

AI System Technologies (EAI 6020)

Spring 2021

Final Project (Using OCR to obtain discrete data from images)

Avinash Tripathi & Sumit Gorey

**Introduction**

OCR stands for optical character recognition, this technology manages the issue of perceiving a wide range of characters. Characters with handwritten and printed form can be converted to a digital data format & a format which a machine can read.

Consider any kind of serial number or code involving numbers and letters that you need digitized. By using OCR, you can change these codes into a computerized output. The advancement uses different methodology. Put forward obviously, the image taken is processed, the characters extracted and are then perceived.

What OCR doesn't do is consider the real idea of the object that you need to scan. It's anything but "a look" at the characters that you mean to change into a digital format. For instance, in the event that you scan a word it will learn and recognize the letters, however not the meaning of the word.

What OCR doesn't do is consider the real thought of the article that you need to check. It's anything but a look at the characters that you intend to change into a digital configuration. For example, if you scan a word it will learn and perceive the letters, be that as it may not the significance of the word.

**Image Pre-Processing in OCR**

OCR programming regularly pre-processes images to work on the odds of fruitful recognition. The point of image pre-processing is an improvement of the genuine image information. Along these lines, undesirable distortions are smothered, and explicit image features are improved. These two processes are significant for the accompanying advances.

**Character Recognition**

For the character recognition, we need to understand what "feature extraction" is. At the point when the data is too enormous to be processed, just a reduced arrangement of features is chosen. The features chose are required to be the significant ones while those that are suspected to be repeated are overlooked. By utilizing the reduced set of data rather than the underlying enormous one, the performance is increased.

For the person acknowledgment, get what "include extraction" is. Exactly when the information is too huge to ever be prepared, only a decreased course of action of highlights is picked. The highlights picked are needed to be the critical ones while those that are suspected to be rehashed are neglected. By using the reduced arrangement of data instead of the underlying enormous one, the performance is increased.

**Post-Processing in OCR**

Post-preparing is another error adjustment strategy which guarantees high precision of OCR. The precision can be additionally improvised if the output is limited by a vocabulary. That way, the algorithm will be short of words that are permitted to happen in the scanned record for instance. For the process of OCR, this is huge as the algorithm needs to recognize unequivocal parts or shapes of a digitized picture or video stream.

OCR isn't simply used to recognize appropriate words however likewise understand codes & numbers. for distinguishing the long series of letters & numbers this is distinguishing, for example, serial numbers utilized in numerous ventures.

To all the more likely arrangement with various kinds of input OCR, a few providers began to foster explicit OCR systems. These systems can manage the extraordinary images, and to further develop the recognition accuracy, considerably more, they joined different techniques of optimization.

For instance, they utilized standard expressions, or fruitful data contained in the color picture. This procedure of consolidating different optimization method is designated to "application-oriented OCR" or "customized OCR". It is utilized in applications, for example, business card OCR, ID card OCR, and receipt OCR.

In this project we have used the OCR Technology, we have extracted all the data from the desired given images.

In this given assignment, we have got a Burial records dataset, We have fetched the data from the scanned images, All the images are in text & handwritten format. There were following information provided within the data such as Age, date of burial, the name of the deceased person, burial location information.

We have installed all the necessary libraries first

Graphical user interface, application

Description automatically generated

We have uploaded the dataset which is in the image(.jpg) format



After loading the dataset, we have received an output of the scanned burial records, we can easily see that the image shows the age as 41 years , name of the person, date of burial & so on. Graphical user interface

Description automatically generated with medium confidence

In the next step we are going the convert an image into the text format, we tried opening an image from the source path,we have installed the tesseract module which will convert the image to the result & will save it into the result variable, Moreover we have written a text in the text file & saved it to a source path. We have used the tesseract to extract printed text from images.

Graphical user interface, text, application

Description automatically generated

We can easily see from the output below that after using the tesseract we have extracted the data from the image format & converted it into a text format.

A picture containing background pattern

Description automatically generated

Furthermore, after this we have started the preprocessing of the data, we have run some commands to the extracted data to make it more visual, we have used the grayscale image, noise removal, thresholding & canny edge detections.

We have used the grayscale image as it is single dimensional & it has been used as it decreases training complexity of models in different issues & algorithm in Canny edge location.

Moreover, we have used the noise removal to remove the noise in the image such as patches, small dots which is higher in intensity as compared to the rest of the image. It can be used for binary & colored images.

Furthermore, thresholding can be used to divide the image into two segments of pixels, so there will be more clarity in pixels. It is widely used in image processing tasks.

Whereas we have used canny edge detection to extract helpful structural data from various vision objects and drastically lessen the measure of data to be handled.

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

We have use the python tesseract as it will recognize & read the text embedded in images.

As we can see the output which has been

Chart

Description automatically generated

Chart

Description automatically generated

In the next step we will be extracting the useful information from the image file we will be using the split function to make the data in the right alignment

Text

Description automatically generated

As we can see from the output, we can easily understand that the extracted text is not split & is looking in a much formatted & aligned manner.

Text

Description automatically generated

Moreover, after this we tried converting the image properties & have a full overall description, we found the conf interval differs based on if the actual text is present. We have done image to data in this, we can easily understand that whatever the numbers which are shown in the positive range such as 69,69 & 96 in the left column are exactly the showing that a data was there, whereas the values which are shown as -1 & actually null values.

Table

Description automatically generated

After moving forward, we have to find out the location of the elements in boxes using Conf attribute of the image data, if you check the code block above this you can see that most of the text elements shows Conf attribute value above 55.

Text

Description automatically generated

Output:

A picture containing calendar

Description automatically generated

Furthermore, we have stored the data into the database in this step, we have located the directory for storing the data into database.

Graphical user interface, text, application, email

Description automatically generated

We have received an output for the names of the burial records in the jpg format.

Graphical user interface, text, application

Description automatically generated

After the previous step we have done the categorization of the images based on clusters, The reason we are using the OCR is that we can push the csv file into a databases like hadoop's HDFS or SQL server, we can also use this file to retrieve the data stored for a particular person, we can create a dataframe and use it to create data visualization story using tableau or other visualization platforms.

The VGG models are no longer best in class by a couple of percentage points. In any case, they are exceptionally amazing models and helpful both as picture classifiers and as the reason for new models that utilization image inputs. Keras gives an applications interface for loading and utilizing pre-trained models. In this step we have used the VGG-16 prediction which will be showing the top 5 classifications. This code also classifies that what is the probability of the image and its respective sources.

**Clustering Burial records**

According to the slides there are different forms of the burial record and extractig information from each type takes a little bit of different approach. thus the idea behind clustering the similar types of image is to be able to write a cluster specific code to pull out information more accurately.

Graphical user interface, text, application

Description automatically generated



Moreover, As a rule you can utilize any clustering mechanism, we have used a popular k-means clustering. To set up your information for clustering you need to change over your collection into an array X, where each line is one model (image) and each column is a component/feature. if your pictures are little and of a similar size you can essentially have each pixel as a feature. In the event that you have any metadata and might want to sort utilizing it - you can use it as a feature of having each tag in metadata.

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text

Description automatically generated

Output:

Table

Description automatically generated with low confidence

We will be creating a separate csv file for each of the cluster, after this we will be converting the image data into text & append the file in csv.

The future goal is to use this cluster and implement cluster specific preprocessing to have more focused way of editing on the retrieved text, we tried to run our first preprocessing code on the cluster 1, Now here comes the question that why have we used the cluster 1? We have a very crisp & clear answer for this that the initial sample image belongs to the cluster 1, so if we want to preprocess the data for different cluster we need to build a different preprocessing code for them since they vary significantly.

Graphical user interface, text, application, email

Description automatically generated

Output:

Graphical user interface, text

Description automatically generated with medium confidence

**Conclusion:**

OCR innovation is getting extremely broad in professional sector like museums, historical centers, and libraries. This is an incredible method to protect old texts or pictures in a computerized design. All the more critically, these reports can likewise be inspected in the advanced area without upsetting the first actual materials. Taking everything into account, OCR is an entirely astounding innovation that holds a ton of potential. Nowadays, such tools are now very progressed. Nonetheless, Optical Character Recognition will glance far and away superior later on. Artificial intelligence is enroute to getting perhaps the most compelling trends in the coming years, altering data as we probably are aware OCR (Optical Character Recognition) is an innovation that digitizes actual records. It turns images, manually written texts, and printed reports into completely accessible advanced documents. OCR is an extraordinary illustration of how AI solutions are driving database modernization, as these instruments are getting progressively affordable and available. The analysis of sequential lines of curves, the conclusion is that an OCR application makes the best character estimate.

**References:**

What is an OCR? (n.d.). https://www.myfreeocr.com/blog/what-is-an-ocr.

*Image clustering by its similarity in python*. NewbeDEV. (n.d.). https://newbedev.com/image-clustering-by-its-similarity-in-python.

Brownlee, J. (2019, August 19). *How to Use The Pre-Trained VGG Model to Classify Objects in Photographs*. Machine Learning Mastery. https://machinelearningmastery.com/use-pre-trained-vgg-model-classify-objects-photographs/.

Zelic, F. (2021, June 16). *[Tutorial] OCR in Python with Tesseract, OpenCV and Pytesseract*. AI & Machine Learning Blog. https://nanonets.com/blog/ocr-with-tesseract/.

*pytesseract*. PyPI. (n.d.). https://pypi.org/project/pytesseract/.

*What Is OCR or Optical Character Recognition*. Docdigitizer. (2021, March 11). https://www.docdigitizer.com/2019/08/01/what-is-ocr/.

*OCR technology for businesses - Its applications and benefits*. shufti\_pro\_img. (2021, May 28). https://shuftipro.com/blog/ocr-technology-for-businesses-its-applications-and-benefits/.

Kumar, S. (2019, October 10). *A straightforward introduction to Image Thresholding using python*. Medium. https://medium.com/spinor/a-straightforward-introduction-to-image-thresholding-using-python-f1c085f02d5e#:~:text=What%20is%20Thresholding%3F,image%20recognition%20accuracy%2C%20segmentation%20etc.