iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** VIII **Month of publication:** August 2021

DOI: https://doi.org/10.22214/ijraset.2021.37377

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: 7.429
Volume 9 Issue VIII Aug 2021- Available at www.ijraset.com

Data Extraction from Images through OCR

Anurag Tiwari[1], Aditya Agrawal[2], Syed Bilal Ahmad[3], Divya Kant Singh[4], Aditya Pratap Singh[5]
[1, 2, 3, 4, 5]*Department of Information Technology, Babu Banarasi Das Institute Of Technology And Management*

Abstract: ***The paperwork used in maintaining various types of documents in our daily lives is tiresome and inefficient, it consumes a lot of time and it is difficult to maintain and remember the concerned documents. This project provides a solution to these problems by introducing Optical Character Recognition Technology (OCR) which runs on Tesseract OCR Engine. The project specifically aims at increasing data accessibility, usability and improving customer experience by decreasing the time spent to process, save, and maintain user data. Another objective of this project is to nullify the human error, which is huge in manual handling of data records, the software used in the solution uses certain techniques to minimize these errors. Optical Character Recognition (OCR) is used for extracting texts and characters from an image. This helps us in maintaining our records and data digitally and securely. In this project we are using the Tesseract OCR Engine which has high accuracy rates for clean images. We have implemented a web version of OCR which runs on TesseractJS; other JavaScript frameworks are also used. The outcome of the project is that it is able successfully to extract text and characters from the provided image using Tesseract OCR Engine. It is observed that for the high resolution images the accuracy is above 90%. This web based application is useful for small businesses as they don't have to install any extra software, all it needs is a file to be uploaded on an online interface making them able to access remotely. It will also help students to save notes and documents online which will make their important documents easily accessible on the web. This whole process is time and memory efficient.***

Keywords: OCR, Tesseract API, HTML, CSS, ReactJS, Node, Express, Heroku, Accuracy, Image processing, Express, Scalability, Accessibility

I. INTRODUCTION

OCR stands for optical character recognition. This technology lets you extract text from an image, it can be handwritten or printed text. In technical terms, OCR defines the process of electronically converting scanned images of handwritten, typed or printed text into machine-encoded text. Accuracy rates are measured by different approaches. OCR is considered as challenging because this technology is always improving. OCR process works in 3 steps:

A. The computer acquires an image of the document and it is submitted as input to the OCR engine. This step is often called image pre-processing in OCR in which image disturbances are repressed and specific image features are amplified.

B. The OCR engine is instructed to recognize certain shapes; it matches portions of images to these shapes. It uses the concept of feature extraction which extracts only selected features when input data is huge and ignores the useless chunk of information.

C. The OCR analysis takes the image in digital format and converts it into machine recognisable text format. It performs a technique named as post-processing in OCR; this ensures high efficiency of OCR results. It is an error correction technique which can identify not only words but also serial numbers and codes.

D. OCR solves real world problems as this software can be combined with a broad range of technologies. Here are a few examples of possible use cases including OCR software - Identification Processes in OCR, Marketing Campaigns with OCR, Payment Processes in OCR, Reverse Image Search, Google Translate, Captcha and many more.

E. The project aims to make a feasible Minimal Viable Product using this technology accessible to students like us and explore the widespread application of OCR in different areas of our lives through a web based application.

II. AIMS AND OBJECTIVES

The aim of this study is to discover how digital assets management reduces the problem of managing digital assets in one place without compromising the data quality and also how to share the metadata to the reliable third party.

1) *HTML, CSS:* HTML is what your browser understands. When we browse a webpage, we see html, which is similar to the bone. Html is what provides a web page structure and form. CSS (Cascaded Style Sheet) modifies the look of html data. CSS is like skin, texture. It gives color, width ,height, padding, margin, background to the html element. Main job of CSS is to give "STYLE" to the html element.

2) *JavaScript:* JavaScript is a scripting language that is mostly used to create interactive web pages. There are a lot of fantastic things you can accomplish with your website with it's help. There is no need to waste time compiling the code. JavaScript code can automatically execute in the browser window itself without compilations. It is quicker than the Java programming code.
3) *React Web:* Engineers may utilize React to fabricate gigantic web applications that can adjust information without reloading the page. Responds significant objective is to be speedy, adaptable, and simple to utilize. React.JS is just simpler to grasp right away. The component-based approach, well-defined lifecycle, and use of just plain JavaScript make React very simple to learn, build a professional web (and mobile applications), and support it.
4) *Node. Js:* Node.js is an open-source, cross-platform, JavaScript runtime environment. It is built on Chrome's V8 JavaScript engine, it lets developers run server-side scripts outside of the browser. With NodeJS, you can develop the backend of apps easily by integrating express server and mongoDB in it. It is occasion driven and has non-hindering I/O, making it ideal for planning web programs that are lightweight, proficient, and speedy. With its amazing and useful nature, Node.js has been a good playground for developers.
5) *Tesseract OCR Engine Library:* Tesseract is an open source text recognition (OCR) Engine, available under the Apache 2.0 license. It can be used for programs using an API to extract characters from images. It supports a wide variety of languages.
6) *Javascript:* Javascript is a scripting language which helps in dynamically control events on the web and helps in interacting with web components. It is an event-driven, functional and imperative language, it also has object-oriented and pure functions paradigms which makes up pretty much everything you need to work on the web.

III. SYSTEM REQUIREMENT

1) *VS Code:* Visual Studio Code is the best editor out in industry to develop applications of any scale, it the most widely used editor because of its amazing features like autocomplete, intellisense, support for tons of languages, smooth developer experience and much more. The editor's intelligence makes it a seamless and rich experience for developers to build applications and deploy it in production.
2) *Codesandbox:* Codesandbox is an online editor for collaborative development with our team and sharing it with anyone on the web. The main highlight of codesandbox is that it acts like a local server when you develop an application and you can share the app link instantly. The productivity boosts up by using this tool.
3) *Web Browser:* A web browser may be a software system application for retrieving and presenting data on the planet Wide internet. This method is hosted exploitation of the Apache domestic cat internet server and might be accessed via an online browser by getting into the computer address of the hosted JavaServer Page. All the user-system interactions are done through the net browser.
4) *Minimum Machine Requirements:* Intel Pentium 4 processor or later that's SSE2 capable to run a chrome instance, Browser Environment based on Microsoft Edge, Safari, or Chrome, Minimum 2GB of RAM, Recommended 8GB of RAM, Recommended OS Environment: Windows 7 or higher, Mac OS 10.12.6 or higher. Ubuntu 14.04 or higher.

IV. PROBLEM STATEMENT

Maintaining records, identity verification cards, and important documents is very difficult if done by paperwork, it consumes time, storage and compromises in efficiency and security of the overall system. These factors slow down the working process in popular industries like:

1) *Healthcare:* Medical service providers need to take and record the medical histories of their patients in one click. Most of the activities involved in this process are done manually as of today and to optimize storage the data is frequently deleted as it ages which results in loss at times.
2) *Banking:* This industry is mostly paper based; OCR has vital applications in terms of maintaining customer records, monthly statements and processes like signature verification. With the help of OCR, it makes the data management of banking systems secure and efficient.
3) *Traveling:* The tourism industry needs identity verification of customers for their smooth experience, all the processes like booking, check-in and passport verification will need OCR to reduce human error.
4) *Government And Legal System:* The legal system has a huge pile of paperwork and documents which they need to maintain with time; if done manually it would result in human error which will be a loss to both citizens and government. The accessibility of data and it's usability decreases with human hands.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VIII Aug 2021- Available at www.ijraset.com

V. PROPOSED SOLUTION

This project aims at solving these problems faced in the real world by implementation of an OCR System to optically convert a digitally captured image into machine readable text form which will help these sectors optimize on specific workflows and enable longevity of user records by enabling digital storage and compression techniques for physical entities.

This proposed solution will make use of the Tesseract Optical Character Recognition Engine being implemented through a JavaScript Port of the Tesseract API. The setup is supported by a Node.js server configuration and the User Interface is implemented as a React Web app. For the scope of this project, we'll be limiting it to some specific use cases and mainstream languages like ENG (US) and ENG (INDIA) to match the timeline.

VI. DESIGN APPROACH

An agile based approach will be followed with a weekly sprint to achieve milestones for the week. A Scrum master (Divya Kant Singh) will manage the design process while the team goes through the planned activity chart as discussed above. This will ensure us to focus on outcomes with short feedback loops and leave space to explore any new feature we would like to include later but was not discussed in this synopsis at the point of writing it.

We plan to keep the solution web based to allow for anyone to use the solution without much dependence on specific hardware or software. This will ensure a similar experience across devices of all shapes and sizes.

VII. WORKFLOW AND CODEBASE

Once the user opens the application, the front page of the app is open which asks the user to upload an image from their device or from the web. After successfully uploading the image, the app starts to extract the data which is in the image by implementing the Object Character Recognition algorithm. Once the processing is complete, the desired output is shown to the user. We have tried to make the interface of the app minimal and seamless to enhance the accessibility and user experience.

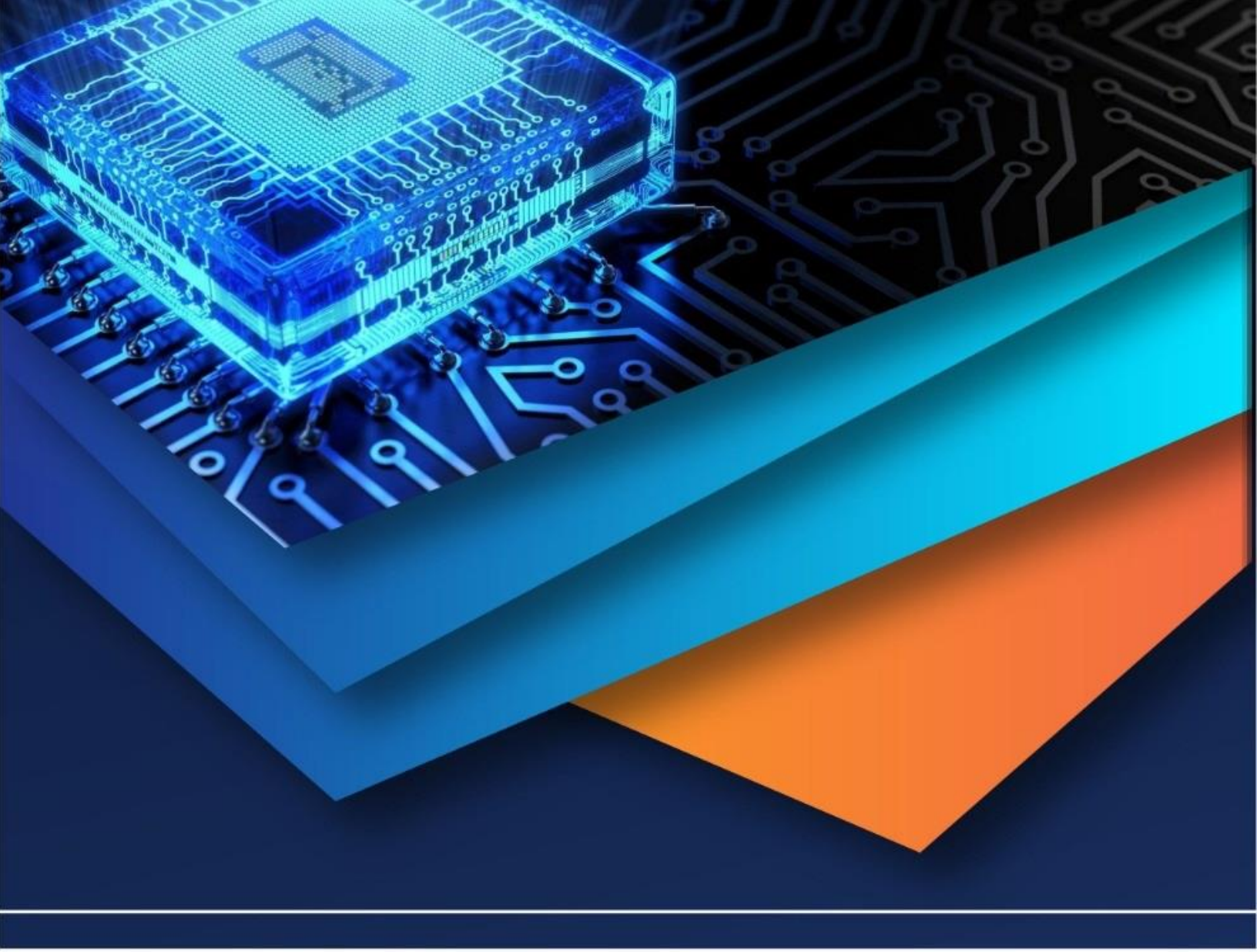
REFERENCES

BOOKS

[1] "Optical Character Recognition by Open Source. OCR Tool Tesseract: A Case Study" by Chirag Patel, Atul Patel(PhD) and Dharmendra Patel .
[2] "An Overview of the Tesseract OCR Engine" by R. Smith.
[3] Using Neural Networks to Create an Adaptive Character Recognition System © 2002, Alexander J. Faaborg Cornell University, Ithaca NY. 6.
[4] Digital Image Processing by A.Gonzales

WEBSITE

[1] http://www.ieeexplore.com
[2] http://www.stackoverflow.com



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY