Preprocessing for Offline Handwriting Recognition

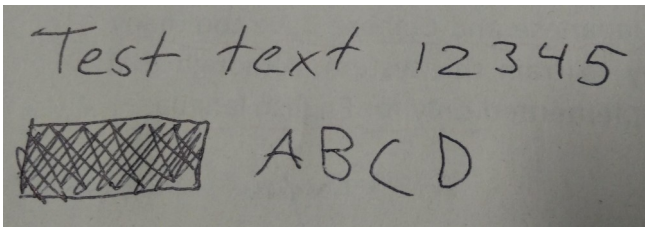
Pitkänen Perttu
Academic advisors: Kawamata Masayuki, Abe Masahide
Kawamata/Abe laboratory
Department of Electronic Engineering

Introduction

Optical Character Recognition (OCR) is the process of analyzing a picture of text and recognizing and extracting the characters to digital from. This process can be applied to for example document digitalization systems. Typical OCR systems consist of three phases: Preprocessing, feature extraction and classification. This phase of research concentrates on preprocessing. For quick development the implementation is done with MATLAB programming language and tools including image processing toolbox.

In preprocessing stage the quality of image is enhanced and the area of text is located. The feature extraction stage captures the distinctive characteristics of the digitized characters for recognition. Lastly in during the classification stage the feature vectors are processed to identify the characters and words.

Methods and implementation



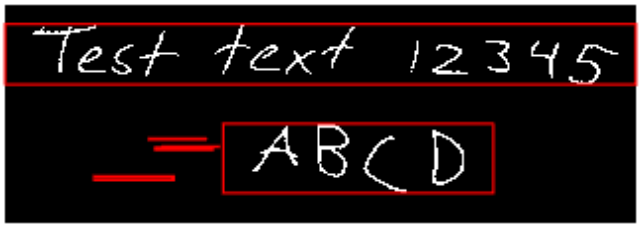
Original image with two rows of text and one irrelevant object

Firstly after the image is acquired as much as possible noise is to be removed. For this purpose adaptive Wiener filter is used. The adaptive filter can recognize the amount of variance and adjust the smoothing according to it.

After noise removal the image must be binarized. This is one of the most crucial parts of the preprocessing as information can be lost if it's not done properly. For this purpose the Sauvola binarization algorithm was chosen as it was designed for document binarization purposes.

For detecting text regions from the image the stroke width variation is observed. The image objects having only a little amount of variance in stroke width can be concidered text region compared to high variance regions such as photographs or drawings.

Afterwards a bounding box is calculated for the remaining areas. To detect vertical tex rows the bounding box is expanded horizontally and overlapping boxes are combined into bigger boxes. The resulting area is considered to be a text row.



Detected text rows (some noise remains)

The individual words should be extracted after the row is detected. This can be done by observing the horizontal projection histogram of the row and finding the areas with no white pixels in them. The row image can be split if the space exeeds pre defined threshold. At this point the space analysis hasn't yet been implemented.

Method evaluation and results

The aforementioned methods for preprocessing were chosen after several tests with different methods. Some additional steps were tested and discarded as they resulted in worse performance. For example histogram equalization was considered. After histogram equalization the noise caused by paper texture was increased and it had negative effect on the process as a whole.

All of the implementations of these methods need some pre-defined parameters to function. These parameters have to be chosen for each case individually which makes the system work only on one case at time. Case sensitive parameters include Wiener filter size, Sauvola neighbourhood size, Sauvola threshold, stroke width threshold, and the box expansion amounts.

Conclusions and future work

Several preprocessing methods were considered and tested. Many of them proved to be functional considering the process of handwriting recognition. For future work remains to make the system less dependent on the constant parameters and to improve the noise reduction even more. For fully functional handwriting recognition system it is important to implement the feature extraction and classification phases.