Preprocessing and Layout Analysis for Offline Handwriting recognition

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

Introduction

Handwriting recognition is a process to extract the textual information from image containing handwritten characters, into computer readable form. This is done by applying various image processing and classification methods to the input image. In this research experients are made with varying methods regarding preprocessing and layout analysis phases. 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 used with machine learning to identify the words.

Methods

Firstly 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 [1].

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 [2].

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 considered text region compared to high variance objects such as photographs or drawings. The objects are discarded if they have larger variance than pre-defined threshold [3].

Afterwards the process proceeds to layout analysis phase and in this case line detection process. The implemented method was proposed by Louloudis et.al. and it provides a sophisticated approach to detect handwritten text lines even if they overlap each other. This is done with so called block based Hough transform mapping [4] .

Hough transform is used to find lines in cartesian space using data points as input. In this case the data points are multiple block centroids aquired by splitting the component into smaller sections. The Hough transform results in accumulator array that shows which θ and ρ values realize most lines. When a line is found all components assigned to it are removed from the accumulator array and the process is continued until no new lines can be found. Afterwards the smallest objects which represent accents and punctuation marks are assigned to nearest line and the largest objects are split into smaller objects if they overlap multiple lines.

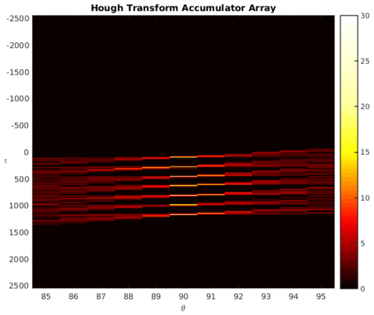

Figure 1: Hough transform results in an accumulator array. Seven lines can be detected with this accumulator array.

Evaluation

Several tests were conducted in order to gain insight on which methods work best with which parameter values and what kind of effect the changes have to the output. The methods were tested with 100 random images from IAM handwriting database containing different English language texts by different writers [5].

All tests were done using MATLAB R2016a with Intel(R) Core(TM) i5-2400 CPU 3.10GHz. Tests were conducted for the preprocessing methods Wiener filtering and Sauvola binarization algorithm. For layout analysis phase the stroke width variance threshold and several constant parameters for Hough transform mapping were evaluated. The tests used the number of detected lines as the metric to measure the accuracy of the system compared to the true amount of lines.

Conclusions

Aforementioned parameters were tested and their effect was studied. The binarization parameters had a significant effect on the recognition process. However the Hough transform mapping is only dependant on the paramters that affected the initial line detection from the accumulator array and the other parameters didn't have noticeable effect on the output. When best parameters were applied the system gained around 97% accuracy regarding text line amounts. Some constraints regarding the line detection were found, such as the method can detect lines from single column text.

As for future work the word detection should be implemented. At this point the system is only capable of recognizing individual lines of handwritten text. The feature extraction and classification phases should also be implemented for the whole recognition process to be complete. This would allow more accurate evaluation of the whole system.

References

[1] The MathWorks, "Noise Removal." `http://www.mathworks.com/help/images/noise-removal.html#buh9ylp-72`.

[2] J. Sauvola and M. Pietikäinen, "Adaptive document image binarization," *Pattern Recognition*, vol. 33, no. 2, pp. 225–236, 2000.

[3] The MathWorks, "Automatically Detect and Recognize Text in Natural Images." `http://www.mathworks.com/help/vision/examples/automatically-detect-and-recognize-text-in-natural-images.html?s_tid=gn_loc_drop`.

[4] P. H. Louloudis, Gatos, "Text line and word segmentation of handwritten documents," *Pattern Recognition 42*, 2008.

[5] U. Marti and H. Bunke, "The iam-database: An english sentence database for off-line handwriting recognition.," *Int. Journal on Document Analysis and Recognition, Volume 5, pages 39 - 46*, 2002.