College of Engineering Chengannur
Department of Computer Engineering
M. Tech. Computer Science (Image Processing)
03CS7903 Seminar II
Abstract of Proposed Seminar Topic

Weakly Supervised Precise Segmentation For Historical Document Images

19/MCS/2019 CHN19CSIP01 ANJANA RAMACHANDRAN

September 8, 2020

Keywords: Weakly supervised learning ,Recognition-guided segmentation, Historical document images segmentation.

Abstract

With the passing of history, precious cultural heritage was left behind to tell ancient stories, especially those in the form of written documents. After years of storage, historical document collections encounter serious degradation via staining, tearing, ink seepage, etc. The problem of how to preserve this priceless culture heritage for the next generation has received intense interest from numerous researchers. Historical document digitization is one way to protect such valuable information on historical knowledge and literary arts.

Historical documents are digitized through photographing, followed by document segmentation, recognition, preservation, management, and research. Among all the above-mentioned stages, document segmentation is conducted as a first step and the over- all digitization performance of the system heavily depends on the segmentation quality. Here,the paper proposes a novel method for segmenting historical document characters for classification and recognition purpose there by decoding the contents in the documents.

Document segmentation consists of three principal stages: Document layout analysis, Text line segmentation and Character segmentation. Document layout analysis (i.e., page segmentation), is to separate a document image into regions of interest. Text line segmentation is the division of document images or paragraph images into individual text line images for subsequent character segmentation. Character segmentation is an operation that seeks to decompose text images into sub-images of individual symbols.

Here a recognition-guided segmentation problem from Bayesian decision theory perspective is formulated and different types of algorithms is provided to search for the segmentation paths from coarse to finer. The weakly supervised precise segmentation system for historical document images mainly consists of four stages, including Preprocessing, Boundary Box Segmentation (BBS), Incremental Weakly Supervised Learning and Recognition-guided Attention Boundary Box Segmentation (Rg-ABBS). The character segmentation problem is formulated from the perspective of Bayesian decision theory. The problem is how to locate every character inside the image by providing each character a bounding box.

In the pre-processing stage, vertical projection is applied to slice the page image into line images, so that line image-annotation pair is obtained. At first the image page is vertically projected onto the x -axis to derive the projection profile, then following Text line segmentation based on morphology and histogram projection, line images are extracted. Next, boundary detection is adopted to over-segment the line images into strokes or radicals.

Through Boundary Box Segmentation(BBS) algorithm, bounding boxes are efficiently merged into characters for subsequent research. The recognition result of BBS is not as precise as those of recognition-guided, it performs fast segmentation without consulting character recognizer, and constitute a key component of Rg-ABBS. Another segmentation algorithm called Recognition-guided boundary box segmentation (Rg-BBS) is proposed to incorporate text line annotation as well as recognition score of character recog-

nizer to facilitate character segmentation. But this algorithm waste much time on consulting the character recognizer.

In order to utilize the advantages of BBS and Rg-BBS while discarding their drawbacks, another algorithm called Recognition-guided Attention Boundary Box Segmentation (Rg-ABBS) is developed to help system focus only on the confusing parts. The idea behind Rg-ABBS is to integrate character recognition precisely on the 'attention' area of the text line image, where mis-segmentation problems usually occurs. It is observed that the proposed Rg-ABBS successfully integrates the recognition information of character and line-level annotation to facilitate the segmentation result while consuming much less time and effort than Rg-ABBS.

Furthermore a novel judgment gate mechanism is proposed to train a high-performance character recognizer in an incremental weakly supervised learning manner ,so that it can provide reliable character recognition score to improve character segmentation results.

The proposed Rg-ABBS algorithm significantly reduces time consumption by performing recognition-guided segmentation only on 'attention' area and achieves comparable performance in comparison to performing recognition-guided segmentation on the whole text line image as in Rg-BBS. The system is comprehensive, including line and character segmentation for historical images, that provides sufficiently precise bounding boxes for characters. Experiments show that the proposed Rg-ABBS system significantly outperforms traditional segmentation methods as well as deep-learning-based segmentation and detection methods under strict intersection-over-union (IoU) requirements.

References

[1] C.-L. Liu Q.-F. Wang , F. Yin. Handwritten chinese text recognition by inte- grating multiple contexts. *IEEE Trans. Pattern Anal. Mach. Intell*, 34(8):1469–1481, May 2012.

[2] T.I. Ren G.D. Cavalcanti R.P. dos Santos, G.S. Clemente. Text line segmentation based on morphology and histogram projection. In *Proceedings of the Tenth International Conference on Document Analysis and Recognition, IEEE*, page 651–655, Feb 2009,.

[3] E. Lecolinet R.G. Casey. A survey of methods and strategies in character segmentation. *IEEE Trans. Pattern Anal. Mach. Intell*, 18(7):690–706, May 1996.