Reading between the Lines: Image-Based Order Detection in OCR for Chinese Historical Documents

Hsing-Yuan Ma[1], **Hen-Hsen Huang**[2], **Chao-Lin Liu**[1]

[1]Department of Computer Science, National Chengchi University
[2]Institute of Information Science, Academia Sinica,
hsingyuanma@gmail.com, hhhuang@iis.sinica.edu.tw, chaolin@g.nccu.edu.tw

Abstract

Chinese historical documents, with their unique layouts and reading patterns, pose significant challenges for traditional Optical Character Recognition (OCR) systems. This paper introduces a tailored OCR system designed to address these complexities, particularly emphasizing the crucial aspect of Reading Order Detection(ROD). Our system operates through a threefold process: text detection using the Differential Binarization++ model, text recognition with the SVTR Net, and a novel ROD approach harnessing raw image features. This innovative method for ROD, inspired by human perception, utilizes visual cues present in raw images to deduce the inherent sequence of ancient texts. Preliminary results show promising reductions in page error rates. By preserving both content and context, our system contributes meaningfully to the accurate and contextual digitization of Chinese historical manuscripts.

Introduction

Chinese historical documents, with their intricate layouts and unique character formations, encapsulate a vast array of cultural and historical knowledge. Digitizing and transcribing these documents accurately requires specialized tools, particularly when deciphering the inherent reading order of such texts. Traditional OCR systems, while adept at text detection and recognition, often grapple with the challenge of deducing the authentic reading order of Chinese manuscripts(Clausner, Pletschacher, and Antonacopoulos 2013).

This paper introduces our OCR system designed specifically for Chinese historical documents, with a predominant focus on ROD. What sets our approach apart is its reliance on raw image features to approximate reading order acquisition, a method inspired by human perception. As we delve into the realm of visual processing, this work aims to contribute meaningfully to the digitization of these ancient texts, ensuring their legacy persists in the digital age.

Background

Chinese historical documents, enriched by millennia of cultural evolution, present unique layouts and calligraphic styles that make their transcription both intriguing and challenging. A central aspect of this transcription challenge lies in deducing the inherent reading order of these documents.

- OCR for Chinese Texts: Over the years, many OCR systems have been developed for Chinese documents, with a prime focus on text detection and recognition. While techniques ranging from pattern matching to deep learning have achieved notable success for modern Chinese scripts, historical manuscripts have always posed a distinct set of challenges.
- Complexity of Reading Order in Chinese Texts: Reading order in Chinese historical documents can differ significantly from their modern counterparts. Instead of a consistent left-to-right or top-to-bottom pattern, these ancient texts can exhibit varied sequences, influenced by their era, purpose, and the medium on which they were inscribed(Liu 2007; Wei 2004).
- Visual Cues and Reading Order: Reading order, especially in historical documents, is not just influenced by the text itself but by myriad visual cues. Elements such as boundary lines, the spacing between characters, clusters of text, and the relative location of layout elements all provide crucial hints. These visual cues are often more pronounced in the raw images of the documents, making them invaluable for deducing reading order.

In light of these challenges and the potential of raw images, our work endeavors to create an OCR system tailored for Chinese historical documents. Our primary focus lies in enhancing ROD by leveraging the untapped potential of raw image features, aiming to capture the authentic essence of these ancient texts.

System Overview

Our OCR system for Chinese historical documents is meticulously designed to efficiently and accurately transcribe and structure the content of ancient manuscripts. Comprising three primary modules, the system streamlines the transcription process from raw image to structured text, as shown in figure 1.

System Input

The system begins by accepting an image of a historical Chinese document. This image serves as the foundation upon

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

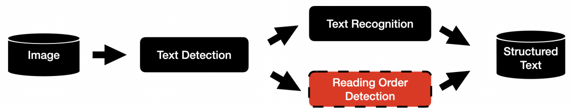

Figure 1: Our system pipeline

which subsequent processes are built.

Text Detection

Upon input, the first task is to identify and delineate the layout elements present in the document image. This detection of text bounding boxes is facilitated by the Differential Binarization++ model (Liao et al. 2022). This model, renowned for its accuracy, effectively pinpoints areas of interest, preparing the stage for subsequent recognition and reading order identification.

Text Recognition

With the layout elements (text bounding boxes) identified, the system employs the SVTR Net(Du et al. 2022) for text recognition. This model is adept at converting image-based Chinese characters into machine-readable text. The recognized text forms the basis for generating structured output.

Reading Order Detection

ROD is essentially a sorting challenge, where given a set of layout elements from a book page, the aim is to rearrange them in the correct reading order. The conventional pairwise learning-to-rank method uses a binary classifier function, learned from data, to determine the sequence of layout elements(Quiros and Vidal 2021). However, due to potential inconsistencies in transitive relations, deriving the optimal sequence often relies on less efficient algorithms based on a pairwise matrix.

In our approach to ROD,as shown in figure 2, we focus on visual cues present in document layouts, such as character size and spacing. We've developed a probability model indicating the sequence likelihood between layout elements on a page. This study introduces a multimodal method that incorporates both image data and spatial embeddings into a CNN. This process produces a pairwise probability matrix, which when decoded, reveals the reading order of elements. The integration of image data and spatial embeddings allows our model to factor in both the positioning and visual characteristics of layout elements. We use small version of MobilenetV3(Howard et al. 2019) as our CNN model.

For the final reading order determination, we construct a pairwise matrix and employ the First Decide then Decode (FDTD) algorithm(Quirós and Vidal 2022). It offers an advantage in accuracy and efficiency over traditional methods. The FDTD approach identifies relative placements of pairs in the matrix and utilizes the matrix's structure to deduce the overall reading sequence.

System Output

The culmination of this processing chain is twofold:

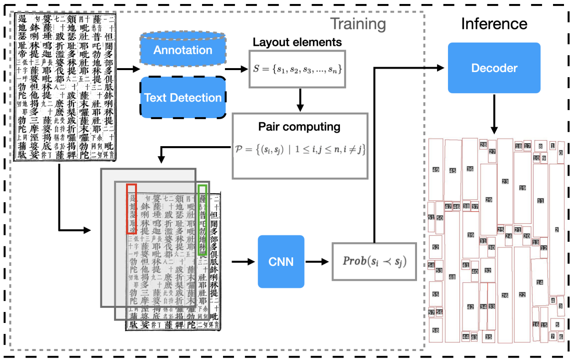

Figure 2: Proposed Reading Order Detection Model

- A structured text derived from the input document image, representing an accurate transcription of the original content.
- A visualized version of the original image, enriched with overlaid text bounding boxes, the transcribed text, and indicators highlighting the correct reading order.

Training Details

We employed the MTHv2 dataset for training our models. The MTHv2 dataset, as described in (Yang et al. 2018; Ma et al. 2020), is an amalgamation of the Tripitaka Koreana in Han (TKH) Dataset and the Multiple Tripitaka in Han (MTH) Dataset. It encompasses a total of 3,199 images.

Text Detection and Recognition

- Test ratio: The dataset was divided with a test ratio of 0.1, ensuring adequate representation for validation.
- Performance:
  - Text Detection: Achieved an F1 score of 0.95.
  - Text Recognition: Attained an accuracy of 0.83.

Reading Order Detection

- Test ratio: A different split ratio was utilized for the ROD model, with a test ratio of 0.3.
- Performance: Our ROD model exhibited a page error rate of 5% on test dataset, emphasizing its performance in determining the reading order.

Conclusion

This study introduced a specialized OCR system for Chinese historical documents, primarily emphasizing the ROD. Our system merges three vital components: text detection, text recognition, and ROD. Distinctively, the ROD model harnesses raw image cues, mirroring human reading patterns. The outcome was commendable: a 0.95 F1 score in text detection and 0.83 accuracy in text recognition. Furthermore, our ROD's innovative multimodal approach, integrating image data and spatial embedding, yielded a mere 5% of page error rate. This work not only enhances our understanding of these invaluable documents but also suggests exciting avenues for future research in visual processing.

References

Clausner, C.; Pletschacher, S.; and Antonacopoulos, A. 2013. The Significance of Reading Order in Document Recognition and Its Evaluation. *2013 12th International Conference on Document Analysis and Recognition*, 688–692.

Du, Y.; Chen, Z.; Jia, C.; Yin, X.; Zheng, T.; Li, C.; Du, Y.; and Jiang, Y.-G. 2022. SVTR: Scene Text Recognition with a Single Visual Model. arXiv:2205.00159.

Howard, A.; Sandler, M.; Chen, B.; Wang, W.; Chen, L.; Tan, M.; Chu, G.; Vasudevan, V.; Zhu, Y.; Pang, R.; Adam, H.; and Le, Q. 2019. Searching for MobileNetV3. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 1314–1324. Los Alamitos, CA, USA: IEEE Computer Society.

Liao, M.; Zou, Z.; Wan, Z.; Yao, C.; and Bai, X. 2022. Real-Time Scene Text Detection with Differentiable Binarization and Adaptive Scale Fusion. arXiv:2202.10304.

Liu, Z.-Y. 2007. *Understanding of Printed Ancient Book and Book Collectors*. studentbooktw. ISBN 9789571513546.

Ma, W.; Zhang, H.; Jin, L.; Wu, S.; Wang, J.; and Wang, Y. 2020. Joint Layout Analysis, Character Detection and Recognition for Historical Document Digitization. In *2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 31–36. Los Alamitos, CA, USA: IEEE Computer Society.

Quiros, L.; and Vidal, E. 2021. Learning to Sort Handwritten Text Lines in Reading Order through Estimated Binary Order Relations. In *2020 25th International Conference on Pattern Recognition (ICPR)*, 7661–7668.

Quirós, L.; and Vidal, E. 2022. Reading order detection on handwritten documents. *Neural Computation and Applications*, 34: 9593–9611.

Wei, L. 2004. *Simple Organization and Version Study of Ancient Books*. Macao: Macao Library & Information Management Association. ISBN 9993731056.

Yang, H.; Jin, L.; Huang, W.; Yang, Z.; Lai, S.; and Sun, J. 2018. Dense and Tight Detection of Chinese Characters in Historical Documents: Datasets and a Recognition Guided Detector. *IEEE Access*, 6: 30174–30183.