CHARACTER RECOGNITION

CHENCHU ARAVIND V - 20MIA1126
chenchuaravind.v2020@vitstudent.ac.in
+91 73584 31959

NITHYASRI S - 20MIA1017
nithyasri.s2020@vitstudent.ac.in
+91 79049 89850

SARVESWARAN MG - 20MIA1128
sarveswaran.mg2020vitstudent.ac.in
+91 99524 40626

JAIGANESH S - 20MIA1055
jaiganesh.s2020@vitstudent.ac.in
+91 73586 74041

ABSTRACT

Text recognition in photos is a current area of study that aims to create a computer programme that can automatically read text from images. The objective of this paper is to generate text from the different images and extract particular text only. In this way we don't have to search explicitly search for the text needed. The python program using OCR and EasyOCR and to recognize particular type of text Regular expressions are used. Optical Character Recognition (OCR) of documents have invaluable practical worth. Optical character recognition is a science that enables to translate various types of documents or images into analyzable, editable and searchable data

KEYWORDS: EasyOCR, Bounding Boxes, Regular Expression, CNN, LSTM

INTRODUCTION

It is normal and customary for us to urge the creation of pattern-recognition software and hardware. From automatic facial and optical character recognition to fingerprint

It is obvious that precise and trustworthy pattern recognition by machines would be very helpful in a variety of situations, including identification, speech recognition, DNA sequence identification, and much more. In an ongoing effort to create a computer system that can automatically extract and analyse text from photographs, optical character recognition is a study field. These days, there is a great need to store information from printed or handwritten documents to a computer storage disc so that it can be used again by computers in the future. A quick and easy method of transferring data from these paper documents to a computer system would be to scan the papers first and then store the images as files. However, it would

be highly challenging to read or query text or other information from these image files in order to use this information again. It is therefore necessary to develop a method for automatically retrieving and storing data from image files, particularly text. Obviously, this is not a simple task. To successfully automate, a few significant obstacles must be identified and overcome. The quality of photographs and the font properties of characters in paper documents are just two recent issues. These difficulties sometimes prevent computer systems from correctly recognising characters. Optical character recognition (OCR) is the process of modifying or converting any type of text or text-containing document, including handwritten, printed, or scanned text images, into an editable digital format for additional and more thorough processing. Text in these papers can be automatically recognised by a machine thanks to optical character recognition technology. In the real world, it is comparable to the union of the human body's intellect and eye. Although the human brain processes the detected or extracted text read by the eye, an eye can detect, view, and extract text from images . Of course, OCR technology is still too primitive to match human talent. The calibre of the input documents directly affects the efficiency and accuracy of OCR.

BACKGROUND STUDY

1.Recognition and Evaluation of Optical Character Recognition (OCR) by Deepak Kadam, Prathamesh Chavan, Prashant Pandhara:

Here, documents are printed on paper. Because the document is repeatedly copied and changed during subsequent processing steps, it exists in a variety of formats copies. Because this method of processing is inefficient, we decide to digitise their documents. Working with files is less expensive than working with traditional documents because there is no need for document storage. Furthermore, because each document exists in a single copy, all changes or notes are visible to all document users. Document digitalization consists of three major steps: scanning, indexation (data entry), and presentation of digitalized documents. In this paper, we concentrated on indexation, particularly the use of automatic systems in this process. Character recognition has grown in popularity and has emerged as an important research area. Character recognition is a research area in which techniques are used to classify character inputs into

predefined classes. Several algorithms have been proposed, but the choice of algorithm and classifier may produce different results for different problem domains.To evaluate and compare results, we employ three classifiers and two feature extraction techniques. To compare the results of Local Binary Pattern (LBP) with Support Vector Machine (SVM) and Neural Networks, we use a 1-m approach (NN). The following is how the paper is structured: In Section I, we present the overall definition and description of OCR, and in Section II, we present our approach as well as the problem statement. Section III presents our evaluation, and Section IV discusses our findings. Section V concludes with our conclusion and future work.

2.Object Detection and Identification by Rohith Sri Sai, Mukkamala Rella, Sindhusha Veeravalli, Sainagesh:

Computer Vision is the branch of science that studies computers and software systems that can recognise and understand images and scenes. Image recognition, object detection, image generation, image super-resolution, and other aspects are all part of computer vision. Face detection, vehicle detection, pedestrian counting, web images, security systems, and self-driving cars are all applications of object detection. We are using highly accurate object detection algorithms and methods such as R-CNN, Fast-RCNN, Faster-RCNN, RetinaNet, and SSD and YOLO in this project. Using these deep learning-based methods and algorithms, which are also based on machine learning, necessitates a thorough understanding of mathematical and deep learning frameworks such as TensorFlow, OpenCV, imageai, and others.

3.EasyOCR by pianalytics:

EasyOCR Software will assist anyone who needs to convert documents. Keeping the original content important when converting will help retain the information in order to obtain an easy OCR software. Assume you need to OCR a 200-page document, but you don't know anything about OCR software. The task may be daunting, but the software procedures will be clear-cut and simple to follow with a simple OCR software.The Concept of Converting a Large Number of Pages Will No Longer Imitate You. Employees save a lot of time when they use simple OCR software.

4. Handwriting Optical Detection by Jamshed Memon, Maira Sami, Rizwan Ahmed Khan And Mueen Uddi:

Given the prevalence of handwritten documents in human transactions, optical character recognition (OCR) of documents is extremely useful. Optical character recognition is a science that allows various types of documents or images to be translated into analyzable, editable, and searchable data. Researchers have used artificial intelligence / machine learning tools to automatically analyse handwritten and printed documents in order to convert them to electronic format over the last decade. The goal of this review paper is to summarise research on character recognition of handwritten documents and to provide research directions. In this Systematic Literature Review (SLR), we collected, synthesised, and analysed research articles published between 2000 and 2019 on the topic of handwritten OCR (and closely related topics). We used widely used electronic databases and followed a predefined review protocol.

PROPOSED METHODOLOGY

One of the frequent applications of deep learning is optical character recognition. Examples include converting handwritten prescriptions, identifying cars by their licence plates, converting PDFs or images to text, and verifying signatures.

There are a number of excellent paid OCR API services available on the market, including Google Cloud Vision, Amazon Textract, and Microsoft's Cognitive Service. At the same time, Keras-OCR, Pytesseract, and easyocr are excellent free open-source APIs. Similar to other for-profit API providers, they likewise produced good outcomes.

Existing System only detects all the characters in the given image, In this project Regular Expression is also added which helps to filter out the characters as per our need and Subprocess is also added to search for the file in the system and open them.

IMPLEMENTATION

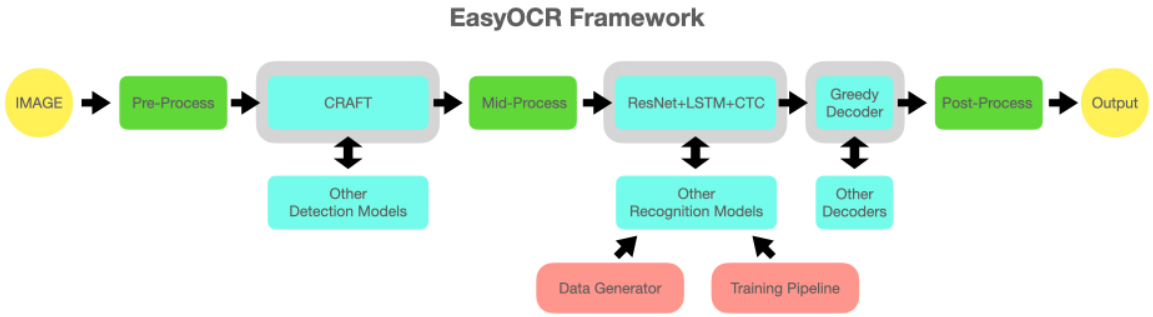
With the aid of opencv, the webcam will capture the image. With the aid of easyocr, the text is then extracted. The file is searched using the text that was extracted. The file will be found and opened by OS subprocesses.

OPENCV

A computer vision and machine learning software library called OpenCV is available for free use. A standard infrastructure for computer vision applications was created with OpenCV in order to speed up the incorporation of artificial intelligence into products. OpenCV makes it simple for businesses to use and modify the code because it is a product with an Apache 2 license.OpenCV is a library of computer vision and machine learning algorithms. It can be used to detect and recognize faces, identify objects, classify human actions in videos, track camera movements, track moving objects, extract 3D models of objects, etc. It has C++, Python, Java and MATLAB interfaces and supports Windows, Linux, Android and Mac OS. A full-featured CUDAand OpenCL interfaces are being actively developed right now.

EASYOCR

OCR, originally known as optical character recognition, is a modern-day digital revolution. OCR is essentially a whole process through which digital photos and documents are processed to extract the text and produce editable text that can be used in other applications. With the use of OCR technology, you can turn a variety of documents into editable and searchable data, including scanned paper documents, PDF files, and digital camera photographs. While using EasyOCR, We discovered that it is the most simple method for detecting text from photos even when a high-end deep learning library is supporting it on the backend, which is the case with every other OCR enhances its accuracy and credibility. 42+ languages are supported by EasyOCR for language detection. Jaided AI Company is the company that developed EasyOCR.



SUBPROCESSES

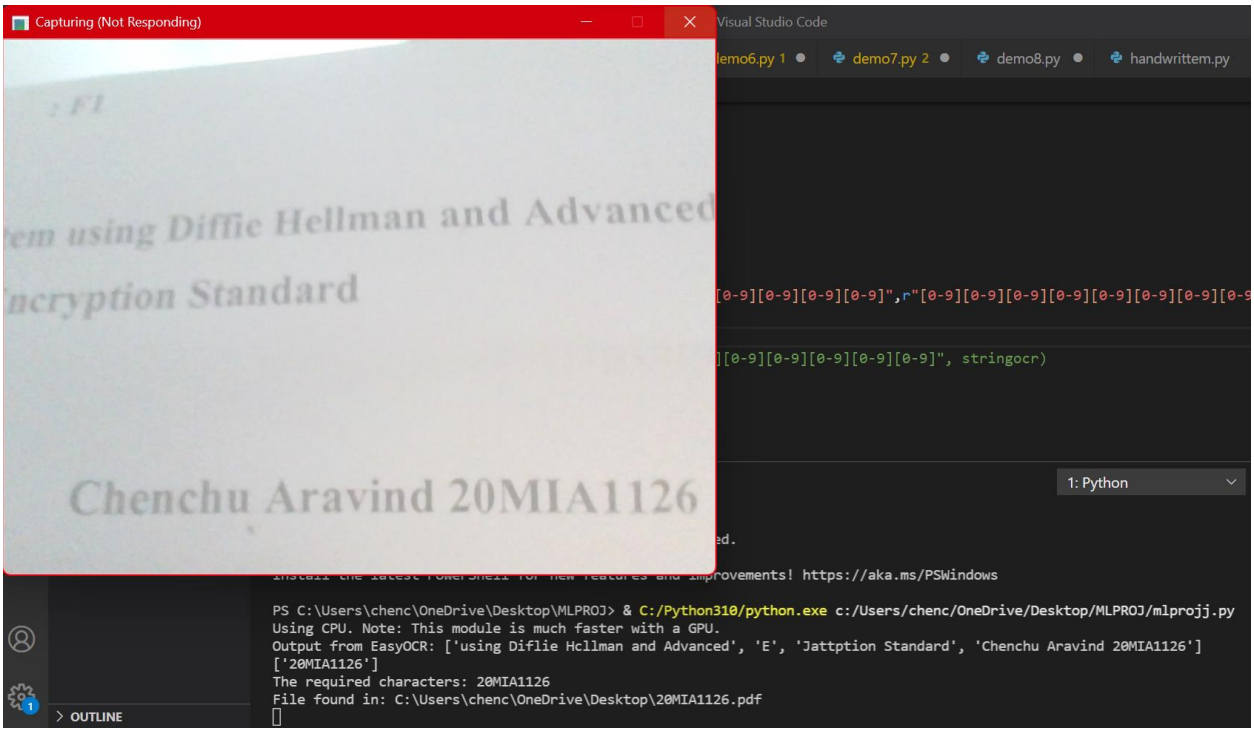
By spawning new processes, subprocesses in Python are utilised to run new applications and scripts. The subprocess module also makes it possible for the user to start new processes directly from a Python programme that is already running. So, you may run code from C or C++ programmes or external apps from a git repository using a subprocess in Python. You can also get exit codes and input, output, or error streams in Python by using subprocesses. If you've ever wanted to simplify your command-line scripting or use Python alongside command-line apps—or any applications, for that matter—Subprocess Python's can be helpful. The Python subprocess module may support a variety of tasks, including launching graphical user interfaces and running command-line programmes.

RESULTS AND DISCUSSION

When the Image is provided to the EasyOCR model by capturing using system webcam. The characters are recognized based on few qualities namely brightness, angle, etc. Result is successfully passed to the subprocess to search for the file named exactly as the text needed to be recognized and the file is opened in any of the pdf readers, in this project the file is opened in Acrobat Reader.

Result-1(When the character 1 is differently printed)

Result-2(When Image with Multiple words)



CONCLUSION AND FUTURE WORKS

The file opened automatically, indicating that it had successfully identified the specific text from the image. Additionally, we may expand the task by taking a picture with a mobile device and opening it on a computer. this can be done by giving the computer the identified object; the computer will then open the file automatically.

REFERENCES

1. https://www.jaided.ai/easyocr/tutorial/
2. https://www.analyticsvidhya.com/blog/2021/06/text-detection-from-images-using-easyocr-hands-on-guide/
3. https://www.guru99.com/python-regular-expressions-complete-tutorial.html#7
4. https://developers.google.com/edu/python/regular-expressions
5. https://www.analyticsvidhya.com/blog/2021/06/text-detection-from-images-using-easyocr-hands-on-guide/
6. https://pianalytix.com/using-easyocr-library-for-text-extraction/
7. https://www.researchgate.net/publication/343273822_Handwritten_Optical_Character_Recognition_OCR_A_Comprehensive_Systematic_Literature_Review_SLR
8. https://www.ijtsrd.com/papers/ijtsrd41099.pdf
9. https://analyticsindiamag.com/hands-on-tutorial-on-easyocr-for-scene-text-detection-in-images/
10. https://ceur-ws.org/Vol-2870/paper15.pdf