Smart Document Management System using OCR

Mihir P [1], Ritik D [1], Vivek V [1], Shrinit P [1], Chandrayani Rokade [2]

[1]*Student, Department of Computer Engineering, Government College of Engineering, Nagpur, Maharashtra, India.*

[2]*Professor, Department of Computer Science and Engineering, Government College of Engineering, Nagpur, Maharashtra, India.*

Corresponding Author: shrinitpatil21i@gmail.com

Abstract: - Web applications are becoming more and more prevalent these days; it is normal for organizations to create specific applications for custom needs such as filling up forms and storing various kinds of information in a digital medium. Digitalizing these documents greatly contributes to the efficiency of organizations as it eradicates the need to maintain documents in their physical forms. The extra space created due to the absence of file storage equipment can be used in many beneficial ways, such as for additional equipment, a larger work area for employees, or any other use needed by the organization. This digitized information is stored on a web-based server and is accessible to anyone with proper authentication, this reduces the access time of these documents as all the information is searchable and eliminates the manual search time.

Key Words: — ***Digitalized Information, OCR, Authentication.***

I. INTRODUCTION

Nowadays, it is still very common in companies or organizations to have a large number of documents or forms that need to be filled manually by individuals and have to be kept as a record in huge stockpiles of files, be it typed or handwritten, for reasons such as to keep track of processes in their organization, to keep track of customers or vendors or stakeholders, etc. It becomes very difficult to store hard copies of such documents as their number keeps on growing every day. However, it is possible to store such information digitally through computer/web-based applications. PDFs have made it much easier to manage and share documents it is a reliable format to export documents or images that you want to share so that the information that you want to transfer is in the same manner as it was shared it helps you to maintain the format of documents pdfs are widely used and easily accessible on any device so nearly all institutions are digitalizing their document management system.

Manuscript revised May 10, 2022; accepted May 11, 2022.
Date of publication May 12, 2022.

This paper available online at www.ijprse.com

ISSN (Online): 2582-7898; SJIF: 5.59

As technology is evolving at a rapid pace, AI has enabled us to read information or text from the images and hence it is possible to digitalize such old documents so that we wouldn't have to worry about storing physical copies anymore.

We can also build a platform that will also help us search through documents using computers much quicker than the traditional way of searching through old documents by skimming. It is also possible for the platform to have a way to verify the legitimacy of the documents. This will be very helpful as it will save a lot of time and manpower behind these tasks and ensure that all the old documents get converted to a digitized form.

1.1 OCR and its working

Optical character recognition enables the transformation of various types of documents or images into analyzable, editable, and searchable data. During the last decade, researchers have used artificial intelligence and machine learning tools to automatically analyze handwritten and printed documents to convert them into electronic format. OCR processed digital files, such as receipts, contracts, invoices, financial statements and more can be searched from a large repository to find the correct document and viewed with search capability within each document and can be edited when corrections need to be made repurposed.

OCR analyses the patterns of light and dark pixels of an image that make up the letters and numbers to turn the scanned image into text. OCR systems need to identify characters in

various fonts, so rules are applied to help the system match what it sees in the picture to the right letters or numbers. For optimal OCR performance, it is important to scan the sharpest version of your document. Blurry text or marks on the copy can cause errors.

OCR consists of many steps: preprocessing, segmentation, feature extraction, classification, and recognition. The output of one step is the input of the next step. Preprocessing involves removing noise and handwriting variations. Some areas where OCR is used, including mail classification, banking, document reading, and email address recognition, require standalone handwriting and image recognition systems.

1.2 What Affects the Accuracy of OCR

Various Factors affect the accuracy of OCR Engines, some noteworthy and important factors which affect the Accuracy are:

Quality of the Original Source: It is recommended that the scanned image should have a resolution of at least 300 dpi. The higher the resolution, the better the results.

Image Characteristics: Bit depth and Image Binarization play a vital role in OCR accuracy, if the scanned image is scanned as grayscale or bi-tonal it significantly increases the accuracy of results. 50% brightness is recommended for scanning. Characters present in the Image should have sharp focus. Marked, mouldy, faded source, characters not in sharp focus negatively affects the identification of characters

Skewing: A text is called skewed when it is not properly aligned with the orientation of the page while text extraction is being performed. In general, perfectly horizontal text yield the best results. Pages should not be skewed prior to OCR so that the word lines are horizontal other than the mentioned factors, other factors such as inconsistent use of font faces and sizes, texts published before 1850, and using multilingual text on the page can lower OCR accuracy.

Table.1. Comparison between Tesseract and Google Vision:

Language	Books			Web		
	#Lines	N-CER [%]		#Lines	N-CER [%]	
		Tesseract	Google		Tesseract	Google
Arabic	946	14.0	4.8	4208	54.8	19.4
English	1000	1.0	0.6	4868	44.0	15.6
Hindi	1067	5.4	2.5	3726	49.3	20.6
Japanese	773	28.0	4.9	3256	57.5	17.1
Russian	864	1.7	1.2	3883	36.2	16.7

Tesseract vs Google Cloud Vision API (Walker et al. 2018)

1.3 Digital Signature

Signatures are generally a way to authenticate documents to verify that they belong to a genuine entity. A digital signature resembles a normal signature with the only difference being that a digital signature is a combination of numbers generated from a digital signature generating algorithm. The following fig illustrates how a digital signature is generated and validated:

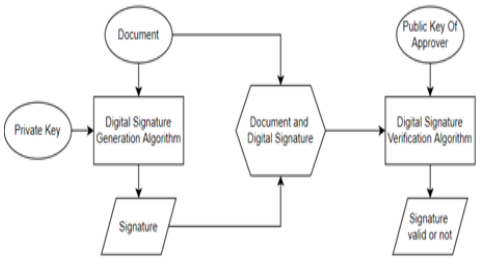

Fig.1. Working of Digital Signature

The whole process of digital signature generation and verification involves public and private keys of the authenticator. The digital signature generating algorithm accepts two parameters: Document and Private keys of the authenticator. Once the signature is generated it is then sent along with the document to the document generator. Once the document generator receives the document with digital signature it can then be verified by the document generator using the public key of the authenticator. The proposed system follows the same concept of digital signature for the authentication of documents.

1.4 Named-entity Recognition (NER)

Natural language processing is a subfield of linguistics, computer science and artificial intelligence concerned with the interactions between computers and human language. It includes techniques, which are used to analyze, synthesize or operate on the natural languages that humans use in day-to-day life. Named-entity recognition or NER is an NLP technique which helps us identify the named entities in a given text. Named entity consists of names or nouns that help us identify a person, an organization, a place, a product, etc. which can be given a proper name. It may be abstract or may have a physical form. Thus, NER is an information extraction technique which can help us extract important real-world objects from a text.

II. PROPOSED SYSTEM

A traditional file system is cumbersome, in that it does not allow users to easily edit files or send information to others. Paper files often cannot be edited directly, forcing users to make new copies to update old files. To distribute data on paper files, users must mail, fax, or scan the data. This proposed system allows users to scan, upload or take photo of document which then gets processed and text extraction is performed on it. After the text is extracted, it is placed in an editable pdf file using the bounding box coordinates. This file is stored in database and user is able to edit information fields directly, and as the information is stored digitally, it is already in a form that can be easily transmitted.

The UI of the system is simple to understand for the users so that it is easier to adopt the system for them and in turn, this would make the complete process effective and smoother. We also provide admin access to the system, so that the admin in the institution would be able to create multiple different users based on their needs. The admin user is able to see all the documents in the system regardless of whether they are approved or currently under the approval process. We can also allow the admin user to map certain initiators to certain approvers for further control over the process. Hence, the exact functioning of the application can be controlled by the admin user as per their requirements.

2.1 Google Cloud Vision OCR

Google Vision API is said to be consisting of 5 Major steps and every image it receives goes through these 5 stages. The steps are as follows:

*Text Detection***:** The first step is using a Conventional Neural Network (CNN)-based model to detect and localized lines of text and generate a set of bounding boxes.

Direction Identification: This step classifies the direction per bounding box. If necessary, some bounding box will be filtered out as it is erroneously detected as text.

Script Identification: This step identifies the script per bounding box. It is assumed that there is one 1 script per bounding box but allows multiple scripts per image.

Text Recognition: This is the core part of OCR which is recognizing text from images. It does not only include a character-based language model but also an Inception-style optical model and custom decoding algorithm.

Layout Analysis: This step includes determining the reading of order and distinguishing title, headers, etc.

2.2 Report labs

It is a library that helps you create documents in Adobe's Portable Document Format (PDF) using Python. It also provides high-end functionality like creating charts and data graphics in multiple formats like bitmap, vector, and PDF. The Report Lab library directly creates PDFs based on your graphics commands. There are no intervening steps. Additionally, report labs have the advantage that we can get input data in any format and make pdf as it is a python library. We have used report labs to generate pdf from the output of our Google Vision OCR. The output from OCR contains each word and each letter along with its bounding box coordinates, so using report labs each word is placed according to its bounding box coordinates. So, to make a digitalized and editable format of scanned image or document, report labs help to place the extracted text at its position in a new editable pdf.

2.3 Named-entity Recognition (NER)

The proposed system uses Natural Language Entity (NER) to extract important and unique keywords from the documents by which the document can be uniquely identified and stored in a database. These keywords extracted can be used to search a particular file and eases the work of managing documents. The keywords extracted can also be used for classification of documents on their type and content. This reduces the effort of classifying documents manually.

III. RESULTS AND CONCLUSION

Smart Documentation App is a platform which is proposed for legalization of documents using text extraction and digital signature. The heart of the project lies in the Google Cloud vision OCR which is used for text extraction from documents and digital signatures that are being used for authentication of these documents. The whole idea behind the proposed system was to reduce the risk and man power that was previously required for management of the physical documents.

Shifting from physical documents management system where there are risks of misplacement or damage to these documents to a digitalized management system would reduce such risks as these documents are being saved in a secured database with proper authentication and validations. the concept of Digital Signature Used in the proposed system makes sure that the documents are being verified by a legitimate or a genuine user. Through this proposed system we wanted to demonstrate how the idea of document digitalization can be a solution to many

problems that exists with the current system of managing documents in physical format.

REFERENCES

[1]. OCR Accuracy, Rose Holley, DLib, March 10, 2022
[2]. OCR Best Practices, Scholarly Commons, Illinois Library, January 17, 2022.
[3]. Google Vision Working, Edward Ma, Towards Data Science, December 12, 2021.
[4]. Natural Language Processing, Wikipedia, December 13, 2021.
[5]. Named-entity Recognition Wikipedia, February 18, 2022.
[6]. Report labs official Documentation: Report Lab, Report Lab PDF Library User Guide, Version 3.5.56, 2020.