**Weekly report**

Downloaded Tesseract-OCR and a corresponding python wrapper for windows, pytesseract.

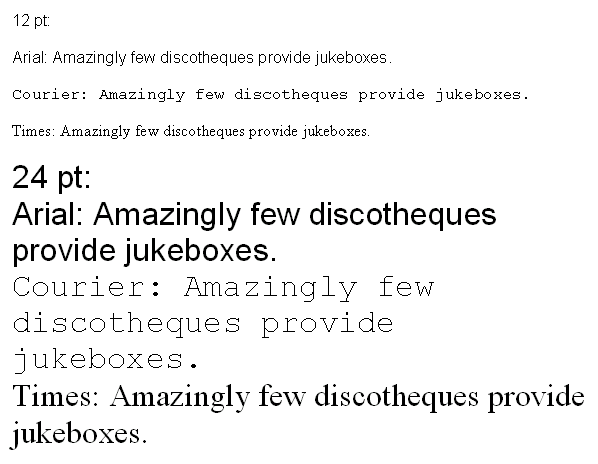
Created a python script to test Tesseract, feeding in images of example typeset text in different languages and fonts to Tesseract.

Figure 1: Font test image for Tesseract, read text with complete accuracy, but misidentified some line breaks.

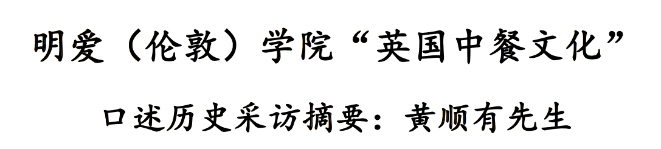


Figure 2: Chinese character test for Tesseract, read with complete accuracy.

Created a slack channel for discussions with other machine learning (ML) MPhys projects about ML.

Tested Tesseract-OCR on a piece of Zhao Mengfu’s ‘Poetry on the Baotu Spring’.

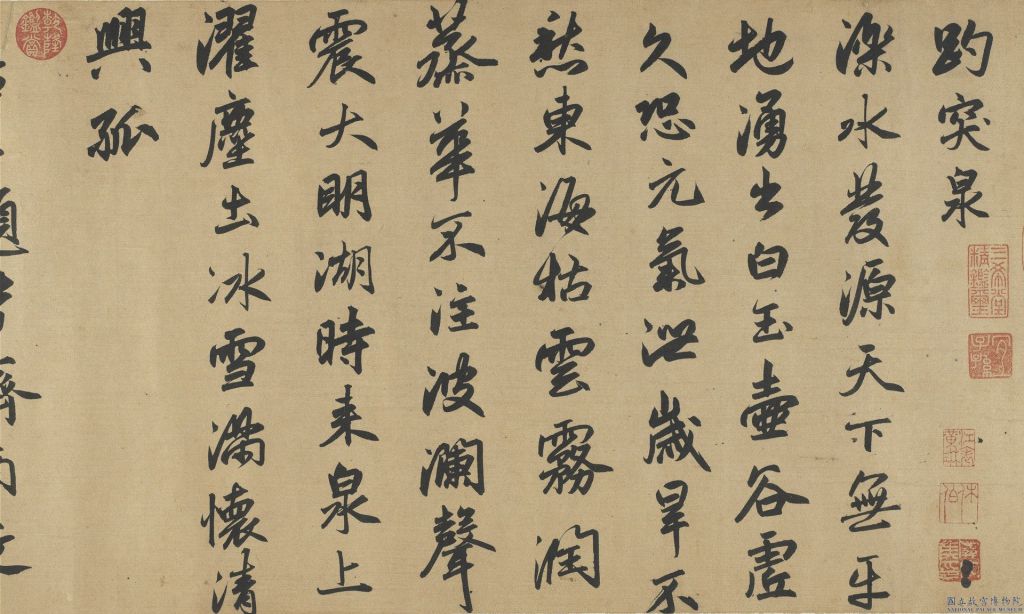


Figure 3: Extract of 'Poetry on the Baotu Spring'. This was first grey-scaled and then binarized using a threshold level.

Tesseract was unable to recognise nearly all of the characters, but it was found that changing the threshold even slightly had a large impact.

Considered a method of using ML to standardize characters to a form similar to ASCII, instead of trying to label each character, then using Tesseract-OCR to perform the final recognition.

Generated .png image of an ASCII Chinese character in the ‘SimSun’ font using Python. Used thresholding to remove the blur at the edge of characters, to help Tesseract.



Figure 4: Comparison of different levels of thresholding. a) has no thresholding, leading to a ‘blur’. b)-d) have decreasing levels of thresholding.

Took the listed 3740 Chinese characters codes in the CASIA 1.0 database and created corresponding .png images of each one. Ran Tesseract on a sample of 100 of these to test its accuracy. It only had accuracy of ~10%, even when threshold levels, image sizes and white border sizes were varied.

This was a surprising result considering Tesseract’s success on the much more varied and stylized characters in Fig. 2. If it is true, we need to think of a new approach.

**Action points for the next week**

1. There are many methods in ML. We will take two methods each, research them in detail and their application specifically to writing and discuss which one we should pursue. We will look at: convolutional neural networks, random forest, wavelet decomposition, and eigenface approaches.

2. Download full CASIA database, test accuracy of Tesseract on it.

3. Research different thresholding techniques and their effect on Tesseract accuracy.

4. Research work done in recognising Chinese handwriting/calligraphy so far. Look at ML techniques for handwriting.