OCR Handwriting Project Outline

Matthew Mulhall

matthew.l.mulhall@uconn.edu

May 28, 2019

1 May 14, 2019

1.1 Summary of Design Decisions

The project will follow an abstraction based design: letters, words, lines, and entire documents. Every document can be broken down into these respective groups of abstraction.

(i) An entire document.

(ii) A collection of lines in a document.

(iii) A collection of words that are consecutively placed on each line.

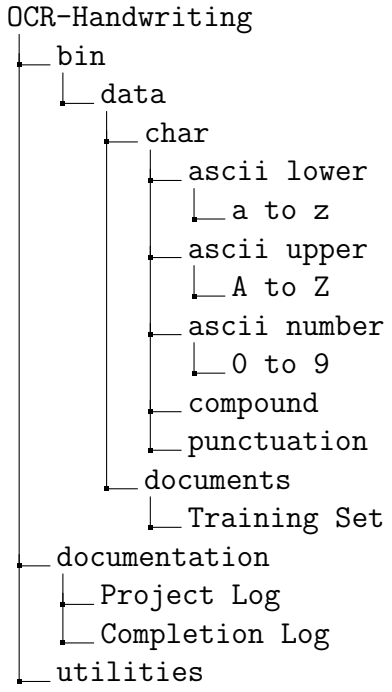
(iv) Single characters that make up the words.

It can be seen that each level abstraction relies on the previous, going all the way down to the individual letters that are on the document. Given the nature of that abstraction Dr. Johnson suggested we start from the ground up, meaning first we will be building the data set for letters, and training a model to recognize other letters of similar (1800's English) style. Our current priority is to build this large data set of characters for our neural network to pull from. After this set is built up we will work on figuring out the optimal design of our model and start to train it. After this section is completed we will have a network that can identify individual characters. From this base level we will then work on the next level of abstraction, that will be able to identify the words in a line. The project will follow a similar style of abstraction based progress until we can use every level to read an entire document.

1.2 Some specifics

We currently have 7 documents that have been allocated for our project. The first 4 will be used to create the data set of images. On top of simple screenshots, we will also employ GPUs to transform the images to get the most mileage out of each photo. The last 3 will be later allocated into development and strict testing sets. These will be allocated as the training set is developed.

1.3 Description of file system



(i) Bin contains all of the 'raw' data such as images, and documents where the images come from. Each sub directory is ordered.

(ii) The section 'compound' has been added due to the nature of John Quincy Adams handwriting. There are several small phrases like 'Mr' and 'Dr' that appear more as one character than 2. This is why it is denoted as 'compound', meaning more than one letter interpreted as a single unit.

(iii) Documentation contains this document, as well as any other documents that are needed to explain the project.

(iv) Utilities contains all scripts, programs, or software that we use as a supplement in order to complete the project.

1.4 Significant Developments

1.4.1 May 14, 2019

Total images taken: 296

Matt created a python script that renames the pictures in the subdirectories according to a naming scheme, this allows for saving files without having to worry about typing the name into the save box. Doing this means the whole process takes 10x less time. When taking photos one can either: focus on a letter saving several in a certain directory (fastest), or save all photos to a "dump" folder and place them afterwards in their correct directory.

1.4.2 May 15, 2019

Letters completed: 'a' , 'e' .

Total images taken: 2,075

Process for quickest imaging (Modified 5/21):

(i) Pick a letter that has a lower than needed sample size (¡ 500).

(ii) Using Lightshot, take a screenshot of a letter, and click the save button. Navigate to the respective directory and save.

(iii) For all subsequent letters, take the photo and use shortcut CTRL-S and it will auto-save to the same directory.

(iv) After you find as many as you can on the page, or several pages, move on to the next page.

(v) After around 125-150 letters from a given set of pages, go onto another set of pages.

(vi) Before pushing to git, run the renameUtilityScript.py file which will rename all of the files to the appropriate schema.

1.5 May 16, 2019

Letters completed: 'b'.
Total images taken: 864

(i) Made changes to the script so that it can be run on any machine without needing to edit the path in the file. If anybody wants to run it, python must be installed and they can either manually run it or write a bat file. There is a provided bat file skeleton, all that needs to be added is the path to the .py utility, and it can be run from anywhere.

1.6 May 21, 2019

Letters completed: 'c', 'd', 'i', 'f', 'g', 'h'
Total images taken: 3,054

(i) Purchased the font "Old Man Eloquent" that we will use to diversify our samples. The current plan is to photograph the font in various contexts and use CUDA to transform the images to extract a large amount of diverse images from one example.

(ii) Mike and Matt had a conversation outlining the plan for hardware to be able to transform images. Once the types of image transformations are chosen, Matt will create software that will be able to be used without programming experience.

(iii) Matt suggested using an image normalization algorithm to give each image the same scale. It would involve locating the global min and max for width and height, and setting each photo to those dimensions. This could be important for making sure the network does not pick up on unintended scale related differences between letters.

1.7 May 22, 2019

Letters completed: 'j', 'l', 'm', 'n'
Total images taken: 1,928

(i) Matt created a completion log complete with all characters so that we can more easily keep track of completed characters.

1.8 May 23, 2019

Letters completed: 'k', 'o', 'p'
Total images taken:1,004

(i) No important developments today. Good progress on imaging.

1.9 May 28, 2019

Letters completed: 'q', 'r', 's', 't', 'u'
Total images taken: 2,244

(i) Today we completed 5 characters towards the end of the alphabet. I have high confidence that by tomorrow we will complete all lowercase imaging. This means we are slightly under halfway to completing the data set

(ii) Another important development is that we passed 10,000 images in just 7 working days, a great achievement. We are currently working with: **11,465 images.**