基于监督式深度学习的中文古籍版本鉴定法

A Deep-Learning Model for Edition Identification of Premodern Chinese Rare Books①

韦胤宗 / 南洋理工大学（第一作者）

付一茗 / 都柏林城市大学

王思启 / 中国科学院大学

高树伟 / 北京大学

刘天成 / 香港科技大学（广州）

王泽宇 / 香港科技大学（广州）（通讯作者）

佟 馨 / 香港科技大学（广州）（通讯作者）

Wei Yinzong[1] / Nanyang Technological University, Singapore, First Author

Fu Yiming+ / Dublin City University, Dublin

Wang Siqi+ / University of Chinese Academy of Sciences, Beijing

Gao Shuwei / Peking University, Beijing

Liu Tiancheng / The Hong Kong University of Science and Technology (Guangzhou)

Wang Zeyu* / The Hong Kong University of Science and Technology (Guangzhou), Corresponding Author

Tong Xin* / The Hong Kong University of Science and Technology (Guangzhou), Corresponding Author

ABSTRACT: Edition identification is a crucial task in the field of bibliography and

①This paper is supported by the "Fundamental Research Funds for the Central Universities" (No. 413000217).

+ These authors contributed equally.

history of the book. However, traditional methods need to improve in terms of efficiency and accuracy. This paper uses deep learning technology to construct a more suitable framework for the edition identification of premodern Chinese books. We preprocess the scanned pages of rare books using global threshold binarization and heuristic rule-based methods. Subsequently, we build a deep learning framework based on InceptionResNet-V2 that is better suited for the task of Chinese rare book edition identification. We conducted experiments on eleven categories of datasets from different periods and regions, comparing our method with the state-of-the-art deep network structures. The results show that our constructed framework is effective in the task of Chinese rare book edition identification, with an accuracy rate of 91.3% and a recall rate of 83.4%.

Keywords: Chinese Rare Books; Edition Identification; Machine Learning; Deep Learning

Introduction

The Study of Book Edition is a discipline that examines the characteristics of the physical medium of literature, as well as the issues related to its production, circulation, evolution, identification, and collection. Its scope of study is paper as the medium, including the contents of the "Printing Era" such as woodblock printing, movable type printing, and manuscripts after the invention of printing.[1] This period spans from the Middle Tang Dynasty to the Republic of China, covering approximately 1,000 years. Edition identification, as the fundamental aspect of edition study, involves verifying the period, region, and pattern of printing, as well as the edition inheritance of rare books. This process is crucial for assessing the documentary and cultural relic values of rare books. There are two major approaches of book edition identification: a) textual evidential research, relying on textual evidence such as publication information (records in the front pages, prefaces, and postscripts of rare books), authorial information; and b) a method traditionally called "observing the wind and watching the air", focusing on the physical and visual features of rare books. Recently, more and more researchers tend to hold the idea that the second approach is more reliable.[2] However, traditionally, edition researchers often adopt a qualitative approach to appraise the editions of rare books based on characteristics such as fonts, relationships

[1] Yinzong Wei, “Bibliography and Philology, Textual Studies, and the History of the Book, in the Context of Chinese and Western Scholarship,” *The Documentation*, vol.3, 2023, pp.112-129.

[2] Xiang Shi, “Observing the Wind and Watching the Air, Typology, and Evidential Research: On the Methodology of Edition Identification,” *Wenshi*, vol.4, 2020, pp.191-256.

between characters, papers, seals, etc.[1] This research method heavily relies on the prior knowledge and experience of edition researchers, consuming a significant amount of time and human resources, and is usually influenced by subjective factors to a certain extent. Different researchers often hold divergent opinions on the identification of the same book. Therefore, the use of modern technological tools for the identification of rare books has become particularly important. These tools can assist edition researchers in evaluating identification results and proposing new hypotheses.

In the current stage, there has been some research on the analysis of the years, scribes, and regions of manuscripts from ancient and medieval times, before the advent of the printing era, with the assistance of digital technologies. Most of these studies have employed pattern recognition-based methods. Dhali et al. proposed a method combining feature extraction and regression, using a support vector regression model based on handwritten character shape features, and texture features such as curvature and slope, to predict the year of rare Hebrew manuscripts from a small sample dataset. The final mean absolute error was 23.4 years.[2] However, the feature extraction process still requires costly time and effort from edition researchers. Christlein et al. proposed an unsupervised feature-based scribe identification method based on a Greek dataset written on parchment. The study used Scale-Invariant Feature Transform (SIFT) to extract local features from sampling points such as note contours, and then clustered them into global features using k-means clustering, applying support vector machines for classification.[3] This method can be more efficiently expanded to unknown datasets. However, unlike tasks on scribe identification, specific information such as engraving period, region, and publisher plays a crucial role in edition studies. Methods based on deep supervised learning show promise in this

[1] Peng Shen, Lijun Wang, Weiyong Shao, Huanxin Zhang, "Research Progress on the Technology of Rare Books Year Identification," *Paper Science & Technology*, vol.42, no.5, 2023, pp.20-26.

[2] Maruf A. Dhali, Camilo Nathan Jansen, Jan Willem de Wit, Lambert Schomaker, "Feature-Extraction Methods for Historical Manuscript Dating Based on Writing style Development," *Pattern Recognition Letters*, vol.131, 2020, pp. 413–420.

[3] Vincent Christlein, Isabelle Marthot-Santaniello, Martin Mayr, Anguelos Nicolaou, Mathias Seuret, "Writer Retrieval and Writer Identification in Greek Papyri," In Cristina Carmona-Duarte, Moises Diaz, Miguel A. Ferrer, Aythami Morales, eds., *Intertwining Graphonomics with Human Movements: 20th International Conference of the International Graphonomics Society*, IGS 2021, Las Palmas de Gran Canaria, Spain, June 7-9, 2022, pp.76–89.

regard. Cilia et al. built an end-to-end manuscript scribe identification system, using MobileNet-V2 for text line segmentation, classifying text lines based on deep neural network transfer learning, and determining the author category of each page of text through majority voting. This research achieved an accuracy of 96.48% in 12th-century Latin manuscripts.[①] However, unlike the limited 26 characters in Latin, Chinese contains a vast number of single characters, with over forty-seven thousand recorded in the Kangxi Dictionary alone. There are also significant variations in the shape and form of Chinese characters in periods and regions, and there is a lack of effective feature captures. This presents a significant challenge for our identification task.

In recent years, there have been numerous attempts at utilizing deep learning techniques in Chinese rare books-related research, but most of them have focused on areas such as layout segmentation, text region detection, and text recognition.[②] In comparison to these tasks, edition identification imposes more stringent requirements on preserving font features during the image preprocessing stage. Considering the presence of numerous illustrations in Chinese rare books, as well as the complexity of mixed layouts and potential overlapping of seals and text (as shown in Figure 1), complete text extraction often leads to significant metadata loss containing font features.

In this paper, we propose a method based on supervised learning with deep neural networks to assist in the edition identification of Chinese rare books. Firstly, we utilize a global threshold selection method for binarization, separating foreground ink pixels from the background and the potential noises in the scanned image dataset. Then, heuristic rules are applied to remove paratext areas. Finally, we employ supervised learning to compare different neural network structures in search of a more suitable method for identifying Chinese character font features in edition identification.

[①] N. D. Cilia et al., “An End-to-End Deep Learning System for Medieval Writer Identification,” *Pattern Recognition Letters*, vol.129, 2020, pp.137–143.

[②] Weiqi Wang, “Research on Text Detection and Recognition Algorithms for Tangut Rare Books Based on Deep Learning,” Master’s thesis, Ningxia University, 2023；Shanxiong Chen, Xiaolong Wang et al., “A Deep Learning-Based Method for Rare Yi Script Recognition,” *Journal of Zhejiang University (Science Edition)* , vol.46, no.3, 2019, pp.261-269；Shanxiong Chen, Han Xu et al., “A Character Detection Method for Rare Yi Script Documents Based on MSER and CNN,” *Journal of South China University of Technology (Natural Science Edition)* , vol.48, no.6, 2020, pp.123-133.

Dataset

As we know in February 2024, digitalization works related to edition studies have mainly been in scanning and management stages, lacking large-scale standardized datasets applicable for machine learning identification. Therefore, we collected 163 Chinese rare books from multiple sources, all printed in woodblock form on paper, and stored as JPEG images after scanning the entire books. Edition scholars classified them into eleven categories (i.e. editions) based on information such as the dynasty and region of publication. These categories are: 1) Northern and Southern Song Zhe Edition, 2) Southern Song Printing House Edition (in Zhe region), 3) Early Southern Song Fujian Edition, 4) Mid-Southern Song Fujian Edition, 5) Yuan Fujian Edition, 6) Sichuan Edition (small character) in Northern and Southern Song, 7) Southern Song Jiangxi Edition, 8) Jin and Mongol Edition, 9) Yuan and Ming Zhao-Style Edition, 10) Jiajing Edition, 11) Zhe Edition of Yuan Dynasty. The periods covered by each edition category are shown in Table 1, with the number of scanned pages ranging from 2,329 to 27597 in each category. This edition classification will not affect the final model framework and can be modified at any time based on academic consensus.

Table 1. Data range statistics used in this study, including edition identification, year range, and corresponding data quantity.

	Edition Based on Fonts	Year Range	Book Copy	Volume	Page Count
1	Northern and Southern Song Zhe Edition	960-1279	20	308	15917
2	Southern Song Printing House Edition (in Zhe region)	1127-1279	13	47	2329
3	Early-Southern Song Fujian Edition	1127-1189	5	149	7481
4	Mid-Southern Song Fujian Edition	1190-1224	15	115	6215
5	Yuan Fujian Edition	1271-1368	9	70	5059
6	Sichuan Edition (small character) in Northern and Southern Song	960-1279	14	148	13815
7	Southern Song Jiangxi Edition	1127-1279	15	69	3487
8	Jin and Mongol Edition	1115-1368	30	103	4425
9	Yuan and Ming Zhao-Style Edition	1271-1644	15	43	6283
10	Jiajing Edition	1521-1566	15	-	3715
11	Zhe Edition of Yuan Dynasty	1127-1279	9	225	27597

Considering the long history of Chinese rare books and their vulnerability to environmental factors, such as storage conditions, issues like yellowing, damage, and stains are prevalent. After converting the scanned image data to grayscale, we applied binarization processing. This method compares the grayscale levels of pixels, where pixels with grayscale levels greater than or equal to a specified threshold are converted to black (i.e. foreground) pixels, while those below the threshold are converted to white (i.e. back-ground) pixels. The goal is to separate ink parts from irrelevant background components like noise and paper texture features, which are unrelated to the identification task. We employed the global Otsu's method [1] for threshold selection. By computing the frequency distribution of each grayscale level in the image, we scanned all possible thresholds and then calculated the inter-class variance of the two categories of black and white pixels for each possible threshold. The threshold that maximizes the inter-class variance is chosen as the final threshold, ensuring optimal foreground and background effects in the image. This method is efficient, does not require manual parameter setting, and can handle complex layout structures, as illustrated in Figure 1.

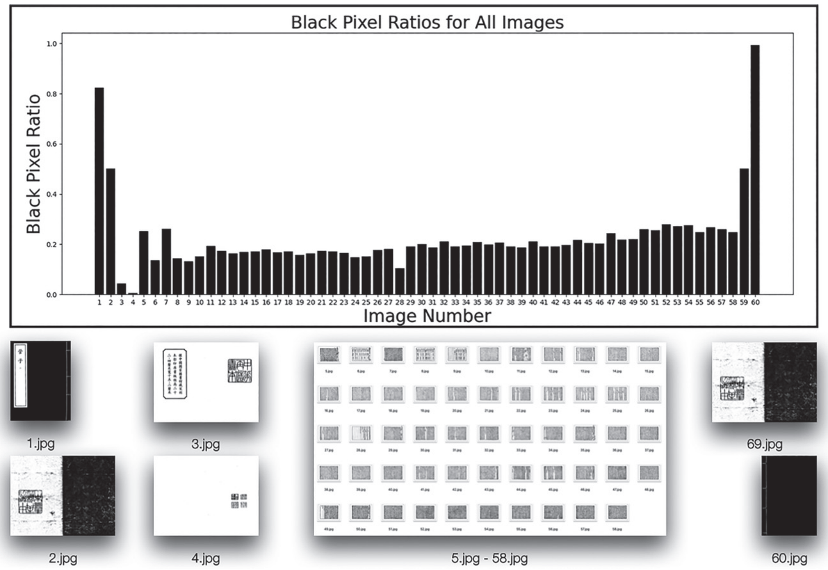

Figure 1. Histogram of the proportion of black pixels in the book page (top) and binarized dataset illustration (bottom). The scanned images used in the example are from the picture of this book: *Guanzi*, Song edition, held at the National Library of China, photo-reproduction in the *Zhonghua zaizao shanben*.

[1] Otsu Nobuyuki et al., “A Threshold Selection Method from Gray-Level Histograms,” *Automatica*, vol.11, 1975, pp.23–27.

Meanwhile, considering that the original dataset comprised entire book scans, which included content irrelevant to the edition identification such as covers, prefaces, and postscripts, we implemented heuristic rules to extract and remove these paratextual areas in scanned images. To analyze the characteristics of text and paratextual areas, we selected a representative rare book and conducted a statistical analysis of the proportion of black pixels in all 60 pages of binary images. As depicted in figure 1, for the main text area in the rare book, the proportion of black pixels in the entire page tends to be relatively consistent, ranging from 15% to 35%. The cover with fewer text contents resulted in a higher proportion of black pixels in the binarized images, while the prefaces and postscripts exhibited a higher proportion of white pixels, showing significant differences. Based on this observation, we established a threshold to filter images with a proportion of black pixels less than 15% or greater than 35% of the total page pixels, effectively eliminating paratextual content from the dataset. The number of pages filtered after removing paratext ranged from 2,073 to 21,821 pages across the eleven categories, representing approximately 87% of the original data. To streamline subsequent machine learning classification tasks, we randomly shuffled the preprocessed datasets for each classification, selected a sample size of 2,073 pages, and divided them into training and testing sets in an 8:2 ratio. Additionally, we encoded the classification labels using the one-hot encoding algorithm.

Model Construction

Most noise and non-text areas in the rare book documents have been effectively removed after the above steps. Subsequently, we proceed with the classification of the cleaned text areas. This paper proposes a deep learning method based on InceptionResNet-V2 for the classification of rare book editions. The specific implementation steps are as follows:

Step 1: Considering the input size limitations of the network architecture, we divided the text areas into fixed square proportions. The final image cropping size has been set to 448 × 448 after experimental validations.

Step 2: To further achieve a more accurate classification of rare book editions, this paper designs a classifier based on the InceptionResNet-V2 network. It combines the Inception and ResNet (Residual Network) architectures. The Inception module can extract features at multiple scales, while the residual connections help the network learn complex mapping relationships more effectively and prevent the gradient vanishing problem that can occur during the training of deep networks.

The training process of the InceptionResNet-V2 model employed a one-phase Adam (Adaptive moment estimation) optimizer [①] with a learning rate set to 0.001 and a learning rate decay factor set to 0.1. The cross-entropy loss function was chosen as the loss function. Training samples were obtained by cropping the fixed-size images after removing non-text areas from the original images.

Result and Discussion

The performance of text area detection was evaluated using the precision and recall metrics defined by the ICDAR 2005 Robust Reading Competition. The experimental environment was: the Windows operating system (Windows 11 Enterprise Edition), Intel Core i7-7700 processor, and NVIDIA GeForce V100 graphics card.

In this section, we compare the classification performance of the proposed method with commonly used models, VGG16 and ResNet50. We also discuss some hyperparameter settings in the experiments and verify the network's feature capture capability using visualized algorithms. Firstly, for text extraction, we used binarization and heuristic rules for preprocessing. Experiments show that the proposed method effectively separates text areas from complex backgrounds.

Subsequently, we compared the overall performance of the proposed method with the widely used VGG16 and ResNet50 networks on similar classification tasks. The training data for these three methods were the rare book images previously used to construct the cropped dataset. During training, all images were normalized to a width of 448 × 448 pixels, with the height proportionally padded.

By comparing the detection results of the three different methods, it can be

[①] Diederik P. Kingma, Jimmy Lei Ba, "Adam: A Method for Stochastic Optimization," Published as a conference paper at ICLR, SanDego, 2015.

observed that due to the limited number of annotated training samples, the final classification results were not very ideal. However, the proposed method showed a better ability to extract detailed character features, achieving the best accuracy and recall rates on the limited annotated data.

Table 2. Comparison of accuracy and recall performance in the rare book edition identification task among VGG16, ResNet50, and the method proposed in this paper.

Method	Accuracy	Recall
VGG16	79. 8	79. 8
ResNet50	86. 4	78. 2
This Paper	91. 3	83. 4

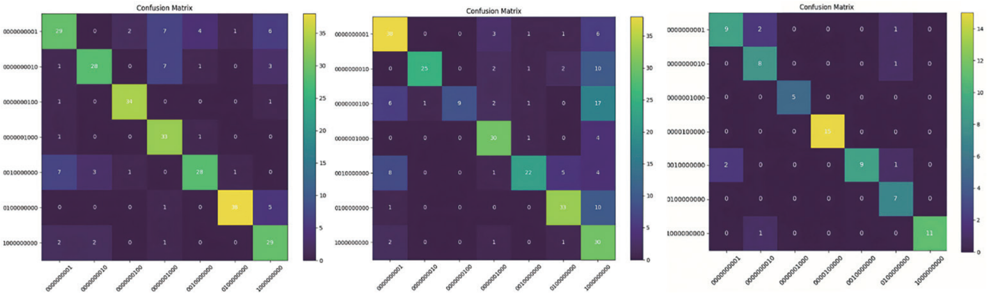

Figure 2. From left to right, the confusion matrix images for ResNet50, VGG, and the method used in this paper. The rows and columns represent the actual and predicted editions of rare books, respectively, with the cell values indicating the number of matches between the actual and predicted editions.

Figure 2 illustrates the pilot study phase, where we compared the classification performance of different network architectures using a smaller-scale dataset with 7 subcategories. The confusion among different editions is visualized as shown in the figure. Under the VGG network, only 25% of samples in edition 8 (i.e., one-hot encoding 0000000100) were correctly classified, while a significant 47% were misclassified as edition 1 (i.e., one-hot encoding 100000000). Additionally, for edition 1, the recall rate was merely 37%, with samples from all six other editions partially misclassified as edition 1. The number of misclassified samples in each edition accounted for 5%–21% of the total recalled samples for edition 1. A similar phenomenon was observed in the ResNet50 model, though with slight improvements. Under ResNet50, the recall rate for edition 1 increased to approximately 66%, and among the six other editions, samples from four were misclassified as edition 1, accounting for 2%–14% of the total recalled samples for edition 1. In contrast, the proposed InceptionResNet-V2 method exhibited superior performance. With this network, edition 1 achieved a classification accuracy of 92% and a recall rate of 100%.

For the final experimental phase, we prepared a more comprehensive dataset with 11 categories, as presented in Table 1. The InceptionResNet-V2 model ultimately achieved an accuracy of 91.3% and a recall rate of 83.4% in the 11-category classification task, significantly outperforming VGG and ResNet50 in feature representation capability. Detailed experimental results are provided in Table 2.

Conclusion

This paper presents a method for preprocessing and classifying scanned rare book images under complex noisy backgrounds. Initially, the original images are preprocessed using global binarization, followed by heuristic rule-based filtering to remove non-text areas. Finally, the rare book editions are identified using an approach based on InceptionResNet-V2. Experimental results demonstrate that the proposed method achieves higher accuracy and recall rates compared to traditional detection methods. Enhancing classification performance in more complex backgrounds, pre-annotating rare book characters, and training using deep learning methods will be the primary focus of future work.