Meaningless Text OCR Model for Medieval Scripts

Adnan Ul-Hasan*
University of Kaiserslautern
adnan@cs.uni-kl.de

Syed Saqib Bukhari*
German Research Center for Artificial Intelligence
saqib.bukhari@dfki.de

Andreas Dengel
University of Kaiserslautern
and
German Research Center for Artificial Intelligence
andreas.dengel@dfki.de

Area: Cutting edge techniques

Abstract: Availability of large amount of ground-truth data for training an Optical Character Recognition (OCR) engine is extremely critical. Training data is usually produced by manually transcribing thousands of document images. In order to augment the limited training data, synthetic training data is also used, where training data is produced by rendering text into images in suitable fonts and styles. The most important part in synthetic training data is the corresponding real world text. If real world text data is unavailable, which could be a case in historical manuscripts, generating synthetic training data is not possible. In this paper, this problem has been addressed for the case of historical manuscripts whose vocabulary and sentence structure is neither available in text form not it is similar to any existing (contemporary) scripts. For such a case, we have introduced a novel meaningless text OCR model, where meaningless words of variable sizes are generated by permuting characters. Meaningless text lines are subsequently produced by randomly choosing these meaningless words. Testing of the meaningless text-line recognizer on real text-lines show good performance.

The rest of the paper answers the following questions in sequence: which types of historical documents are we dealing here?, why a textline-based recognizer is preferable over character-based recognizer?, what is the traditional way of training textline-based recognizers?, what novel technique we are presenting to overcome the limitations of traditional training procedure?, and what initial results we have achieved?

Which types of historical documents are we dealing here?
The “Narrenschiff” is a medieval 15th century German novel. Its first edition was printed in German language in 1494 in Basel and gained a lot of popularity at that time. Afterwards a lot of copies were spread with numerous translations and variations all over the Europe. Almost every edition can be characterized by the use of historical fonts and vocabulary. We are digitizing these documents under the German government funded project, Kalimachos. Some sample document images from this novel are shown in Figure 1.

Why a textline-based recognizer is preferable over character-based recognizer?
Textline based recognizers, in contrast to character based recognizers, produce better recognition accuracies without using any language modeling or any other post-processing techniques [6]. Because, such line based recognizers trained characters with their context. Albeit simple to use, LSTM-based recognizers have shown excellent OCR results for many scripts [6, 7, 2, 4, 5].

What is the traditional way of training textline-based recognizers?
For the development of a text-line based OCR model ground-truth data, that is a set of text-line images and the corresponding text lines, plays an important role. Traditionally, the following paradigms are used for training an LSTM-based OCR model for a particular script: (i) From the scanned document images, extract the text lines along with their ground-truth information and use them for the training of an LSTM network., (ii) If text-lines obtained from above process are not sufficient, then generate synthetic text lines using the available text in that language. However, in case of historical script, both of these traditional paradigms are not applicable, first because a lot of time is required for manual transcription and second because of the non availability of digitally created text.

What novel technique we are presenting to overcome the limitations of traditional training way?
For the training of LSTM networks for non-existing scripts, we have generated meaningless text lines. The process of generating such meaningless data for training line-based recognizer is described in the following. Firstly, we randomly generate a word corpus consisting of all possible permutations of all characters for several word lengths. We referred the set of these words as a bag of meaningless words. For Latin “Narrenschif” novel, there are around 84 characters (small, capital, punctuation marks, digits). The permutations of all characters with...
different word lengths (let say 1 to 8) produce a huge amount of meaningless words, and for limited memory and processing time resources, it is difficult to include all of these words. For the proof of concept, in this paper we have limited ourselves to small length words (i.e. permutations of only small alphabets of 3 and 4 word lengths). Then, we generated the text-line training data by using these bag of meaningless words. After that, we rendered text lines to develop a training database. Some sample rendered text line images from our meaningless training database are shown in Figure 2. Finally we trained a LSTM recognizer using the meaningless training data.

What initial results we have achieved?
To evaluate the trained LSTM model over meaningless training data, we generated 300 textline images from real transcriptions consisting of 3 and 4 length meaningful/real words. For this purpose, we first prepare a word-list from all 3 and 4 length words and then combine them into text lines of length up to 10 words per text-line. The results are shown in Table 1 and Figure 3.

<table>
<thead>
<tr>
<th>Table 1: The Performance Evaluation (Character Error Rate-CER) on the meaningful/real test data.</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Text-lines</td>
</tr>
<tr>
<td>------------</td>
</tr>
<tr>
<td>300</td>
</tr>
</tbody>
</table>

Conclusion:
In this paper, firstly we have introduced a novel process of automatically generating synthetic training data for non-existing historical scripts, i.e. meaningless text. Then, we have showed the performance of meaningless trained model on real-world test samples. Even though we have trained a meaningless model on only a limited lengths of words. The initial results on meaningful/real text data are promising.

References: