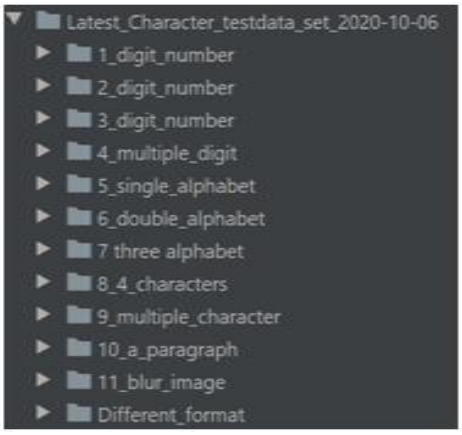
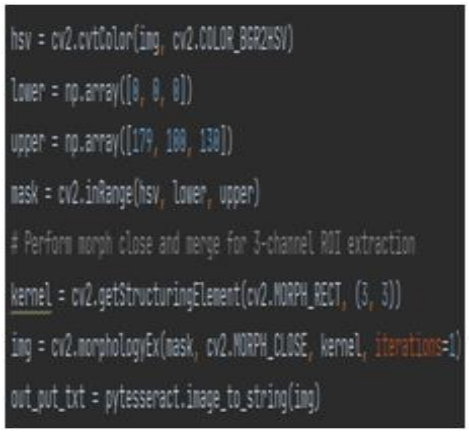
Character recognition in a digital image

By: Chanpreet Singh, Jiayun Feng, Kartik Watts **Supervisor:** Dr. Dax Roberts **Client:** Dr. Waqar khan



There are 12 categories on test data



The main process to handle the data

Pre-processing of the pictures

INTRODUCTION

The **Character recognition in a digital image** is a software which recognizes the characters from the digital image and converts them into editable and searchable data. The aim of developing Character recognition in digital image software is to save the time of the students at the Wellington Institute of Technology (WelTec). In this software, the user needs to upload the image document then he will get content of the image in a digitalized editable format in a short time.

With the transition of storing information using traditional methods, such as papers and books towards using digital platforms. There are still many volumes of data not digitalized yet, which students generally struggle to convert printed resources to digital resources. With our software, students can overcome the problem and save a lot of time.

DEVELOPMENT

At the beginning part of the project, we get the requirement from our client. As the requirement suggested, we need to use Python, OpenCV, CNN to make this Optical Character Recognition (OCR) software. Our team know how to use Python but it's still a challenge to learn OpenCV and CNN for the project. We spend a lot of time learning OpenCV and CNN at the beginning of the project.

After we get to know the main idea of OpenCV and CNN, we try to read the image data with OpenCV and train a CNN for this project. It takes some time to collect the data because the training data is important to get a satisfied CNN.

At first, we focus on the number of characters because recognizing number is easier than the alphabet one. After we finish training number-recognize CNN, we test it and it works fine.

Then we try to train the alphabet recognizing CNN, we realize that we need lots of data and it's kind of hard to collect that amount of data. Our client suggests we use existing pre-trained CNN to recognizing all the alphabet in the image because he thinks there is no need to re-create the wheel and there are already some good pre-trained CNN on the internet.

So, we begin to research different types of pre-trained CNN. After testing and research, we find the tesseract is a good pre-trained Long Short-Term Memory (LSTM) CNN which is an open-source by Google. To use this pre-trained CNN, we need to pre-process the image before feeding it to this LSTM CNN.

After working out the pre-processing part, we begin to collect data from real-life with our phone to test the whole OCR system. We create some test cases on different conditions such as zoom in or zoom out and take the pictures from a different angle (75,90,120). Finally, the success rate of the test is 85%. If the user takes the pictures in zoom-in mode and 90 degrees, the success rate will be much higher. At last, the software is good enough to be used.

CONCLUSION

The conclusion of the project is that the software recognizes the characters very well in most conditions. So, students at Weltec can easily convert their printed resources into digital resources.