ICDAR 2019 Historical Document Reading Challenge on Large Structured Chinese Family Records

Foteini Simistira Liwicki∗, Derek Dobson †, Jon Morrey †, Rajkumar Saini∗, Marcus Liwicki∗,
∗Machine Learning Group, Luleå University of Technology, Sweden, {firstname}.{lastname}@ltu.se
†FamilySearch, USA, {firstname}.{lastname}@familysearch.org

Abstract—We propose a Historical Document Reading Challenge on Large Chinese Structured Family Records, in short ICDAR 2019 HDRC CHINESE. The objective of the proposed competition is to recognize and analyze the layout, and finally detect and recognize the textlines and characters of the large historical document collection containing more than 20 000 pages kindly provided by FamilySearch.

I. COMPETITION PROTOCOL AND DATA

We invite all researchers and developers in the field of document layout analysis to register and participate in the new Historical Document Reading Challenge on Large Chinese Structured Family Records.

We propose 3 different tasks for this competition:

• Task 1 Handwritten Character Recognition on extracted textlines
• Task 2 Layout Analysis on structured historical document images
• Task 3 Complete, integrated textline detection and recognition on a large dataset

A. Dataset

The dataset is provided by FamilySearch[1] and consists of the following collections:

• The test set consists of in total 1.757 images selected from 12 separate books.
• The training set consists of in total 19.360 images selected from another set of 37 separate books.

FamilySearch-DB 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[1], 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 [2].

1https://www.familysearch.org/

B. Task Description

1) Handwritten Character Recognition on Extracted Textlines:

The scope of this competition is to recognize (OCR) given extracted textlines and, if possible, to find the segmentation points of the characters. The advantage of the character competition is that we would be able to generate synthetic historical images, once we have the characters segmented and recognized. Training data will be also available in PAGE-XML format.

We will have at least 100 different characters to be recognized, having at least 20 samples, each. The distribution of characters will be according to a typical distribution, so there are actually some characters having more than one thousand instances and thousands of characters having only a few instances. We plan to map less frequent characters to the class label unknown.

2) Layout Analysis:

The scope of this competition is to segment the page in different classes by assigning a different pixel value for each class:

RGB(0,0,1)=#000001: background/non text
RGB(128,0,4)=#800004: don’t care regions/boundary regions
RGB(0,0,8)=#000008: foreground/text

The training data will be available as pixel labeled images.

To avoid unfair penalties for the boundary regions, we add a value for boundary pixels: RGB(128,0,4)=#800004. Mislabeling between the foreground and background in the boundary region will not be penalized in the final evaluation (see Section [1]).

3) Textline recognition:

The scope of this competition is to detect and recognize (OCR) a given textline image. The training data will be available also in PAGE-XML format. The PAGE-XML file will contain the information of the textlines’ location and their corresponding text.

C. Submission Types

We allow for four different submission formats, either an executable file (or a bash script), a virtual machine, or even
a docker image (preferably). The following instructions were given:

- **Executable**:
  - All dependencies should be in the same (sub)directory;
  - Code compiled against Ubuntu-14.04 or 16.04 (64bit) is preferred;
  - Windows binaries are also possible;
  - Provide a Link for downloading the specific zip file.

- **VirtualBox Image**:
  - Provide a VirtualBox-Image as download link;
  - Provide instructions how the method can be executed inside the VirtualBox.

- **Docker Image**:
  - Provide the reference image name as hosted on docker hub (see https://hub.docker.com);
  - Provide instructions how the method inside the docker image can be executed.

II. Evaluation

A. Task 1: Handwritten Character Recognition on extracted textlines

The evaluation of Task 1 will be based on typical textline recognition accuracy (Hits-Insertions)/numberOfCharacters.

B. Task 2: Layout Analysis on structured historical document images

The evaluation of Task 2 will be similar as in our previous competition. All evaluation tools are freely available as open source on GitHub and possibly RESTful API web services on DIVAServices. Providing the evaluation methods on DIVAServices allows users to run them using our infrastructure and do not need to install anything locally. More information about executing methods on DIVAServices can be found in [3].

The evaluation of the layout analysis 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:

\[
IU = \frac{TP}{TP + FP + FN}
\]  

(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, don’t care regions) and then averaged. The final evaluation of a system is then obtained by averaging the IU of all pages of the dataset.

In order to provide the user a more exhaustive evaluation of prediction quality, the tool outputs several other standard metrics, including F1-score, precision, and recall — for each class and averaged over the classes. Additionally, a human-friendly visualization of the results is provided in form of a output image obtained by overlapping the evaluated prediction with the original image. This is useful to get a quick estimation of the results and to detect the area of improvement for the evaluated method. More information about this evaluation tool can be found in [5].

C. Task 3: Complete, integrated textline detection and recognition on a large dataset

For the evaluation we will compute several metrics which are useful for measuring the OCR-accuracy combined with the text line positions. The most important one is the following: If all characters are correctly found and the grouping into text lines is as it is logically performed by human experts (depends on the layout structure), the score would be perfect. For every deviation from the perfect score, a penalty (in terms of graph edit distance) will be taken into account. The winner of this competition will be awarded with some award money provided by FamilySearch.

REFERENCES

[1] C. Clausner, S. Pletschacher, and A. Antonacopoulos, “Aletheia—an advanced document layout and text ground-truthing system for production environments,” in Document Analysis and Recognition (ICDAR), 2011 International Conference on. IEEE, 2011, pp. 48–52.

[2] S. Pletschacher and A. Antonacopoulos, “The page (page analysis and ground-truth elements) format framework,” in Pattern Recognition (ICPR), 2010 20th International Conference on. IEEE, 2010, pp. 257–260.

[3] M. Würsch, I. R. Simistira, Foteini, and M. Liwicki, “Turning Document Image Analysis Methods into Web Services – An Example Using OCRopus,” International Workshop on Open Services and Tools for Document Analysis, 2017.

[4] U.-V. Marti and H. Bunke, “Using a statistical language model to improve the performance of an hnn-based cursive handwriting recognition system,” International journal of Pattern Recognition and Artificial intelligence, vol. 15, no. 01, pp. 65–90, 2001.

[5] M. Alberti, M. Bouillon, M. Liwicki, and R. Ingold, “Open Evaluation Tool for Layout Analysis of Document Images,” International Workshop on Open Services and Tools for Document Analysis, 2017.