

Group 146 Project Report: textOCR

Dhruv Chand, Roshaan Quyum, Arin Khandelwal
{chandd9, quyumr, khanda3}@mcmaster.ca

1 Introduction

- We are creating an application that extracts text from an image, otherwise known as Optical Character Recognition (OCR). It is a multiclass classification task using images with text on it as its input data. It can also be seen as a single label classification task where each data point will be classified into only one class representing the character that it is. We will create a command line application that takes an image with English text as input and outputs the text on the image into the terminal. It involves implementing a number of different parts of the machine learning and model training process such as image preprocessing, text recognition, and character recognition. We shall use machine learning libraries such as OpenCV, NumPy, Scikit-Learn and PyTorch to implement them. If we have the time, we would like to extend the OCR to recognize text in natural photographs (e.g. recognize text in a photo of a stop sign).

2 Dataset

- from the sheet. check the data set from. refer to the new data set.

We chiefly used two datasets: [OCR-Dataset](#) and [Words MNIST](#).

OCR-Dataset consists of the 26 letters of the English alphabet in both lower and uppercase along with all 10 digits; across the 62 classes, there are roughly 210 000 images in total. The data uses 3475 different fonts from publicly available Google Fonts. It is the dataset we trained our model on.

There are a number of preprocessing steps done to the images from this dataset. Images are converted to a specific type (np.uint8) to ensure consistency in the following preprocessing steps. Then, all white pixels on the edge of the letter are cropped

out (the cropping is similar to a bounding box being placed around the letter/digit and then all pixels outside of it removed from the image). The image is then scaled to have dimensions of 64x64, and finally, it is turned into a tensor. After this preprocessing, the images from this dataset can be inputted to the model. This preprocessing here assumes the image contains a single character; however, for images with more than one character, extra preprocessing was needed.

The Words MNIST dataset contains about 10 000 images of various words. All characters from the alphabet (lower and uppercase), all 10 digits, as well as an assortment of special characters/punctuation are in the words in the images. The images have variable sizes and need to be resized to be uniform. The images mostly come from scanned documents and synthetic generation. Data was synthetically generated to have lesser seen characters included in the dataset. Some of the images were labelled manually and some of it was labelled using tesseract OCR and then manually checked after for errors. Our OCR model is built only to handle individual characters; so before running the preprocessing from the previous paragraph on it, the characters in each word needed to be segmented. We had two different attempts at this and will describe both.

Our first, less successful segmentation, first grey scaled the image, applied median blur to remove “salt and pepper” noise, applied inverse binary threshold using Otsu’s method, and dilated the image to make the characters larger in the vertical direction. We then found the contours of each character and drew bounding boxes around them. The dilation was meant to preserve the gap between characters, which manifested horizontally, and remove any “intra-character” space, such as the hole in an “a” or the large amount of space in the column a “u” takes up so that all characters looked roughly like rectangular blobs. Finally, we used

the dimensions of each bounding box to crop out the characters from the original image and return an array of images containing each character. One large flaw of this approach is that if the dilation did not remove all the space in a character's rough column location, then multiple contours would be drawn in that column. This would mean, for example, that both the bounding box containing "a" as well as the circle in "a" would be returned as distinct characters. With the variation in font, this occurred quite often with "o"s and "a"s.

The second, more successful segmentation function implements the Potential Segmentation Columns (PCS) method from the paper: [A New Character Segmentation Approach for Off-Line Cursive Handwritten Words](#). The same rough process is done as before with the exception of blurring before thresholding and instead of dilating the image, we use Zhang Suen's method to thin the characters. After thinning, all columns of pixels in the images are scanned. If there is less than or exactly 1 pixel (this was generally the PCS threshold/limit for the images in the Words MNIST dataset) of white (meaning less than 1 pixel of character), then the column is marked as being a PCS. After the image is fully scanned, the array of marked segmentation columns is looped through. For each set of consecutive PCSs, the middle column of the set is chosen as the segmentation column and the rest are unmarked. Finally, the original image is cropped according to these columns and the individual images are returned. This method worked much better than the first one and is near guaranteed to work when there are gaps between characters. When there aren't any gaps, it still performs well and better than the first method, but clearly not as good. This method was chosen to character segmentation due to its consistency (it is easier to find out why it fails) and better quality.

3 Features and Inputs

- diff features and inputs for pre processing.

The

4 Implementation

5 Evaluation and Progress

- Copy the description of the old model.

6 Error Analysis

- Need to figure out

Team Contributions

7 Figures and Tables