A Comparative study on OCR Engines for Invoice Analysis

Shreya Ekande[1], Yash Chavan[2], Vaishnavi Gurav[3], Rajat Joshi[4], Vandana Rupnar[5]
[1,2,3,5]*Department of Computer Engineering MMCOE, Pune, India*
[4]*Emergeflow Technologies Pune, India*

***Abstract*—OCR is a technology used to extract text from images and documents via mechanical or electronic means. OCR engines have been developed into many kinds of domain-specific OCR applications. Thus, commercial off-the-shelf (COTS) OCR software packages have become powerful tools for quick development. This paper studies major COTS OCR software. In today's world where most of our mundane and repetitive tasks havebeen automated, still in a B2B or even B2C industry we have one task which is still majorly done with manual labor that is invoice analysis and bookkeeping. As a result, automating invoice analysis aids in the task automation.**

***Index Terms*—OCR, Text Detection, Table Extraction, Entity Extraction**

I.INTRODUCTION

Optical character recognition or optical character reader (OCR) is the electronic or mechanical conversion of images oftyped, handwritten or printed text into machine-encoded text, whether from a scanned document, a photo of a document, a scene-photo (for example the text on signs and billboards in a landscape photo) or from subtitle text superimposed on an image (for example: from a television broadcast).[1]

In the 2000s, OCR was made available online as a service (WebOCR), in a cloud computing environment, and in mobile applications like real-time translation of foreign-language signs on a smartphone. With the advent of smart-phones and smart glasses, OCR can be used in internet connected mobile device applications that extract text captured using the device'scamera. These devices that do not have OCR functionality built into the operating system will typically use an OCR API to extract the text from the image file captured and providedby the device.[2][3] The OCR API returns the extracted text, along with information about the location of the detected text in the original image back to the device app for further processing (such as text-to-speech) or display.

OCR engines have been developed into many kinds of domain-specific OCR applications, such as receipt OCR, invoice OCR, check OCR, legal billing document OCR. In recent years, the major OCR technology providers began to tweak

OCR systems to deal more efficiently with specific types of input. This strategy is called "Application-Oriented OCR" or "Customized OCR", and has been applied to OCR of license plates, invoices, screenshots, ID cards, driver licenses, and automobile manufacturing.

Thus, many commercial off the shelf (COTS) OCR software which can be further customised to application specific needs are now widely available. Google Cloud Vision, ABBYY FineReader, Adobe Acrobat, OmniPage, Tesseract OCR, Ama- zon Textract, Microsoft Azure Computer Vision OCR engine, ABBYY Flexicapture, Docparser, Kofax OmniPage, Microsoft office Document Imaging, Nanonets, Readiris, Rossum are frequently mentioned in Top OCR engines. We had the opportunity to experiment with the SDKs of several leading COTS OCR Packages such as: Google Cloud Vision, Tesseract, Amazon Textract, ABBY Flexicapture, Nanonets.

II.OCR SOFTWARES

A. Tesseract

Tesseract is a package contains an OCR engine - libtesseractand a command line program - tesseract. [2] Tesseract was originally developed at Hewlett-Packard Laboratories Bris-tol UK and at Hewlett-Packard Co, Greeley Colorado USA between 1985 and 1994, with some more changes made in 1996 to

port to Windows, and some C++izing in 1998. In 2005 Tesseract was open sourced by HP. From 2006 until November 2018, it was developed by Google. Tesseract has unicode (UTF-8) support and can recognize more than 100 languages "out of the box". Tesseract supports various image formats including PNG, JPEG and TIFF.

Tesseract supports various output formats: plain text, hOCR (HTML), PDF, invisible-text-only PDF, TSV and ALTO (the last one - since version 4.1.0). This project does not includea GUI application. Tesseract can be trained to recognize other languages. Developers can use libtesseract C or C++ APIto build their own application.[3] If you need bindings to libtesseract for other programming languages, please see thewrapper section in the AddOns documentation. Tesseract usesLeptonica library for opening input images (e.g. not documentslike pdf). It is suggested to use leptonica with built-in support for zlib, png and tiff (for multipage tiff).

Subjective opinion on using Tesseract is that it's very primitiveand cannot do feature extraction, rather read text line by line, to use it properly we would need to couple it with a deep learning model which would perform the segmentation.

B. Google Cloud Vision OCR

Google Cloud Vision OCR is part of the Google cloud vision API to extract text from images. 1) You essentiallysend an image (remote or from your local storage) to the Google Cloud Vision API. 2) The image is processed remotely on Google Cloud and produces the corresponding JSON formats with respect to the function you called. 3) The JSON file is returned as the output after the function is called.[1] Google Cloud Vision give out a JSON containing information about character positions,Just as for Tesseract, based on this information one could try to detect tables, but again, thisfunctionality is not built in and gets complicate

C. Amazon Textract

Amazon Textract makes it easy to add document text detection and analysis to your applications. Amazon Textract removes the complexity of building text detection capabili- ties into your applications by making powerful and accurate analysis available with a simple API. [4] Detect typed and handwritten text in a variety of documents, including financial reports, medical records, and tax forms, Extract text, forms, and tables from documents with structured data, using theAmazon Textract Document Analysis API, Specify and ex- tract information from documents using the Queries feature within the Amazon Textract Analyze Document API, Process invoices and receipts with the AnalyzeExpense API, Process ID documents such as driver's licenses and passports issuedby U.S. government, using the AnalyzeID API are its features.The Free Tier lasts for three months, and new AWS customerscan analyze up to Detect Document Text API: 1,000 pagesper month. [5] To develop products using Amazon Textract, First Set Up an AWS Account and Create an IAM User. Set Up the AWS CLI and AWS SDKs. Upload an image that contains a document to your S3 bucket. Code snip is available. Finally, it provides you with the JSON output for the operation.Subjective opinion on using AWS Textract is that it's a all-round tool which can do feature extraction, Table extraction and text, with a high enough accuracy, though it lags when it comes to handwritten text.

D. Nanonets

A cognitive capture automation tool for intelligent documentprocessing, Nanonets is an AI-based OCR programme. It is generally used to process ID cards, receipts, invoices, and other types of paperwork. Nanonets extracts pertinent data from unstructured data using cutting-edge OCR, machine learning, and deep learning approaches[7]. There is an API available for our usage, and it is a full software package. Simply unstructured invoices from numerous clients are uploaded by the user, but he can extract the fields he requires. After using the API, we get the extracted data in JSON format. Nanontesis paid. Despite having a powerful engine and being simple to use, the nanonet platform requires us to establish a model for processing, and if we were to diversify by using a newor different invoice type, the current model would fail. As a result, nanonet would have a lengthier workflow and would be generally unreliable for service providers in various industries. Even though Nanonets has a feature of table extraction, it's accuracy is low.

E. ABBYY Flexicapture

Similar to Nanonets, ABBYY flexicapture is an entire soft- ware for document recognition. It uses

NLP, Machine Learning and advanced recognition techniques for easy processing. ABBYY provides the output in various formats such as Excel, CSV, XML. Even though ABBYY Flexicapture provides goodtable extraction, it is not free.

II.COMPARISON

Based on the investigation on the five selected OCR tools, the comparisons are divided into several key features. The comparison of the features can be seen in Table 1.

IV.METHODOLOGY

The 5000+ OCR samples were obtained from various com- panies. The samples were divided according to whether the invoices were handwritten, UPI payment screenshots, system generated invoices and system generated but scanned invoices.We have limited our scope to just system generated invoices. The data was used as input for the above listed OCRs and results were manually cross checked for the comparative study.

Fig. 1. Invoice receipt example

V.CONCLUSION

Both Nanonets and Textract OCR tools are perfect for invoice analysis. When the cost per document is taken into account, Textract wins out because it is roughly 10 times less expensive than Nanonets.

Table- COMPARISON OF OCR SOFTWARES

OCR Software	Multi Language recognition	Multiple files conversion/bundled	Multiple pages conversion	Identification of Email	Hand-writing identification	Table Extraction	License	Output Format	Fee
Tesseract	NO	NO	NO	YES	NO	YES	Apache	Text	Free
ABBYY Flexicapture	NO	NO	NO	YES	YES	YES	Proprietary	CSV, XML	Paid
Google Cloud Vision OCR	YES	NO	NO	NO	YES	NO	Proprietary	JSON	Paid
Nanonets	YES	YES	YES	YES	YES	YES	Proprietary	JSON,CSV,Text	Paid
Amazon Textract	YES	YES	NO	NO	YES	YES	Proprietary	JSON,CSV	free tier

REFERENCE

[1] "Vision AI — Cloud Vision API — Google Cloud," Google Cloud. https://cloud.google.com/vision (accessed Nov. 03, 2022).

[2] tesseract-ocr, "GitHub - tesseract-ocr/tesseract: Tesseract Open-Source OCR Engine (main repository)," GitHub, Nov. 01, 2022. https://github.com/tesseract-ocr/ tesseract (accessed Nov. 03, 2022).

[3] tesseract-ocr, "Releases tesseract-ocr/ tesseract," GitHub. /tesseract- ocr/ tesseract /releases (accessed Nov. 03, 2022).

[4] " What is Amazon Textract? - Amazon Textract. https://docs.aws.amazon.com/textract/latest/dg/what-is.html (accessed Nov. 03, 2022).

[5] "Intelligently Extract Text Data with OCR - Amazon Textract Pricing - Amazon Web Services," Amazon Web Services, Inc. https://aws.amazon.com/textract/pricing/ (accessed Nov. 03, 2022).

[6] "tesseract-ocr-Revision458:/trunk/doc," tesseract-ocr-Revision458:/trunk/doc. https:// web.archive.org/web/20100916041654/http://

tesseract- ocr.googlecode.com/svn/trunk/ doc/ (accessed Nov. 03, 2022).

[7] Louis, S., Sonar, P., Kaul, P. (2022). An Image-based Intelligent System for Data Extraction. Asian Journal for Con- vergence In Technology (AJCT) ISSN -2350-1146, 8(2), 1-4. https://doi.org/ 10.33130/ AJCT. 2022 v08i02.001