International Journal of Research in IT and Management (IJRIM)
Available online at: http://euroasiapub.org
Vol. 12 Issue 12, December- 2022
ISSN(o): 2231-4334 | ISSN(p): 2349-6517 | Impact Factor: 8.106

Enhancing Text Extraction Accuracy from Images using Tesseract Algorithm: Insights and Implications

Chennakesh S[1], Dr. Anand Kumar[2]
Department of Computer Science
[1,2]Capital University, Koderma (Jharkhand)

Abstract
In the modern world, one of the areas of study and technology that is experiencing the most rapid expansion is image processing. A computer system that is capable of storing the information that is available in newspapers and other hard copy paper documents is in high demand; this demand is expected to continue. One of the most straightforward methods for storing textual information in computer systems is to scan the paper containing the information. Following that, it is possible to save it on the computer, and if necessary, modifications can be made to it. However, extracting text from the image that was obtained is a task that is fraught with difficulty. The Tesseract algorithm, which simplifies the process of extracting text from images, has been utilized in an effort to accomplish this extraction.
Keywords: *Tesseract Algorithm; Image; optical character recognition (OCR)*

Introduction

The majority of people in today's society find it more convenient to take their mobile phones with them rather than their laptops or desktops. As a result, new applications are required in preparation for the expanding world. Image recognition, often known as optical character recognition (OCR) **(Yang, 2020),** is a powerful tool that can turn printed text into text that can be edited. It is used for scanned documents. On the other hand, achieving the desired result is a lengthy process because of the variations in the characteristics, dimensions, and hues of the many characters. The use of Tesseract has been used in order to solve this issue. Open source optical character recognition (OCR) engine Tesseract was created by HP **(Vogtlin et al., 2021).**

International Journal of Research in IT and Management (IJRIM)
Available online at: http://euroasiapub.org
Vol. 12 Issue 12, December- 2022
ISSN(o): 2231-4334 | ISSN(p): 2349-6517 | Impact Factor: 8.106

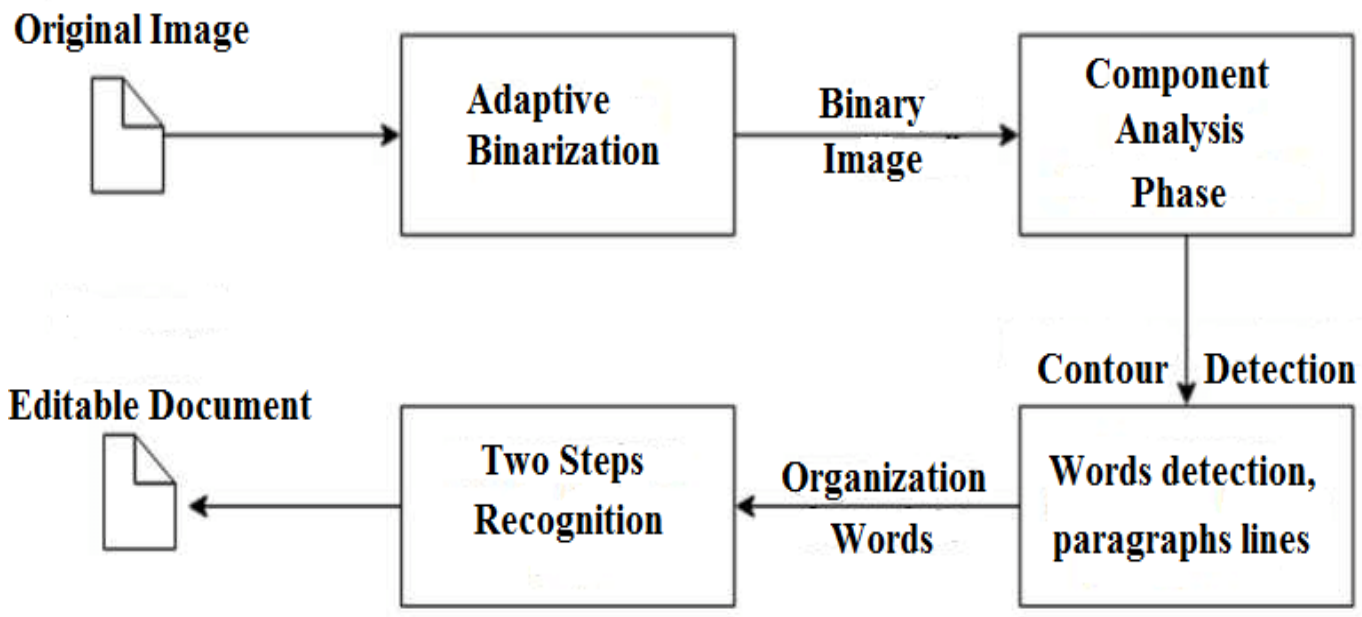


Figure 1. Flow chart of Tesseract OCR system

According to the flowchart of the OCR system that is presented below, the picture that was captured by the camera serves as the input image. Following the completion of the preprocessing step, this input image is subsequently sent to the Tesseract **(Ma et al 2021).** Essentially, Tesseract is the application's central processing unit. A piece of software that is utilized for the purpose of image processing and analysis is included in the open-source library known as Leptonica. A data set that has been trained is provided to the Tesseract engine. Following the completion of the processing, the Tesseract provides a printed text of the image that was taken as the output.**(Balasooriya, 2021).**

Methodology

The process of converting printed text into text that can be edited is known as optical character recognition (OCR), and it is a powerful tool that is utilized for scanned documents.**(Kaló, & Sipos, 2021).** Utilizing technology to discriminate between printed and handwritten text characters included inside photographs of paper documents is what this term refers to. The procedure begins with the examination of the text included inside a document, followed by the translation of that text into a code that may subsequently be utilized for the analysis of data. Tesseract is an open-source optical character recognition (OCR) engine that was initially developed by HP and is currently being maintained by Google. Without the need to de-skew the page, it is possible to identify a page that has been skewed with the assistance of Tesseract. First, a quadratic spline is applied to the baselines in order to fit them. Tesseract examines the text lines to determine whether or not they have a fixed pitch. Following that, it utilizes pitch to split the words down into individual characters or letters. Word recognition is the next step in the process. The purpose of this study is to develop an algorithm for the correction of skew angles that occur during the scanning of paper documents for the purpose of text segmentation.

International Journal of Research in IT and Management (IJRIM)
Available online at: http://euroasiapub.org
Vol. 12 Issue 12, December- 2022
ISSN(o): 2231-4334 | ISSN(p): 2349-6517 | Impact Factor: 8.106

During the initial phase of the process, an image is loaded into the system. For this particular photograph, you have the option of selecting it from the gallery or taking a live shot from a camera. When it comes to the output, the clarity of the image that is loaded has a considerable impact on both the quality and the accuracy of the output. Directly following the loading of the image, the second stage comprises doing an analysis of the image in order to assess the degree of clarity it possesses. It is possible that the output that is produced will be deficient in clarity and accuracy if it is determined that the image was recorded erroneously or that it was blurry.

In the succeeding stage, which comes after the completion of the picture processing, the image is cropped to the size that is required. This chance allows users to select a specific zone from which text extraction will take place, giving them the ability to control the process. Immediately following that, the procedure of extracting text from the region that was chopped takes place. There is a database that contains each and every character in the language that has been selected, and the characters that are found in the image are compared with the characters that are kept in the database. If a match is found, the character that corresponds to it is displayed as a component of the output. This occurs in the event that a match is found. It is necessary to repeat this process for each and every character that is present inside the region of the image that has been cropped. In the event that a character is unable to find a match, the pointer will go on to the character that comes after it in the sequence.

The employment of this technology ensures that text is extracted from images in a precise manner, which in turn makes the retrieval and processing of information less time consuming and more effective.

Results

A wide range of image types were used to test the program, and the results showed a significant amount of variation in terms of accuracy. The nature of the image, the layout of the website, and the computing capability of the mobile device are all factors that have been seen to influence the accuracy of the application. With the help of the printed text shown in the image below, one of the successfully produced digital texts was obtained.

International Journal of Research in IT and Management (IJRIM)
Available online at: http://euroasiapub.org
Vol. 12 Issue 12, December- 2022
ISSN(o): 2231-4334 | ISSN(p): 2349-6517 | Impact Factor: 8.106

Additionally, it has been noted that Tesseract functions most effectively when the text is clean black and the background is solid white. Additionally, it provides a high level of accuracy when the test is conducted horizontally and the text height is at least twenty pixels.

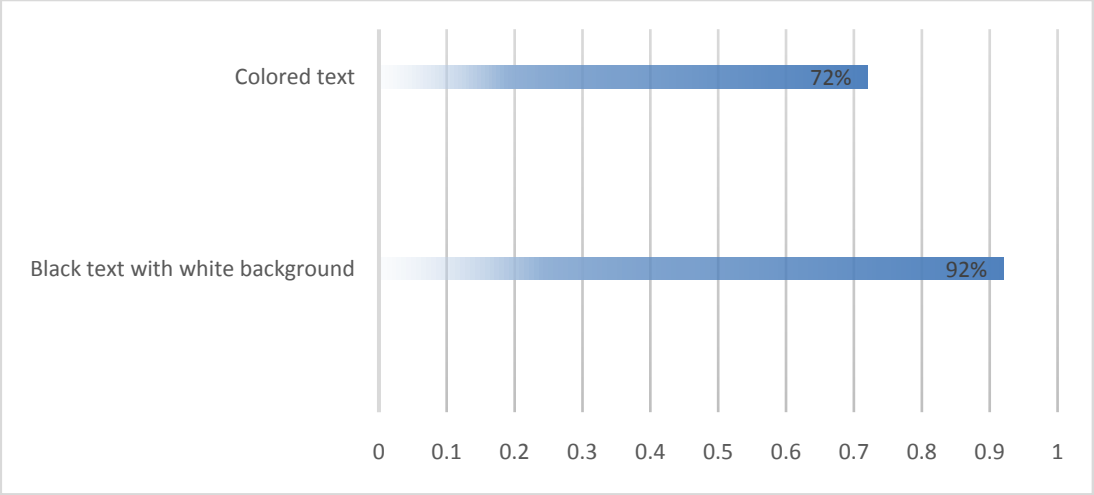
Types of Input Image	Total Words	Successful	Error	Accuracy %
Black text with white background	118	109	9	92%
Colored text	36	26	10	72%

Table 1. Accuracy observed for different types of images

When using this method, the text that is contained within the image that was shot or saved can be recognized with less effort and with greater precision. Tesseract has been utilized in this method in

International Journal of Research in IT and Management (IJRIM)
Available online at: http://euroasiapub.org
Vol. 12 Issue 12, December- 2022
ISSN(o): 2231-4334 | ISSN(p): 2349-6517 | Impact Factor: 8.106

order to acquire digital text from the printable text of a paper document that has been scanned. Processing, cropping, and loading of the image have all been completed. In order to improve the overall performance of the software, a post processor with two levels has been implemented. An increase in the number of processing algorithms could be incorporated into the OCR system.



Conclusion

When it comes to the successful implementation of image-based text extraction systems, our findings highlight how important it is to take these elements into consideration. Particularly noteworthy is the fact that Tesseract exhibits its best performance when the text is a clean black on a solid white backdrop, and when the text is horizontally aligned with a minimum height of twenty pixels. The presence of these parameters contributes to increased rates of accuracy in text extraction process. Tesseract's ability to recognize printed text from scanned paper documents is further demonstrated by our research, which demonstrates its usefulness. We were able to successfully get digital text from printable paper documents by incorporating Tesseract into our process. This was accomplished with minimal effort and increased precision. In order to facilitate text extraction with greater precision, the process of image processing, cropping, and loading proved to be extremely helpful. In addition, in order to improve the overall performance of the software, we created a post-processor that had two tiers. Through the addition of this feature, the capabilities of text recognition and extraction have been significantly enhanced. Taking a look into the future, we are aware of the possibility of further improvements being made by including new processing algorithms into the optical character recognition system. In a nutshell, the findings of our research highlight the significance of Tesseract as a valuable tool for extracting text from photographs, particularly in situations where printed text needs to be digitized for subsequent processing. We intend to expand the capabilities of image-based text extraction systems for a variety of applications in research and technology by first gaining a knowledge of the aspects that influence accuracy and then continuously refining our approaches.

International Journal of Research in IT and Management (IJRIM)
Available online at: http://euroasiapub.org
Vol. 12 Issue 12, December- 2022
ISSN(o): 2231-4334 | ISSN(p): 2349-6517 | Impact Factor: 8.106

Reference

Yang, S. (2020). *Uncertainty Quantification and Estimation on Medical Imaging Classification Tasks* (Doctoral dissertation, Concordia University).

Vögtlin, L., Drazyk, M., Pondenkandath, V., Alberti, M., & Ingold, R. (2021). Generating synthetic handwritten historical documents with OCR constrained GANs. In *Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part III 16* (pp. 610-625). Springer International Publishing.

Ma, T., Yue, M., Yuan, C., & Yuan, H. (2021, July). File text recognition and management system based on tesseract-OCR. In *2021 3rd International Conference on Applied Machine Learning (ICAML)* (pp. 236-239). IEEE.

Balasooriya, B. P. K. (2021). *Improving and Measuring OCR Accuracy for Sinhala with Tesseract OCR Engine* (Doctoral dissertation).

Kaló, Á. Z., & Sipos, M. L. (2021, January). Key-Value Pair Searhing System via Tesseract OCR and Post Processing. In *2021 IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI)* (pp. 000461-000464). IEEE.