Text Recognition in Images

Yogita[1]
19MCA0009
MCA, SITE School (VIT)

Deepali Poddar[2]
19MCA0019
MCA, SITE School (VIT)

Mansi Lohiya[3]
19MCA0154
MCA, SITE School (VIT)

***Abstract*- In the modernized world, most of the information is present either on paper or in other forms such as Images, Videos etc. Large information is stored in images. Knowledge of text in images is very popular nowadays, as it helps in better understanding of the image components. This paper proposes an approach for reading text from the set of images. Data set consists of various images which have text on them. Images in the data set need to be pre-processed, so that we can get the content from the images easily. The goal of fetching or detecting text from various images can be achieved by using Optical Character Recognition software which will convert input images from the data set into a digitized form so that it can be processed further or manipulated by the machine. In this paper, the data set of the images will be taken and information from the images will be detected.**

***Keywords*-** Text recognition, OCR, Optical Character Recognition, Data pre-processing, Denoising, Text Extraction, Pytesseract

I. INTRODUCTION

Number of images that are present on the internet are increasing exponentially each day. These images can be an image of a person or image with some scenery on it, or it can be with some text. Text which is written on these images might have some useful and important information on it. This information can be used in the future for further processing or it can be used for some other purpose. For recognising the text from these images there are several steps that are involved in it such as data pre-processing, detection and extraction. However, there are certain issues that might occur while extracting text from the images. There are various reasons for it such as variation in text due to size difference, styles, orientation, and contrast, bad or noisy background which makes it even more difficult to extract text from the images. Text extraction from the images basically means converting human readable text into a format which can be easily understood by the system or machine for further processing. Nowadays, there is a huge demand for capturing and saving the data in the machine which is gathered from various documents or images. One of the ways to store the data into the machine by scanning the document and storing them in the form of the images. But reutilization of the context which is stored in the form of images can be a tedious task. Here comes one of the advanced technologies known as OCR (Optical Character Recognition) into play. OCR is the advanced software that helps in retrieving and storing the content from the images in particular text files. OCR is the technology which converts human understandable content into machine readable form so that machines could use the content and process it whenever needed.

For doing this task efficiently, there are some challenges that need to be handled to get more accurate results using OCR. Few of these challenges are font characteristics

and quality of the images. Due to these factors images might not be correctly processed by OCR, which will not give us our desired result. So, for overcoming the challenges, images that are given as the input should be of good quality or they should be pre-processed so that they can be easily processed by the system.

Image pre-processing means improving image quality that have unwanted distortions or noise by enhancing some important features for future processing. Pre-processing stage is one of the important stages that is preceding the text extraction process. It maintains output quality for the rest of the stages. The OCR rate of success is dependent on the success percentage of the pre-processing or the previous stage. In addition to this, there are various factors such as watermarks and uneven illumination that influence the accuracy of the OCR output.

II. RELATED WORKS

In [1], techniques are used to analyse documents by reading its content, number plate of vehicle can be recognized, natural scene, etc., which was done on the basis of CNN and LSTM architecture. Hence, various tasks were automated that have been linked with each other because it has better identification reliability.

Nowadays the main creator of information is social media, as it contains huge data that should be pre-processed before getting valuable information from it. In [2], Tesseract OCR technique is used to generate highly accurate documents from the extracted text from images. The pre-processing techniques include- document localization, resolution improvement, text localization.

In [3], the K-means algorithm is used to reduce the process complication. In this, two different algorithms are evaluated and then compared to opt out the better accuracy between classification and efficiency. In [5], MKL framework is used to transform text detection into pattern classification. The Kernel method is finer than the traditional SVM method.

Digital image processing is increased very quickly and plays a major role in the fields like AI, robotics and many more. This paper also helps in providing direct links between humans and computers which establish easy understanding between them.

According to the survey, text recognition in images haven’t been explored much. As there is still some researches are on-going. In [5], major challenges of OCR are talked about and then their further important role is discussed. OCR technique also provides accurate information as its comprehensive algorithm focuses on a dataset, which is the main source of text extraction.

Apart from the traditional text detection and recognition models, a scalable feature learning algorithm is used to produce desired results from the given dataset. This learning algorithm is more reliable and gives more refined output. It also enhances the ability to achieve performance through scalability.

To reduce the workload of humans in rectifying text and transform it into useful information, machine learning algorithms

came into existence. These algorithms help in providing more reliable and relevant information [9].

In [8], extracted information from text is shown in directed acyclic graphs, so that it would be easy to understand. But, there are few errors which may not be rectified by this graph algorithm, so that various statistical models were used to overcome this problem. By using the concept of focus, text extraction is done. Initially, focus is made and then with the help of HCL distance measure the dataset is compared and targets the relevant text [9].

Text can be recognised easily in images but difficult in videos because sometimes videos are of low quality, so to fix this challenge holistic approach and SVM is used [10]. These approaches combined transform distortion into accurate shape.

Techniques such as- face detection and OCR are used to better update the text information which is collected from image sources [11]. By this, generated text gives relevant and useful information. Similarly, video OCR is used to identify the grey level character images [12], this helps in reducing the level of noise or any kind of distortion among the video.

III. PROPOSED ARCHITECTURE WITH DETAILED DESCRIPTION

A. Architecture -

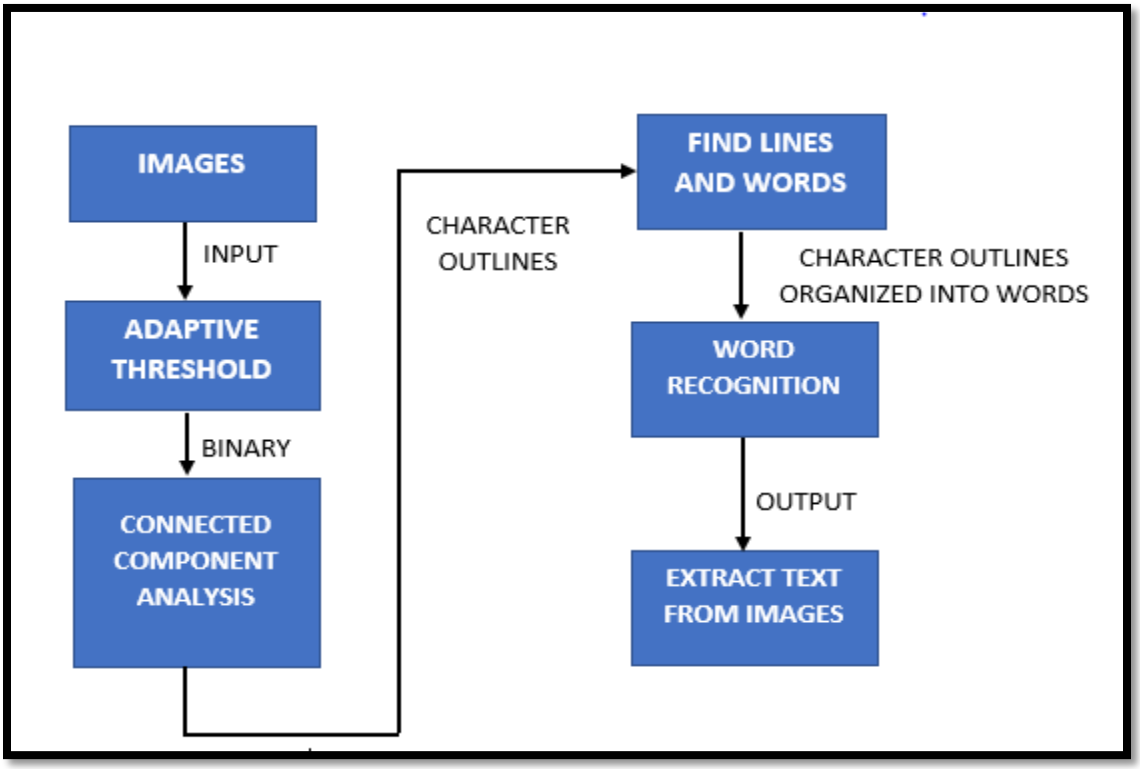

Fig 1. Image Extraction Using OCR

B. Description -

Tesseract OCR works in the steps mentioned below.

It can read text from one image or more than one image and extract the written text from the images and store it in the text or pdf depending on the type of file provided for storing the output.

Steps of this procedure is as follows:

Step1: First it read the images from the specified path.

Step 2: Now Adaptive Thresholding converts the images into binary images.

Step 3: Next step is connected component analysis which is used to extract character outlines.

Step 4: Then outlines are converted into Blobs and they are organised in Text.

Step 5: Now this blob gets organised into text lines. These text lines are divided into words according to the spacing character type.

Step 6: In this step recognition of text and machine learning algorithms is applied.

Step 7: Finally extracted text is stored in the specified path and in specified format.

IV. EXPERIMENT

A. Data Set:

Fig 2. Image Dataset

In this paper, text extraction from images is done using Python inbuilt packages and functions. Used image from PIL (Python Imaging Library) which supports opening, manipulating and saving images in different image file formats such as jpeg, png, etc.

B. Libraries:

1. `from PIL import Image`

We used Pytesseract which is an open source OCR software. It is free and needs to be installed on the machine as Pytesseract depends on it. It is the OCR tool for python which used to read text from images.

2. `import pytesseract`

Next used library is OS which provides modules to interact with the operating system. It comes under Python's standard utility modules. It grants permission to interface with the underlying operating system.

3. `import os`

Imported Pandas to create data frames. It will store file name and corresponding extracted text in tabular format.

4. `import pandas as pd`

We need to give a full path of the directory where tesseract is installed so that it can open it and can use its functionality. All the images which need to be processed should be stored in a file so that tesseract can take images as an input and process it further. Now pytesseract converts images in binary form and extracts text from it using the built in function pytesseract.image_to_string (). It saves the extracted text in specified file format. The stored file can be accessed and we can see the extracted text.

V. OUTPUT

Text gets extracted from the set of images and stored in a file in specified format. The directory of the stored file will be the same as the input directory.

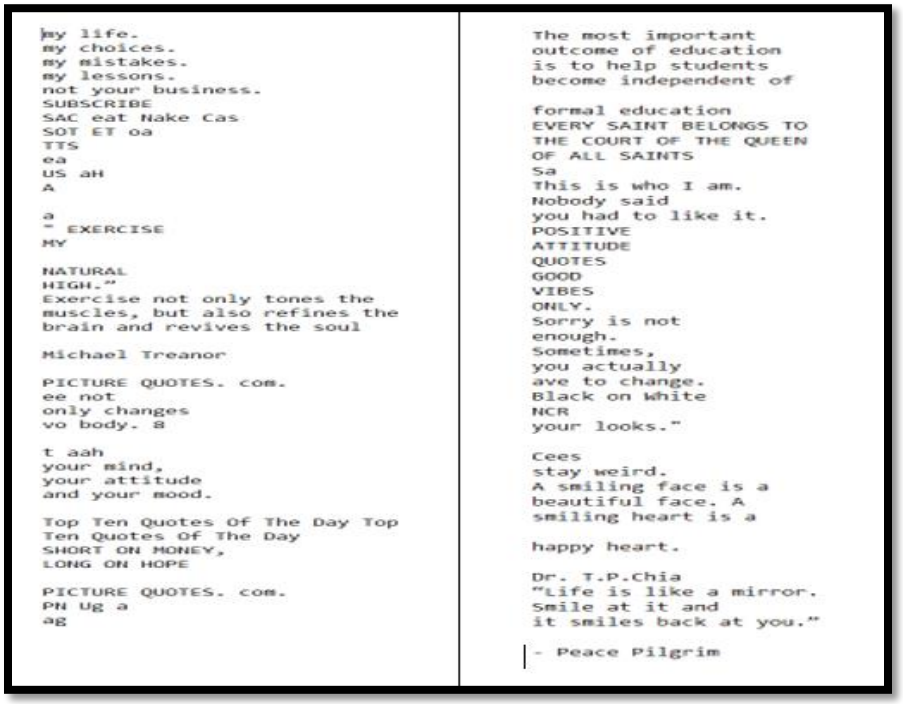

Fig 3. Extracted Text

VI. CONCLUSION

This paper concludes that nowadays, images contain more textual information or we can conclude that more information comes in the format of documents or images. To read that information from images, various researches have been done and various techniques were introduced to extract that textual data from an image document or from a normal image which contains some kind of text. Also this technique can work on both black and white and coloured images. Here we have used tesseract OCR technique to extract text information from various images. In this, multiple images at once can be taken as an input dataset and then after processing of in-built functionalities of OCR the output text will be generated. This output is in the actual form of text rather than images which is easy to understand and is more simplified than images.

VII. FUTURE WORK

In this paper, multiple coloured, black and white images are taken as the input to be processed with the OCR software and as the output we are obtaining the text content that is written on the images into our desired format such as a notepad file, word document, data frame etc. For future work, this extracted text from these images can be used as another text data set for further classification. For example, we can classify the content or quote written on images whether it is a motivational quote, travel quote, lifestyle or fitness quote etc. The classification will be totally dependent on the type of images that you are including in the dataset.

VIII. REFERENCES

[1] Shrivastava, Anupriya, et al. "Deep Learning Model for Text Recognition in Images." *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. IEEE, 2019.

[2] Akopyan, M. S., et al. "Text Recognition on Images from Social Media." *2019 Ivannikov Memorial Workshop (IVMEM)*. IEEE, 2019.

[3] Zhao, Xiaofan, et al. "Image Preprocessing Algorithm Based on K-Means." *2019 20th International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT)*. IEEE, 2019.

[4] Hamad, Karez Abdulwahhab, and Mehmet Kaya. "A detailed analysis of optical character recognition technology." *International Journal of Applied Mathematics Electronics and Computers* Special Issue-1 (2016): 244-249.

[5] Lu, Shen, et al. "Text detection in images based on Multiple Kernel Learning." *2011 International Conference on Machine Learning and Cybernetics*. Vol. 4. IEEE, 2011.

[6] Ikica, Andrej, and Peter Peer. "An improved edge profile based method for text detection in images of natural scenes." *2011 IEEE EUROCON-International Conference on Computer as a Tool*. IEEE, 2011.

[7] Wasankar, Subodh L., et al. "Self intelligence with text recognization." *2010 International Conference on Signal and Image Processing*. IEEE, 2010.

[8] Saidane, Zohra, Christophe Garcia, and Jean Luc Dugelay. "The image text recognition graph (iTRG)." *2009 IEEE International Conference on Multimedia and Expo*. IEEE, 2009.

[9] Kim, Egyul, SeongHun Lee, and JinHyung Kim. "Scene text extraction using focus of mobile camera." *2009 10th International Conference on Document Analysis and Recognition*. IEEE, 2009.

[10] Lee, SeongHun, and JinHyung Kim. "Complementary combination of holistic and component analysis for recognition of low-resolution video character images." *Pattern Recognition Letters* 29.4 (2008): 383-391.

[11] Ye, Qixiang, et al. "Text detection and restoration in natural scene images." *Journal of Visual Communication and Image Representation* 18.6 (2007): 504-513.

[12] Kim, Jinsik, et al. "Stroke verification with gray-level image for Hangul video text recognition." *18th International Conference on Pattern Recognition (ICPR'06)*. Vol. 2. IEEE, 2006.