EN.600.461 Computer Vision
Final Project
Recognizing and Translating Text from Images

Joon Hyuck (James) Choi
(Senior Undergraduate)
The Johns Hopkins University
3400 N Charles Street
Baltimore, MD 21218, USA
`jchoi100@jhu.edu`

Abstract

This work explores optical character recognition (OCR) in photos of printed and handwritten documents. It first explores basic preprocessing of photos of documents using `OpenCV` [1] functionalities for blurring, thresholding, and denoising. It next discusses the use of `tesseract-ocr` [2] to perform OCR. We then discuss the incorporation of the Google Cloud Translation API [3] to translate the OCR results to different languages. If finally discusses template matching using the Johns Hopkins University (JHU) logo.

1 Introduction

This work achieved the following main goals stated in the original project proposal: 1) Given a photo of a document, convert it into a clean scanned version; 2) Take the scanned version and perform OCR. Optionally, this work took the OCR output and translated the text into different languages using the Google Cloud Translation API. This work also experimented with template matching using the JHU logo in order to determine whether the input document is an official JHU document or not.

2 Methods

In this section, we discuss the methods we took and external libraries used for each stage of our work.

[1]http://opencv.org/
[2]https://github.com/tesseract-ocr
[3]https://cloud.google.com/translate/

2.1 Image Preprocessing

We implemented our program in Python using `OpenCV`. In order to feed the OCR algorithm clean input to achieve best performance, we used `cv2.medianBlur` to smoothen the input image with an aperture size of 5. Then, we passed the smoothened image through `cv2.adaptiveThreshold` with *adaptiveMethod*= `ADAPTIVE_THRESH_GAUSSIAN_C`, *threshold = binary*, *blockSize=5x5*, and *C=2*. The adaptive method we chose uses the weighted sum of the *blockSize* x *blockSize neighborhood of pixel (x, y) - C* as its threshold value. Parameter values were chosen empirically.

Finally, we performed denoising on the thresholded image to remove noise and make the output clean. We used `cv2.fastNlMeansDenoising` with *templateWindowSize=7*, *searchWindowSize = 21*, and *h=7*. Numbers here were also chosen empirically. The motivation behind denoising in our work was that document scans usually have many random specs here and there. Such noise can greatly reduce OCR accuracy because an OCR algorithm may confuse a random spec with punctuation marks or associate a spec with an actual character near by (e.g. confuse an l with an i).

2.2 OCR

In order to perform OCR on the cleaned inputs, we used the `tesseract-ocr` library through its python wrapper `pytesseract`. We used the raw output from `pytesseract`'s `image_to_string` method to pass into python's file writer and Google Cloud Translation API.

As a side experiment, we used the `Keras` [4] library to train the MNIST [5] dataset of 70,000 handwritten digits on three different convolutional neural networks (CNN). Results will be discussed in section 3.3. We referred to code online [6] to build the three CNNs.

2.3 Translation

We made use of Google Cloud's Translation API in order to translate the OCR processed document into several different languages based on user command line input. We referred to code posted online [7] and modified the details to fit our purpose.

2.4 Template Matching

Aside from OCR, which was the main focus of this project, we experimented with template matching in images. Specifically, we used the JHU logo to determine whether the input document was an official JHU document or not. We naively assumed that official JHU documents contained a JHU logo for our experiment's sake. We used `cv2.matchTemplate` with our input document as the source image, multiple JHU logo images (each with different sizes since `cv2.matchTemplate` is scale sensitive) as templates, and matching method `cv2.TM_CCOEFF_NORMED`. If the normazlied `TM_CCOEFF` value came out to be greater than 0.51 for any of the input JHU logo templates of various sizes, we determined that the input document contained a JHU logo in it. The matching method and determinant value of 0.51 were chosen empirically.

3 Results

Here, we qualitatively discuss OCR results on typed and handwritten input documents.

3.1 Typed Documents

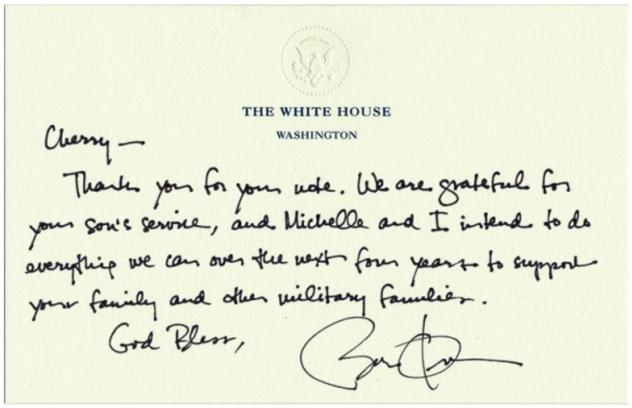
We ran our code on five input document images: `comp_vis_proj_description.png`

[4]https://keras.io/

[5]http://yann.lecun.com/exdb/mnist/

[6]http://machinelearningmastery.com/handwritten-digit-recognition-using-convolutional-neural-networks-python-keras/

[7]http://github.com/mouuff/mtranslate


THE WHITE HOUSE
WASHINGTON

Cherry —

Thanks for your note. We are grateful for your son's service, and Michelle and I intend to do everything we can over the next four years to support your family and other military families.

God Bless,

[illegible]

Figure 1: Sample handwritten document in our dataset

3.2 Handwritten Documents

3.3 MNIST Training with Keras

4 Discussion

5 How to Run the Code

Please extract `FINAL_PROJECT_jchoi100.zip` to view and run the code.

Assuming that the user has Python 2.5+, OpenCV, pytesseract, and tesseract installed, run the following command on the command line in the same directory as `driver.py`.

```
$python driver.py {path-to-image-file} {list of target languages separated by single space}
```

Sample usage:

```
$python driver.py data/doc7.png es fr
```

performs OCR on `doc7.png`, outputs the thresholded image, the text in the original input language, and the text translated in Spanish and French.

- Left and right margins: 1in
- Top margin:1in
- Bottom margin: 1in
- Column width: 3.15in
- Column height: 9in
- Gap between columns: 0.2in

Type of Text	**Font Size**	**Style**
paper title	15 pt	bold
author names	12 pt	bold
author affiliation	12 pt	
the word “Abstract”	12 pt	bold
section titles	12 pt	bold
document text	11 pt	
abstract text	10 pt	
captions	10 pt	
bibliography	10 pt	
footnotes	9 pt	

Table 1: Font guide.

Captions: Provide a caption for every illustration; number each one sequentially in the form: “Figure 1. Caption of the Figure.” “Table 1. Caption of the Table.” Type the captions of the figures and tables below the body, using 10 point text.

Acknowledgments

We thank the python libraries we used for image processing and OCR, the code we referred to for constructing CNNs using `Keras` and making use of the Google Translate API, and MNIST for the handwritten digit dataset.