Chinese Historical Document Image Analysis and Recognition Database

Rajkumar Saini[1], Derek Dobson[2], Jon Morrey[2], Foteini Simistira Liwicki[1], and Marcus Liwicki[1]

[1]Machine Learning Group, Luleå University of Technology, Sweden,
[2]FamilySearch, USA
rajkumar.saini@ltu.se, derek@familysearch.org, jmorrey@familysearch.org, foteini.liwicki@ltu.se, marcus.liwicki@ltu.se

Abstract. Historical documents are a heritage to the society. They give the view of there time for the topics they were written for. The scripts in such documents may differ from modern writing. Moreover, they are often in poor physical condition due to scratches, stains, missing text, etc. However, scholars in the humanities often want to access and extract valuable information from them. With recent improvements in document analysis and recognition, this is often done with the help of computerized processing methods. To train such methods, large databases of manuscripts with labels are needed. In this paper, we present a large historical database of Chinese records for the research. We aim on historical documents to develop robust systems for historical document analysis. In this direction, there will be a competition named Historical Document Reading Challenge on Large Chinese Structured Family Records, in short ICDAR 2019 HDRC Chinese on this database. The objective behind this competition is to boost the research on historical document analysis. The focus of the competition is to recognize and analyze the layout, and finally detect and recognize the textlines and characters of the documents in this database. The database contains more than $10'000$ pages written in Chinese traditional *Han* script.

1 Historical DIA and Importance of the Database

In recent years, there has been an exemplary growth in understanding the cultural heritage like historical documents aiming their Optical Character Recognition (OCR) to preserve their knowledge for ages [3]. For this, it is important to recognize the text and layout structures present in these documents. Historical documents differ from the ordinary documents due to the presence of different artifacts in them [5]. The issues such as poor conditions of the documents (see Fig. 2), texture, noise and degradation, large variability of page layout, page skew, complicated layout, random alignment, specific fonts, presence of embellishments, variations in spacing between the characters, words, lines, paragraphs and margins, overlapping object boundaries, superimposition of information layers, etc bring complexity in analyzing them [5]. Thus, a huge number of doc-

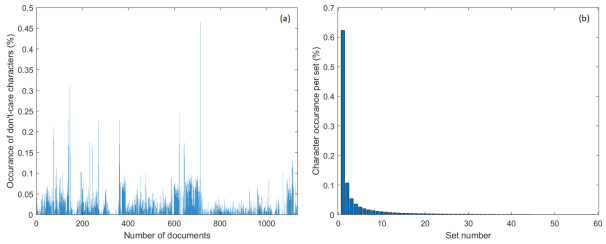

Fig. 1: Distribution of characters in Test data. (a) Occurrence of do-not-care characters per document image (b) Occurrence of characters sets.

uments are required to develop efficient models to recognize the text, and the layout present in the historical documents.

Aiming that, we present a large database of historical Chinese documents for the research. The database consists of more than 10′000 document images collected from various Chinese books mainly written in traditional *Han* script. The database was kindly provided by FamilySearch[1]. There is competition named ICDAR 2019 HDRC Chinese on this database in ICDAR international conference in the month of September this year. The competition is to analyze and recognize the historical Chinese records in the database. The database is one of the biggest databases available for Chinese language analysis and recognition.

2 Database Description

ICDAR 2019 HDRC Chinese database is a collection of Chinese manuscripts that have been chosen regarding the complexity of their layout in semantic structure and font. All manuscripts are annotated using Aletheia[2], an advanced system for accurate and yet cost-effective ground truthing of large amounts of documents. The annotation of the manuscripts are available in PAGE XML format, a sophisticated XML schema which is component of the PAGE (Page Analysis and Ground truth Elements) Format Framework [6]. The database consists of the collections shown in Table 1. The characters having occurrence less than 10 are considered to be do-not-care. Fig.1(a) shows the occurrence of do-not-care characters for each document image in the Test data. It can be noticed that the only few documents have greater than 25% of total characters within them as do-not-care. Fig.1(b) shows the distribution of the characters sets. Sets 1 consists of 79 highest appearing characters in the Test data. Similarly, set 2 is the set of another 79 characters having highest occurrence after set 1, and so on. It can be noticed from the figure that the distribution of characters occurrence is not uniform. Thus, the database is very rich and highly imbalance.

Fig. 2 shows a sample Chinese document where individual textlines are shown in green bounding boxes.

[1] `https://www.familysearch.org/en/`

Table 1: Statistics of the ICDAR 2019 HDRC Chinese Database

Total number of documents in the database	12′850
Number of documents in Training data	11′715
Number of documents in Test data	1′135
Total number of unique characters in the database	8′461
Number of unique characters in Training data	8′319
Occurrence of characters (total) in Training data	25′36′865
Number of characters in Training data having at least 10 instances	4′290
Number of unique characters in Test data	4′661
Occurrence of characters (total) in Test data	2′46′287

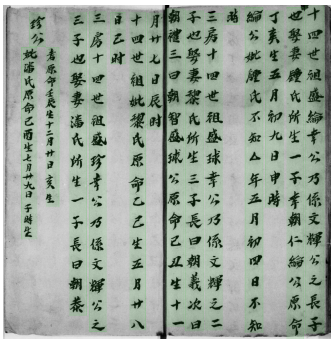

Fig. 2: A sample Chinese document. Textlines are shown in green bounding boxes. Best viewed in color.

3 Task Description

There are 3 different tasks in the competition:

- **Task 1** Handwritten Character Recognition on extracted textlines. Given an image of a textline, the objective is to recognize the text in the image.
- **Task 2** Layout Analysis on structured historical document images. Given an image of the document, the objective is to detect the text-lines in it.
- **Task 3** Complete, integrated textline detection and recognition on a large database. Given an image of the document, the objective is the combination of Task1, and Task2 i.e. detecting the textlines and recognizing them.

4 Evaluation Tools and Metrics

The evaluation of **Task 1** is be based on the edit distance between two text strings as the minimum number of operations (insertion, deletion, and substitution) needed to transform one into the other. More details could be found at: https://web.stanford.edu/class/cs124/lec/med.pdf .

The evaluation of **Task 2** at pixel level is based on the Intersection over Union (IU) as proposed in [4] as ranking metric. The IU, also known as the Jaccard

Index, is defined as:

$$IU = \frac{TP}{TP + FP + FN} \tag{1}$$

where TP denotes the True Positives, FP the False Positives and FN the False Negatives.

For each page, the IU is computed class-wise (background, text, do-not-care care regions) and then averaged. The final evaluation of a system is then obtained by averaging the IU of all pages of the database. The tool is freely available on GitHub[2]. More information about this evaluation tool can be found in [1].

The evaluation of **Task 3** is based on the graph based edit distance operations such as the number of nodes and edges inserted, deleted, number of nodes substituted, error ratio. More details could be found at: https://arxiv.org/abs/1903.03341

The aim of our research is to facilitate historical documents analysis and associated issues as discussed in the Section 1.

Important note. The winner will get an award price of $1′000 USD provided by FamilySearch.

5 Acknowledgements

We thank DIVA Group[3] of University of Fribourg, Switzerland, and especially Michele Alberti, for providing us the open source evaluation tool for Task 2.

References

1. Alberti, M., Bouillon, M., Liwicki, M., Ingold, R.: Open Evaluation Tool for Layout Analysis of Document Images. International Workshop on Open Services and Tools for Document Analysis (2017)
2. Clausner, C., Pletschacher, S., Antonacopoulos, A.: Aletheia-an advanced document layout and text ground-truthing system for production environments. In: Document Analysis and Recognition (ICDAR), 2011 International Conference on. pp. 48–52. IEEE (2011)
3. Jenckel, M., Bukhari, S.S., Dengel, A.: anyocr: A sequence learning based ocr system for unlabeled historical documents. In: 2016 23rd International Conference on Pattern Recognition (ICPR). pp. 4035–4040. IEEE (2016)
4. Marti, U.V., Bunke, H.: Using a statistical language model to improve the performance of an hmm-based cursive handwriting recognition system. International journal of Pattern Recognition and Artificial intelligence 15(01), 65–90 (2001)
5. Mehri, M.: Historical document image analysis: a structural approach based on texture. Ph.D. thesis, Université de La Rochelle (2015)
6. Pletschacher, S., Antonacopoulos, A.: The page (page analysis and ground-truth elements) format framework. In: Pattern Recognition (ICPR), 2010 20th International Conference on. pp. 257–260. IEEE (2010)

[2] `https://github.com/DIVA-DIA/DIVA_Layout_Analysis_Evaluator`
[3] `https://diuf.unifr.ch/main/diva/`