Computers and Artificial Intelligence；Vol.3, No.3; 2026
E-ISSN: 3006 -3728 P-ISSN：3006-371X
DOI: https://doi.org/10.70267/cai.26v3n3.8490
Published by: Zeus Press

Research on Oracle Bone Inscription Detection and Recognition Algorithm Based on YOLO11-ViT

Zhiyuan Lu*

School of Electronic and Information Engineering, University of Science and Technology Liaoning, Anshan, Liaoning, China

**Corresponding author: Zhiyuan Lu.*

Abstract

Oracle bone inscriptions are the earliest mature writing system discovered in China and constitute an important historical source for the origin of Chinese characters and traditional Chinese culture. However, oracle bone rubbing images are usually affected by manual carving variations, long-term underground burial, imbalanced sample distributions, severe noise interference and highly similar character structures, which considerably restrict recognition accuracy. To address these challenges, this paper proposes a two-stage oracle bone inscription detection and recognition model integrating YOLO11-ViT. First, the images are preprocessed using grayscale conversion, Otsu binarization, Gaussian denoising and morphological optimization. Second, YOLO11m is employed to accurately detect and localize oracle bone characters in the rubbing images. Finally, a Vision Transformer model is used to classify the cropped single-character regions. Experiments are conducted on the dataset provided by the 2024 14th MathorCup Mathematical Application Challenge. The results show that the proposed method achieves an mAP@0.5 of 0.92 in the detection stage and an accuracy of 97.60%, a recall of 0.96 and an F1 score of 0.97 in the recognition stage, outperforming the compared methods.

Keywords

oracle bone inscription recognition, yolo11, vision transformer, object detection, image preprocessing

1. Introduction

As the oldest mature writing system known in China, oracle bone inscriptions are not only the source of Chinese character development but also an essential foundation of traditional Chinese culture [1]. Oracle bone rubbings refer to original image materials produced by rubbing excavated tortoise shells, animal bones and other carriers with engraved inscriptions. Due to the diversity and antiquity of these carriers, the obtained rubbing images often suffer from severe damage, strong noise interference and uneven sample distributions. These factors directly limit the recognition accuracy of oracle bone inscription images [2]. Therefore, research on oracle bone rubbing recognition has important academic value and practical significance for activating cultural heritage resources and promoting the inheritance and development of Chinese civilization.

With the rapid development of artificial intelligence, deep learning models have shown increasing advantages in image processing and have consequently been applied to the recognition of this ancient writing system. According to the structural characteristics of oracle bone characters, researchers have developed

various specialized deep learning frameworks, including convolutional neural networks (CNNs) and recurrent neural networks (RNNs), to achieve accurate segmentation and recognition of single characters in oracle bone rubbings [3,4]. Nevertheless, existing automatic recognition systems still face several technical bottlenecks. First, oracle bone characters show high intra-class and inter-class visual similarity, and many semantically different characters differ only in subtle local structures. Second, the sample distribution is highly imbalanced, and large differences in character frequency lead to insufficient training samples for rare classes. Third, oracle bone rubbing images are generally degraded, with blurred character edges and complex background noise, which poses serious challenges to algorithm robustness.

To solve these problems, this paper proposes a two-stage oracle bone inscription detection and recognition algorithm based on YOLO11-ViT. The method first preprocesses the original rubbings, then uses YOLO11m for character detection and localization, and finally applies a Vision Transformer (ViT) model for character classification and recognition.

The main contributions of this paper are as follows. First, a complete preprocessing pipeline suitable for oracle bone rubbing images is designed. Second, the YOLO11m model is introduced into the oracle bone character detection task. Third, the ViT model is adopted for character recognition, which improves the discrimination of visually similar oracle bone characters.

2. Oracle Bone Inscription Image Preprocessing

Oracle bone rubbing images are commonly affected by three types of interference: point-like noise, artificial texture and inherent carrier texture. The rubbings excavated in archaeological contexts also have several common properties. Characters may appear in arbitrary regions of an image and vary greatly in shape and size; image surfaces are often seriously degraded, making inscriptions difficult to identify; and cracks similar to oracle bone character strokes may exist in the rubbings. To improve the accuracy of subsequent detection and recognition, this study constructs a systematic image preprocessing method.

2.1 Grayscale Conversion

The color image is converted into a grayscale image to eliminate the interference of color information. Let the original image be I(x,y) and the grayscale image be I'(x,y), where R, G and B denote the red, green and blue components of a pixel, respectively. The grayscale conversion is expressed as:

$$I'(x,y) = 0.299R(x,y) + 0.587G(x,y) + 0.114B(x,y) \quad (1)$$

2.2 Otsu Binarization

After grayscale conversion, binarization is performed to convert the grayscale image into a black-and-white image. This paper adopts the Otsu algorithm to automatically determine the optimal threshold. The basic principle is to divide the grayscale histogram into two parts using the optimal threshold so that the between-class variance is maximized [5]. Let ω_0 and ω_1 be the proportions of foreground and background pixels, and let μ_0 and μ_1 be the mean gray values of the two classes. The between-class variance is defined as:

$$\sigma\_b^2(T) = \omega_0(T)[\mu_0(T) - \mu\_T]^2 + \omega_1(T)[\mu_1(T) - \mu\_T]^2 \quad (2)$$

The optimal threshold T* maximizes the between-class variance. Based on this method, a binarized image with richer details and finer character contours can be obtained.

2.3 Gaussian Filtering for Denoising

Point-like noise widely distributed in oracle bone rubbing images affects the determination of character segmentation regions. This paper uses Gaussian filtering to smooth the images and suppress noise interference. Gaussian filtering transforms pixels according to a Gaussian distribution and removes image noise as much as possible. The two-dimensional Gaussian function is defined as:

$$G(x,y) = 1/(2\pi\sigma^2)\exp(-(x^2+y^2)/(2\sigma^2)) \quad (3)$$

where σ denotes the standard deviation of the Gaussian distribution. The Gaussian function can decompose a two-dimensional transformation into the product of two one-dimensional transformations, thereby simplifying

the processing procedure and improving efficiency.

2.4 Morphological Operation

This paper uses morphological opening to process the images. Opening operation shrinks object boundaries and eliminates bright spots in the image, thus removing small noise points. Let A be the original image and B be the structural element. The opening operation, which consists of erosion followed by dilation, is defined as:

$$A \circ B = (A \ominus B) \oplus B \tag{4}$$

After the above preprocessing steps, the image quality is significantly improved, and most interference elements in the rubbings are removed.

Figure 1: Preprocessing results of original oracle bone rubbing images.

3. Detection and Recognition Model Design

3.1 Overall Framework

The proposed oracle bone inscription detection and recognition algorithm adopts a two-stage architecture. The first stage is the detection stage, in which YOLO11m is used to locate and detect characters in preprocessed rubbing images. The second stage is the recognition stage, in which the detected character regions are cropped and input into a ViT model for classification. The model architecture mainly consists of four key modules: an input layer, a feature extraction layer, a feature fusion layer and a detection layer.

3.2 YOLO11m Detection Model

YOLO11 is a recent member of the YOLO family and adopts an improved CSPDarknet backbone and a PANet-based feature fusion structure [6]. The feature extraction layer uses the YOLO11 backbone, whose core components include convolutional layers, residual blocks and downsampling layers. The detection stage adopts the SIoU loss function for bounding box regression. The intersection over union (IoU) is calculated as:

$$IoU = (A \cap B)/(A \cup B) \tag{5}$$

where A denotes the predicted bounding box and B denotes the ground-truth bounding box. The detection layer uses an anchor mechanism and applies non-maximum suppression (NMS) to remove redundant detection results.

In terms of network structure, YOLO11m introduces the C2f module to replace the traditional C3 module, achieving more efficient gradient flow and feature reuse through cross-stage partial connections. Meanwhile,

YOLO11 adopts a decoupled detection head that separates the classification branch from the localization branch, effectively reducing mutual interference between the two tasks and improving detection accuracy. In the training configuration, the initial learning rate is set to 0.01, a cosine annealing strategy is used for learning rate scheduling, the batch size is set to 16, the input image resolution is uniformly resized to 640 × 640, and the number of training epochs is 150. Data augmentation strategies include random flipping, color jittering and Mosaic augmentation to alleviate the imbalanced sample distribution in the oracle bone dataset.

3.3 ViT Recognition Model

Vision Transformer (ViT) is a pioneering work that applies the Transformer architecture to image classification tasks [7]. ViT divides an image into a sequence of fixed-size patches, applies linear embedding and position encoding, and then feeds the sequence into a Transformer encoder. The self-attention mechanism is used to capture global features. The self-attention calculation is expressed as:

$$\text{Attention}(\text{Q}, \text{K}, \text{V}) = \text{softmax}(\text{QK\^{}T/sqrt(d\_k)})\text{V} \tag{6}$$

where Q, K and V denote the query, key and value matrices, respectively, and d_k is the dimension of the key vector. Compared with traditional CNNs, ViT demonstrates stronger feature extraction capability on large-scale datasets and is particularly suitable for processing oracle bone characters with complex morphological variations.

This paper adopts ViT-Base as the recognition backbone. The input image is divided into patches of 16 × 16 pixels. The encoder contains 12 Transformer blocks, the number of attention heads is set to 12, and the hidden dimension is 768. To address the relatively limited training samples of oracle bone characters, the model is initialized with ImageNet-21K pretrained weights and fully fine-tuned on the oracle bone recognition dataset. The learning rate is set to 1×10^{-4}, and the model is trained for 50 epochs. To handle class imbalance, a weighted cross-entropy loss function is introduced, assigning higher weights to low-frequency characters to improve the recognition ability for rare oracle bone classes.

4. Experimental Results and Analysis

4.1 Dataset and Experimental Environment

This paper uses the detection and recognition datasets provided by the 2024 14th MathorCup Mathematical Application Challenge [8]. The single-character detection and segmentation dataset is derived from the oracle bone character detection dataset in Yinqi Wenyuan, and the oracle bone single-character recognition data are derived from the OBC306 dataset in Yinqi Wenyuan. The recognition dataset contains 76 commonly used oracle bone characters and their corresponding character forms, with a total of 40,618 images. After preprocessing, 9,134 and 60,248 images are obtained for the corresponding tasks and are randomly divided into training, validation and test sets in an 8:1:1 ratio. The experimental environment is Ubuntu 22.04 with an NVIDIA GeForce RTX 4090 GPU, PyTorch 2.1.3 and Python 3.10.

4.2 Evaluation Metrics

Precision (P), recall (R), F1 score and mean average precision (mAP) are used as evaluation metrics in this paper. They are calculated as follows:

$$\text{P} = \text{TP}/(\text{TP} + \text{FP}), \text{R} = \text{TP}/(\text{TP} + \text{FN}), \text{F1} = 2\text{PR}/(\text{P} + \text{R}), \text{mAP} = (1/\text{N})\Sigma\text{AP\_i} \tag{7}$$

where TP denotes true positives, FP denotes false positives and FN denotes false negatives. mAP denotes the mean of the average precision values over all classes.

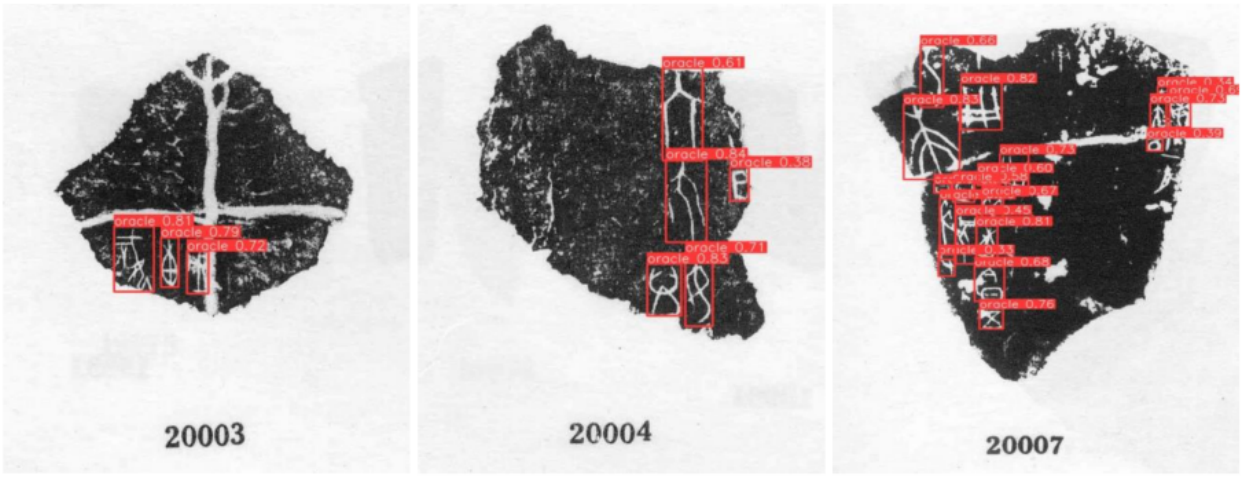
4.3 Detection Model Results

Table 1 presents the performance comparison of different detection models on the oracle bone dataset. YOLO11m achieves the best performance on all metrics, with an mAP@0.5 of 0.92, which is two percentage points higher than that of YOLOv8m.

Table 1: Performance comparison of detection models

Model	P	R	mAP@0.5	F1
YOLOv5m	0.85	0.82	0.87	0.83
YOLOv8m	0.89	0.88	0.90	0.89
YOLO11m (proposed)	0.93	0.89	0.92	0.91

Figure 2: Examples of oracle bone character detection results



4.4 Recognition Model Results

Table 2 presents the performance comparison of different recognition models. The ViT model achieves a validation accuracy of 97.60%, a recall of 0.96 and an F1 score of 0.97, outperforming the traditional CNN models.

Table 2: Performance comparison of recognition models.

Model	Accuracy (%)	R	F1
AlexNet	82.45	0.80	0.81
ResNet50	91.48	0.90	0.91
GoogleNet	93.53	0.92	0.93
ViT (proposed)	97.60	0.96	0.97

Figure 3: Examples of oracle bone detection and recognition results.

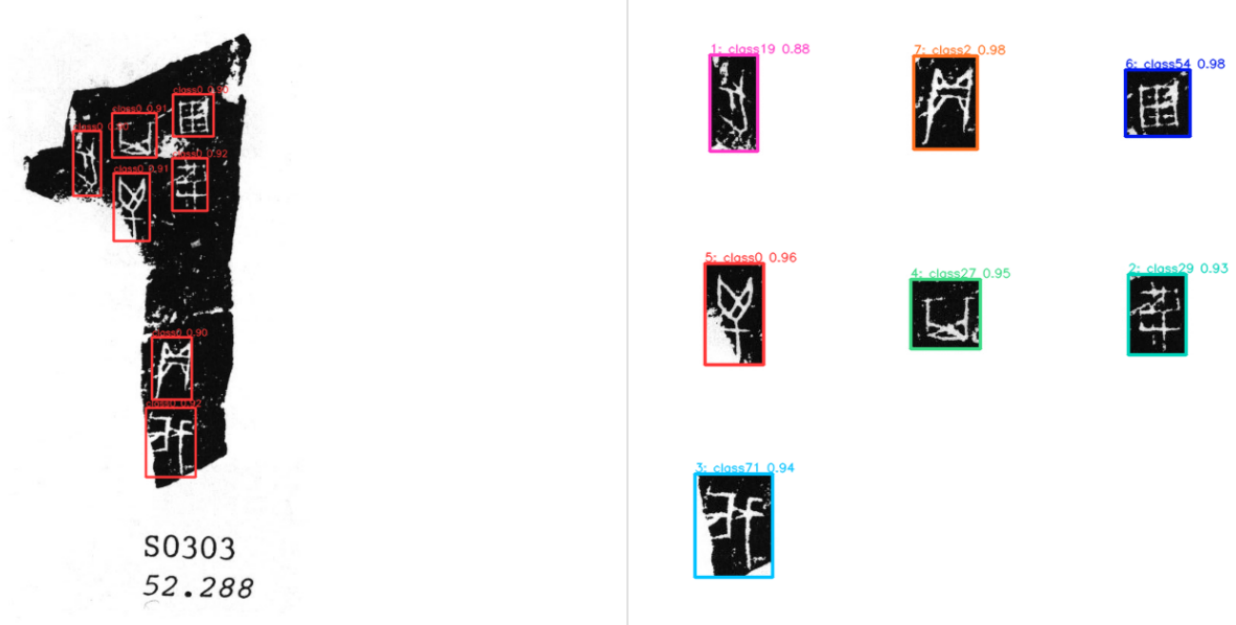


Figure 3 shows the detection and recognition results for the oracle bone rubbing numbered S0303. A total of seven oracle bone characters are recognized in this example, and each character is annotated with its

coordinate position in the rubbing. The corresponding modern Chinese characters are rabbit, prisoner, gate, mouth, ten thousand, field and bing, respectively. The confidence scores are generally high, with most values above 0.90. In particular, the confidence scores for field and bing both reach 0.98, indicating that the recognition results are reliable.

5. Conclusion

This paper proposes a two-stage oracle bone inscription detection and recognition algorithm based on YOLO11-ViT. A complete image preprocessing pipeline, including grayscale conversion, Otsu binarization, Gaussian filtering and morphological operation, is designed to reduce noise interference in rubbing images. YOLO11m is then used for accurate character detection and localization, and a ViT model is used for high-accuracy character classification and recognition.

The experimental results show that the proposed method achieves an mAP@0.5 of 0.92 in the detection stage and an accuracy of 97.60%, a recall of 0.96 and an F1 score of 0.97 in the recognition stage, outperforming the compared methods. Theoretically, this study provides a feasible technical route for combining object detection and Transformer-based visual recognition in oracle bone inscription research. Practically, the proposed model can support digital preservation, archaeological collation and intelligent analysis of oracle bone documents. Compared with conventional CNN-based recognition methods, the proposed approach improves the discrimination of visually similar characters and strengthens robustness under noisy rubbing conditions. However, the study still has limitations, including the scale of training data, the dependence on two-stage processing and the insufficient use of contextual semantic information. Future work will expand the training dataset, explore end-to-end joint training for detection and recognition, and investigate oracle bone semantic understanding based on contextual information.

References

[1] Liu, Y, Lu, Y, Wei, YC, et al. (2023). Research status and prospects of oracle bone inscription recognition technology. *Knowledge Management Forum*, *8*(2), 115–125.

[2] Mao, YF, Bi, XJ. (2023). Oracle bone inscription recognition on rubbings using an improved ResNeSt network. *CAAI Transactions on Intelligent Systems*, *18*(3), 450–458.

[3] Zhang, YK, Zhang, H, Liu, YG, et al. (2021). Oracle bone character recognition based on cross-modal deep metric learning. *Acta Automatica Sinica*, *47*(4), 791–800.

[4] Wang, HB. (2019). R*esearch on Oracle Bone Character Detection and Recognition Based on Deep Learning* (Master's Thesis). South China University of Technology, Guangzhou.

[5] Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, *9*(1), 62–66.

[6] Redmon, J, Divvala, S, Girshick, R, et al. (2016). You only look once: unified, real-time object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. p. 779–788.

[7] Dosovitskiy A, Beyer L, Kolesnikov A, et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. International Conference on Learning Representations. Available from: https://arxiv.org/abs/2010.11929

[8] Huang, S, Wang, H, Liu, Y, et al. (2019).OBC306: a large-scale oracle bone character recognition dataset. International Conference on Document Analysis and Recognition. p. 681–688.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.