 images.

For efficient retrieval, the HUST-OBC is organized and stored by category names. Each image file is systematically named following the format <source>_<label>_<filename>, encapsulating its origin, category number, and sequence number, and is stored in folders named after their category number. For the deciphered categories, we have corresponding category numbers and the corresponding Chinese dictionary stored in a UTF-8 encoded JSON file. Figure 8 demonstrates some deciphered and undeciphered OBCs from the HUST-OBC.

Technical Validation

One of the primary objectives in creating the HUST-OBC is to facilitate future AI-assisted tasks in deciphering OBCs. To this end, we further assessed the quality of the dataset by employing it to train AI models. Specifically, we divided the deciphered section of the HUST-OBC dataset into a training set, a validation set, and a test set using stratified sampling with proportions of 8:1:1, using them for training, validation, and testing in image classification tasks. Due to the limitation of classification models not being able to categorize unseen classes, we allocated all classes with only one sample into the training set. The accuracy of image classification can reflect the quality of the dataset to some extent. If the images in the dataset are of poor quality or have many labeling errors, the classifier’s accuracy will be low, and vice versa. We employed the widely-used ResNet-50¹⁸ as the backbone network for training. We tested the test set using the model that achieved the highest accuracy on the validation

set, ultimately achieving a classification accuracy of 94.6% and a macro-average F1 score of 0.914, which validates the dataset's quality and potential academic value. Table 4 shows the model's recognition accuracy in some categories and provides example input images from different sources.

Licenses

The dataset is released under a non-commercial license, CC BY-NC 4.0 (<https://creativecommons.org/licenses/by-nc/4.0/deed.en>), which permits users to reuse and reproduce the dataset for research purposes.

Usage Notes

The HUST-OBC is available as a compressed archive, comprising three distinct folders. These folders separately house images of OBCs. The first one is for those that have already been deciphered, the second one is for those still awaiting interpretation, and the third one is for unreliable data from GuoXueDaShi. Within each folder, subfolders are organized by categories, containing images of OBCs corresponding to their respective categories. For more information, please see (<https://github.com/Pengjie-W/HUST-OBC>).

Code availability

OpenCV toolkit is used to detect the borders in scanned book pages, which is available at <https://opencv.org/>.

The models and code for Chinese OCR, MoCo, and ResNet50 for Validation are available at (<https://github.com/Pengjie-W/HUST-OBC>).

Received: 22 January 2024; Accepted: 20 August 2024;

Published online: 06 September 2024

References

1. Boltz, W. G. Early chinese writing. *World Archaeology* **17**, 420–436 (1986).
2. Keightley, D. N. The shang state as seen in the oracle-bone inscriptions. *Early China* **5**, 25–34, <https://doi.org/10.1017/S0362502800006118> (1979).
3. Bazerman, C. *Handbook of research on writing: History, society, school, individual, text* (Routledge, 2009).
4. Gao, J. & Liang, X. Distinguishing oracle variants based on the isomorphism and symmetry invariances of oracle-bone inscriptions. *IEEE Access* **8**, 152258–152275 (2020).
5. LeCun, Y. *et al.* Handwritten digit recognition with a back-propagation network. *Advances in Neural Information Processing Systems* **2** (1989).
6. Graves, A. & Schmidhuber, J. Offline handwriting recognition with multidimensional recurrent neural networks. *Advances in Neural Information Processing Systems* **21** (2008).
7. Bhunia, A. K., Das, A., Bhunia, A. K., Kishore, P. S. R. & Roy, P. P. Handwriting recognition in low-resource scripts using adversarial learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4767–4776 (2019).
8. Silver, D. *et al.* Mastering the game of go with deep neural networks and tree search. *nature* **529**, 484–489 (2016).
9. Li, B. *et al.* Hwobc-a handwriting oracle bone character recognition database. *Journal of Physics: Conference Series* **1651**, 012050, <https://doi.org/10.1088/1742-6596/1651/1/012050> (2020).
10. Fu, X., Yang, Z., Zeng, Z., Zhang, Y. & Zhou, Q. Improvement of oracle bone inscription recognition accuracy: A deep learning perspective. *ISPRS International Journal of Geo-Information* **11**, <https://doi.org/10.3390/ijgi11010045> (2022).
11. Yue, X., Li, H., Fujikawa, Y. & Meng, L. Dynamic dataset augmentation for deep learning-based oracle bone inscriptions recognition. *J. Comput. Cult. Herit.* **15** <https://doi.org/10.1145/3532868> (2022).
12. Guo, J., Wang, C., Roman-Rangel, E., Chao, H. & Rui, Y. Building hierarchical representations for oracle character and sketch recognition. *IEEE Transactions on Image Processing* **25**, 104–118, <https://doi.org/10.1109/TIP.2015.2500019> (2016).
13. Li, B. *et al.* Oracle bone inscriptions multi-modal dataset (2024).
14. Wang, P. *et al.* Hust-obc, [figshare](https://doi.org/10.6084/m9.figshare.25040543.v3), <https://doi.org/10.6084/m9.figshare.25040543.v3> (2024).
15. Zhao, L. *Xin Jia Gu Wen Bian (Revised Edition)* (Fujian People's Publishing House, 2014).
16. Liu Zhixiang, L. X. *Jia Gu Wen Liu Wei Shu Zi Ma Jian Suo Zi Ku* (Sichuan Lexicographical Publishing House, 2019).
17. Bradski, G. The opencv library. *Dr. Dobbs' Journal: Software Tools for the Professional Programmer* **25**, 120–123 (2000).
18. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778 (2016).
19. Bökset, R. *Long story of short forms: the evolution of simplified Chinese characters*. Ph.D. thesis, Institutionen för orientaliska språk (2006).
20. He, K., Fan, H., Wu, Y., Xie, S. & Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9729–9738 (2020).

Acknowledgements

The authors thank the Key Laboratory of Oracle Bone Script Information Processing, Ministry of Education, Anyang Normal University for providing ancient text data sources and review of the dataset construction.

Author contributions

Pengjie Wang conducted experiments on Automatic Annotation, Integration, and Technical Validation. Haisu Guan, Jinpeng Wan, Pengjie Wang, and Kaile Zhang collectively obtained data from books. Pengjie Wang and Kaile Zhang collectively crawled data from websites. Xinyu Wang analyzed the results. Shengwei Han and Yongge Liu, as oracle bone script experts, supervised and assisted in the establishment of the dataset. Yuliang Liu provided guidance on the entire project. Xiang Bai provided laboratory resources. All authors reviewed the manuscript.

Competing interests

The corresponding author is responsible for providing a [competing interests statement](#) on behalf of all authors of the paper. This statement must be included in the submitted article file.

Additional information

Correspondence and requests for materials should be addressed to Yuliang Liu.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2024