iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** VI **Month of publication:** June 2023

DOI: https://doi.org/10.22214/ijraset.2023.53364

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue VI Jun 2023- Available at www.ijraset.com

Handwritten Text Recognition from Image

Kalpesh Joshi[1], Swarup Patil[2], Sujal Patil[3], Swami Patil[4], Sushant Patil[5], Tanishaa Patil[6]
Department of Engineering Sciences and Humanities (DESH), Vishwakarma Institute of Technology, Pune - 411037

Abstract: ***A computer vision program called Handwritten Text Recognition (HTR) attempts to recognize and translate handwritten text from scanned or photographed images. In this project, we suggest implementing an HTR system using Tesseract and OpenCV. English, Chinese, and Arabic are all supported by the popular open-source optical character recognition (OCR) engine known as Tesseract. It is employed to find and identify printed text within photographs. On the other hand, OpenCV is a well-liked computer vision library that offers several tools for processing and analyzing images.***

The pre-processing step of the proposed system uses OpenCV to increase the input image's quality and OCR accuracy. After that, Tesseract receives the pre-processed image for text recognition. The extracted text is then saved in a text file after being identified. To enhance the quality of the input image, the project will use several pre-processing techniques, including deskewing, noise removal, and binarization. With the help of a sizable dataset of handwritten photographs, the Tesseract OCR engine is taught to recognize handwritten text more accurately. The HTR system can be used in a variety of fields, including document analysis, historical manuscript digitalization, and postal automation. It can also be applied in academic settings to help students translate their notes and assignments. Therefore, it is anticipated that the proposed HTR system employing Tesseract and OpenCV will offer a reliable and effective method for identifying and transcribing handwritten text from photographs.

I. INTRODUCTION

Offline handwriting recognition is the automated process of converting handwritten text that is present in an image into a machine-readable format that can be used by various computer and text-processing applications. In other words, it involves extracting text from handwritten notes or pages. It is referred to as "offline" because it deals with handwritten text that is not generated digitally using tools like stylus or Apple pencil.

For image processing and optical character recognition (OCR) operations, two well-known open-source libraries are Tesseract and OpenCV. Tesseract is a Google OCR engine that can identify text in scanned documents, pictures, and other sources. It supports a variety of languages, and accuracy can be increased by training with a large training set of data.

The OpenCV package is a potent tool for computer vision applications like object detection, pattern recognition, and picture processing. It offers numerous techniques and tools for modifying and analyzing photos, including feature identification, segmentation, and image filtering.

It is possible to create a variety of applications that include identifying and extracting text from photos by fusing Tesseract and OpenCV. These libraries, for instance, can be used to create handwritten text recognition systems, where Tesseract is used for OCR and OpenCV is used to preprocess the image data.

Due to their open-source nature and capacity to offer potent tools for image processing and OCR tasks, these libraries have grown in popularity in the field of computer vision.

II. LITERATURE SURVEY

Handwritten character recognition (HCR) is the process of identifying characters in photographs, papers, and other sources and transforming them into forms that can be processed by machines. A significant barrier still exists in the accurate recognition of complexly formed compound handwritten characters. [1]

Text detection, text segmentation, and character recognition are just a few of the techniques that go into extracting text from an image. For this, Tesseract OCR and Open CV are coded into Python. For this, an optical character recognition system with multiple algorithms is needed. The most precise optical character recognition engine available right now, Tesseract, was created by HP. A machine learning and computer vision software library is called OpenCV (Open-Source Computer Vision Library). [2]

The methods used to extract text from an image include text detection, text segmentation, and character recognition. Tesseract OCR and Open CV are written in Python to accomplish this.

An optical character recognition system with several algorithms is required for this. OpenCV is a popular software library used for computer vision and machine learning applications. It is important to note that the accuracy of text extraction systems can be significantly impacted by various disturbances. The input image for our suggested system has a text-filled, complicated background. The segmented characters are compared to the character matrices that have been stored in the text recognition step, where the closest match for each character is presented. [3]

We will require the tesseract package, which is a wrapper for the tesseract engine, to recognize handwritten text in Python. We will also require the pillow library, which extends Python's image processing capabilities, as we are working with images. Massive amounts of computer power are becoming widely available locally and, on the cloud, unfathomable amounts of data may be extracted not only from the visual domain thanks to the ongoing development and improvement of machine learning techniques. [4]

One of the most important tasks in document image analysis is text extraction. Without the ability to recognize characters, automatic text extraction is used to extract text-only sections. To extract text from an image, the process involves four main steps: text detection, localization, segmentation, and enhancement. These steps are essential for accurately recognizing and extracting text from an input image. Each character must be separated during text extraction to be sent into the recognition stage. Several text extraction methods, including region-based, edge-based, texture-based, and morphological-based methods, can be used to do this. OCR (Optical Character Recognition) enables us to identify text that has been extracted from text after segmenting and edge-detecting the characters in it.[5]

The current requirement is text extraction from images. It can be used for a variety of purposes, including reading bank checks, converting any handwritten data into structural text, updating outdated documents, etc. As a result, prior to extracting the text, the input is subjected to various procedures, including de-skewing, turning images into black and white, classifying columns, paragraphs, and captions as discrete blocks, and normalizing. Next, we extract some characteristics of the symbols. It improves precision. After that, we categorize the characters, utilize edge detection techniques to gain a thorough understanding of the images, and then use OCR (Optical Character Recognition) to determine the output of the supplied input photographs.

III. PROPOSED METHODOLOGY

Handwritten Text Recognition and Extraction is done using the Python program language. The following python libraries are required for this process.

A. Tesseract and OpenCV

Python-tesseract is an OCR tool that is designed to recognize and extract text from images, and it is implemented in Python. It works by utilizing Google's Tesseract-OCR Engine and is capable of reading various image types supported by the Pillow and Leptonica imaging libraries. OpenCV-Python, on the other hand, is a collection of Python bindings that were created to address computer vision issues, and it was developed by Guido van Rossum, the creator of Python known for his creation of a user-friendly and easily understandable programming language.

Steps

1) Load the handwritten text input image.

2) The image is improved through pre-processing, and OpenCV is used to extract the text sections. Techniques including scaling, binarization, noise removal, deskewing, and segmentation can be used in this context.

3) To identify the text regions and extract the recognised text, send the pre-processed image to the Tesseract OCR engine. Tesseract recognises text using a combination of deep learning and pattern recognition methods.

Put the detected text in a database or text file for later processing or analysis.

VALENTJNE

4) Extract and save the text: The system can be used to identify handwritten text from fresh input photographs after the OCR engine has been trained and tested. The extracted text is then saved in a text file after being identified.

5) Provide a user interface so that people may upload photographs and read the text that has been identified. A desktop application or a web-based interface can be used to do this.

6) Evaluation and system improvement: In the end, assess the system's effectiveness and pinpoint areas for development. The OCR engine may need to be retrained, the image pre-processing methods adjusted, or new features like handwriting recognition or natural language processing added.

IV. RESULT AND DISCUSSION

Tesseract and OpenCV's Handwritten Text Recognition (HTR) system's output is influenced by a number of variables, including the input image quality, the quantity and variety of the training dataset, and the OCR engine's accuracy.

The Tesseract OCR engine can recognize printed text with a generally high level of accuracy. Yet, given the wide range of handwriting styles and fonts, reading handwritten text might be more difficult.

The quality of the input photos can be increased by using several image pre-processing techniques, such as deskewing, noise removal, and binarization. Higher text recognition accuracy may result from this.

Accuracy can also be increased by training the Tesseract OCR engine with a large and varied dataset of handwritten images. The OCR engine may learn to recognize various patterns and features of handwritten text by being exposed to a variety of handwriting styles and typefaces.

It's crucial to check the OCR engine's accuracy after training it using a different set of test photos. This can show how accurate the OCR engine is and show where it needs to improve.

It is possible to extract the recognized text and save it in a text file for later examination or processing. Applications like document analysis, postal automation, and the digitization of old manuscripts can all benefit greatly from this.

Overall, Tesseract and OpenCV can be used to create a Handwritten Text Recognition system that recognizes, and transcripts handwritten text from photos. To get the best results, it is crucial to properly adjust the pre-processing methods and training parameters.

V.CONCLUSION

Using Tesseract and OpenCV for Handwritten Text Recognition (HTR) is a difficult yet intriguing use of computer vision technology. It is possible to recognize and transcribe handwritten text from photos with a respectable level of accuracy by utilizing a variety of pre-processing techniques and training an OCR engine using a big and varied dataset of handwritten images.

HTR systems have the potential to be used in many different fields, such as document analysis, postal automation, and historical manuscript digitalization. Students can also use it to help them translate their handwritten notes and assignments.

The quality of the input photos, the amount and diversity of the training dataset, and the precision of the OCR engine, however, can all affect how accurate HTR systems are. To get the greatest results, it's crucial to properly adjust the pre-processing procedures and training parameters.

Overall, Tesseract and OpenCV may be used to build an effective and accurate HTR system that can recognize and transcribe handwritten text from photos. This system has the potential to be used in a variety of fields.

REFERENCES

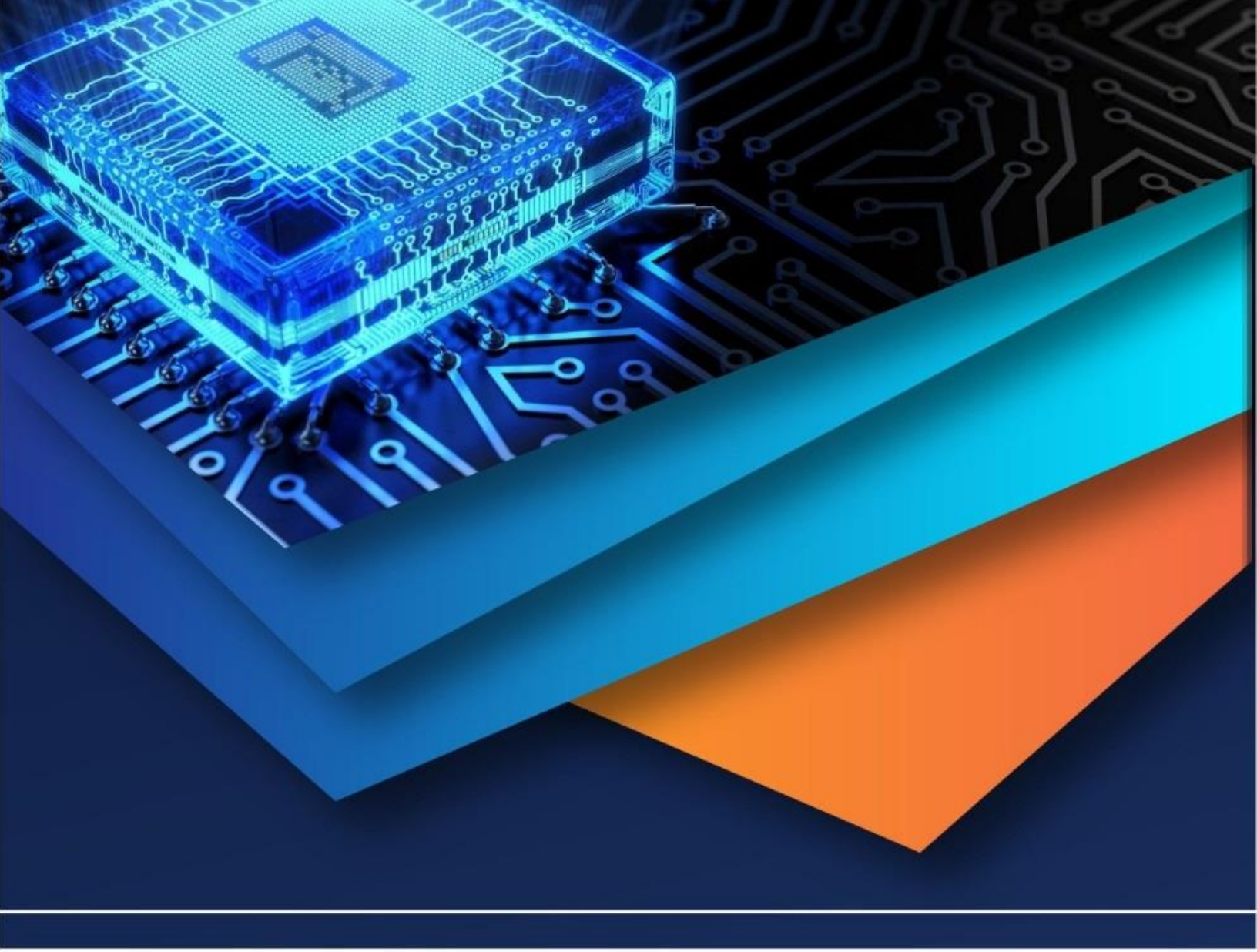
[1] Kaundilya, Chandni, Diksha Chawla, and Yatin Chopra. "Automated text extraction from images using OCR system." 2019 6th International Conference on Computing for Sustainable Global Development (INDIACom). IEEE, 2019.

[2] Mittal, Rishabh, and Anchal Garg. "Text extraction using OCR: a systematic review." 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA). IEEE, 2020.

[3] Zhang, Jian, et al. "Research on the text detection and extraction from complex images." 2013 fourth international conference on emerging intelligent data and web technologies. IEEE, 2013.

[4] Zhang, Honggang, et al. "Text extraction from natural scene image: A survey." Neurocomputing 122 (2013): 310-323.

[5] Saxena, Neeru, and Humera Parveen. "Text extraction systems for printed images: a review." International Journal of Advanced Studies of Scientific Research 4.2 (2019).



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY