**Project Plans Report**

|  |  |
| --- | --- |
| **Goal** | **Deadline** |
| Find a suitable dataset  (for individual characters and words) | 1/28/2018 |
|  |  |
| Run Alexnet on the dataset to get preliminary results and set a benchmark | 02/02/2018 |
| Have a morphology function for testing data sets | 02/02/2018 |
| Separate individual letters within words | 02/02/2018 |
| Identify individual handwritten characters | 02/02/2018 |
|  |  |
| Identify groups of characters as words | 02/09/2018 |
| Output identified, handwritten words as typeset words | 02/09/2018 |
| Convert a series of handwritten words (e.g. sentences, paragraphs, etc.) into a typeset sentence/paragraph. | 02/09/2018 |
|  |  |
| Finish report and prepare for the presentation | 02/11/2018 |

Possible Datasets

Word Image Data Set by [Center for Intelligent Information Retrieval](http://ciir.cs.umass.edu/downloads/old/data_sets.html)

[RETAS OCR Evaluation Dataset](http://ciir.cs.umass.edu/downloads/ocr-evaluation/) Although, this is not directly downloadable and you have to contact the authors for getting access for the dataset (details on the page)

[OCR dataset](http://ai.stanford.edu/~btaskar/ocr/) by Stanford

[IAM Handwriting Database](http://www.iam.unibe.ch/fki/databases/iam-handwriting-database) (you would need to register first)

[Dataset](http://www.ee.surrey.ac.uk/CVSSP/demos/chars74k/) for single letters

The minimum goal of this project is to recognize groups of identified characters as words. The reasonable goal is recognizing a sentence. The stretch goal is to recognize a whole paragraph.

Three different types of dataset have been found, which are: single letter training set, simple words validation set, and more words test set. An advanced dataset has also been found, which contains an entire paragraph for use in the stretch goal.

Before feeding the test set into the net, a few simple morphology methods will be applied on the words to clear the noise and isolate each letter. After this preprocessing, each letter in the word will be separated by using the MATLAB regionprops function, and the bounding box of each letter will be found. Figure 1 shows the basic script (thus far) that will be used to isolate the handwritten characters. Figure 2 shows the progression of the word images from using the script.

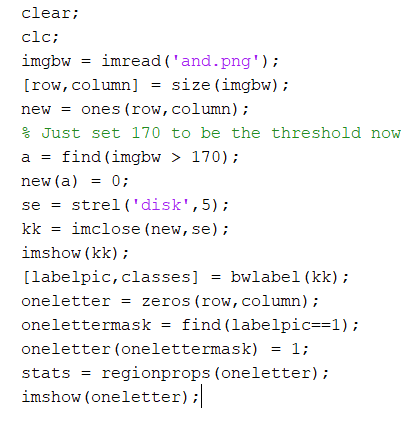


Figure : Basic MATLAB script to isolate letters

https://lh6.googleusercontent.com/02rXfEqBoKxkk6nbJE6mlsRzYkn7DptxH8NT8vig__-V8kk-E4KyYCxXv3BHoWOVtGp79Apw-Cgx9VHwAyGJtSTdUcFlZrewbRKUdsIIhiE8VI7sbH7F9O6VbeRt_qkiqcwg3Ghk





Figure : Progression from original image to isolated letter

The letters “i” and “j” have small dots that are separated from the main part of the letter. When preprocessing images to be fed to the network the small dot will ignored to keep the processing simple.

In the training section of the project, we will feed 62 different classes of character (26 letters, lowercase and capital, and 0-9) to the network. AlexNet will be used as a preliminary classifier. When the project progresses more a more complex pretrained network may be used.

Figure 3 shows the expected structure for our OCR project, and Figure 4 shows the expected completion dates for each section of the project.

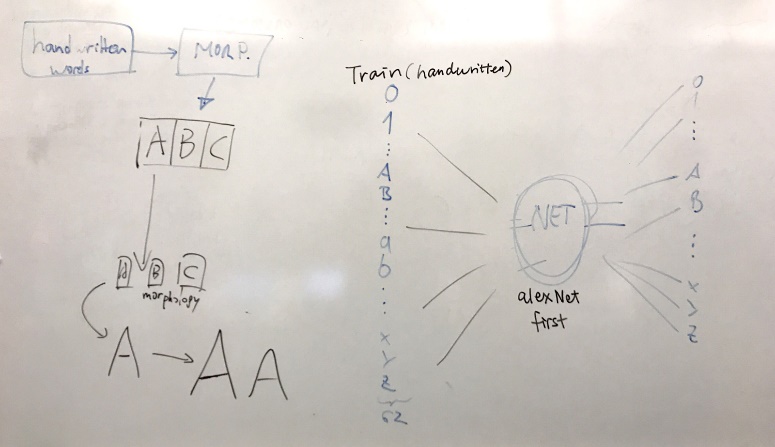


Figure : Expected OCR identification structure

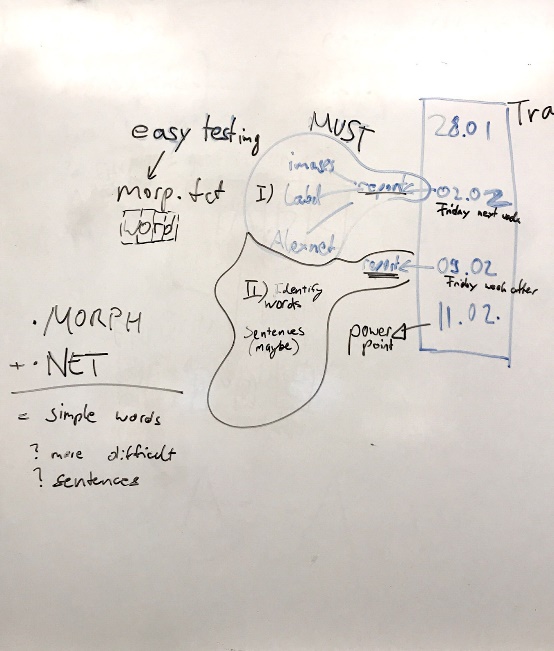


Figure : Expected timeline