Intelligent Information Retrieval: Techniques for Character Recognition and Structured Data Extraction

Akhil Chawla
Student
Electronics and Telecommunication Engineering
R.V College of Engineering
Bangalore, India

Aarushi Gupta
Student
Electronics and Communication Engineering
R.V College of Engineering
Bangalore, India

Mohana
Assistant Professor
Electronics and Telecommunication Engineering
R.V College of Engineering
Bangalore, India

K S Shushrutha
Associate Professor
Electronics and Communication Engineering
R.V College of Engineering
Bangalore, India

***Abstract*—The day-to-day activities of every corporation involve working with a huge amount of varying data formats such as those of work orders, techlogs, maintenance documents, etc. all of which are either vector or scanned PDFs. These activities involve long hours of manual work to extract the required data from these documents for further processing and becomes a costly affair for these organizations. Thus there is a huge scope for the development of a tool that provides intelligent optical character recognition and automates the process of extracting required information from these documents. This work contains a detailed analysis of end-to-end information extraction and proposes a high-quality information extraction tool. The pro- posed tool incorporates vital preprocessing required and a variety of methods for accurate data extraction based on the type of data. The prerequisite work provides an extensive insight into the technologies and presents its comparative analysis and performs the much needed capabilities check that can be utilized to further build on the intelligent information retrieval tool.**

***Index Terms*— Computer vision, Table Extraction, Character Recognition, Commercial OCRs**

I. Introduction

According to various estimates, the world generates 2.5 quintillion bytes of data every day and around 80% of this textual data is stored and shared in PDF format, which is the standard and the most widely used format of data storage. These PDFs include various kinds of structured and unstructured data including those from invoices, bills, receipts, forms, journal papers, emails, official communication letters, scanned PDFs, and many others, with unstructured data occupying more than 80% of this data produced. This quantity of unstructured data has an exponential growth with each day and many companies spend millions of dollars to analyze and interpret information for these documents and integrate it with the information management system used by them which is aimed to provide a huge productivity boost.

The business function of any organization is dependent on the processing of various template-based and non-template-based documents which is carried out manually in most cases as the current extraction procedures are highly prone to errors, especially in case of documents with complex structures, in which the contents of different fields are highly heterogeneous concerning the layout and require a manual supervision layer for every processed document accounting for redundant work. Therefore there has been a need for digital document data extraction tools for industrial use cases.

Several open-source tools such as OCRs are widely used in digital document processing but are only proven to be fit only for simple layout documents and otherwise often misinterpret non-textual information as text when dealing with complex documents, resulting in inefficient data extraction. Other issues with most OCRs include their dependence on high-clarity input images along with an adequately uniform background, which is difficult in cases of most scanned and handwritten documents. OCR's dependence on these among many other criteria for effective information extraction has been the reason for their non-acceptance in industrial use cases.

Therefore to provide an efficient information extraction method, this work proposes important preprocessing techniques to improve the accuracy and precision of OCRs and focuses on the use identifying various document layout elements such as text, table, and image area, which assist in their respective process of extraction and provide high accuracy results. The proposed work also caters to the frequent need for the pre analysis work of performing comparative analysis and capabilities check on the technologies. The work caters to the needs of the developers aiming to build upon these technologies.

II. Literature Survey

Optical character recognition (OCR) has been around for a while now, but its accuracy and its ability to derive meaningful data instead of giving out plain text have been areas of continuous research.

The research done in [23] explains how an OCR works in its various stages. This research aids in identifying the various shortcomings of common OCR system. The paper also discusses how those flaws can be addressed and how a better, future-ready OCR can be achieved. The findings of [23] give insights into the blocks an OCR comprises, which starts with a pre-processing stage and is followed by character recognition, which separates the individual characters on the captured image. This is where OCR systems are primarily distinguished as they can either examine resemblance with characters in the database or use their intelligent character recognition system.

Shortcomings of present-day OCRs have also been discussed in [23]. OCRs cannot currently differentiate text based on the size of the text on the scanned document. They also have an inaccurate judgment of line ends and no font type detection feature is available. The proposed solution explains that this can be avoided by including an additional layer of ”document analysis” that not only scans the document but also maintains the alignment, spacing, and font size.

Image enhancement techniques are discussed in [1] which improve the quality of images for human perception by removing noise, reducing blurring, increasing contrast, and providing more detail. It discusses spatial filtering operations, such as point processing which includes contrast stretching, global image thresholding, histogram processing, log transformations, and power law transformation; followed by mask processing, under which there are concepts around smoothing filters, sharpening filters, median filters, maximum filters, minimum filters, range filters; finally comes local thresholding. Furthermore, [1] also discusses the techniques and algebraic expressions around noise removal, skew detection/correction, page segmentation, character segmentation, image size normalization, and various morphological processes including erosion and dilation, opening and closing, outlining and thinning, and skeletonization which further enhances the image quality to ease the OCR process and help in acquiring better results and higher accuracy.

The work in [8] focuses on the development of a method for extracting text from document images. The strategy has several advantages. To begin, a new model for illumination adjustment based on Contrast Limited Adaptive Histogram Equalization (CLAHE) is proposed to improve the overall contrast of the objects present in the processed document image. Second, for text extraction, the Luminance algorithm is used to optimize grayscale conversion. Third, the Unsharp masking filter is used to enhance text details and edges. As a final enhancement step, the Otsu Binarization algorithm is used to clean and whiten the document background.

The authors of [17] created an offline OCR. This paper achieves greater than 90% accuracy and is ideal for skewing the image for validating methods. This paper presents a com- plete Optical Character Recognition (OCR) system for Calibri English characters. If the input image is not correctly aligned, it first corrects the skew of the image and then performs noise reduction from the input image. This process is followed by line and character segmentation, after which the characters are recognized by the recognition module. Furthermore, the developed method is computationally efficient and takes less time than other optical character recognition systems.

The work in [24] provided an overview of how machine learning techniques, such as the Keras algorithm, can be combined with image classification [19]. According to the author, text extraction can be accomplished by matching strings with invoice text. A section in [13] defines a method of “Text cleaning and Table detection” and “Recreation of Table and Storing into JSON and CSV format”. The usefulness of heading-based and serial number-based methods has been highlighted along with storing the same in a structured format as a template has been a major turning point for this project as it can be used in the data extraction of work orders which will majorly be done as heading based detection and extraction. The template storing method can be used for batch processing, which is again a major use case for the project.

The section “Analysis of OCR Errors” in [9] presents a study on the commonly occurring OCR errors which includes edit operations, including standard mapping, nonstandard mappings, edit distances, and string similarity based on the longest common sequence (LCS). It also discusses length effects such as word length, OCR token length, two-dimensional classification based on word lengths, and edit distances. The most frequently occurring error in OCR outputs is erroneous character positions and the issue of real-word vs. non-word errors.

From the above findings on OCR, we realize that analyzing the layout of the document will play an important role in the final accuracy we receive, and also it will help us in segmenting the document to accordingly process each section as the extraction process differs based on the content being extracted to get the best out of the OCR engines. Since tables form a major part of the project’s use case and they are the most difficult to detect and extract while maintaining the structure, this is the next focus of this literature review.

The authors of [20] present TAble Organization (TAO). It is a system for automatically detecting, extracting, and organizing information from tables in PDF documents. TAO detects and extracts table information from documents using processing based on the k-nearest neighbor method and layout heuristics. This system creates an enriched representation of the data extracted from PDF tables. TAO’s performance is comparable to that of other table extraction methods, but it overcomes some related work limitations and proves to be more robust in experiments with a variety of document layouts. Based on the understanding from [14], intending to keep the overhead as minimum as possible, the idea was to use a method that solely relies on computer vision techniques.
[2] proposes a methodology for doing exactly that. According to their analysis of this approach, the obvious advantages of this approach are that the results are completely interpretable; and the algorithm is general enough to apply to a wide range of tabular images. And, the obvious disadvantages are that it is not (in its current form) a learning-based approach, so the

results are not the best they can be; and a generalized approach can be too vague at times, necessitating extensive pre/post-processing. Whereas contrary to one of the disadvantages mentioned, this generalized approach can be tailored to work better in specific scenarios such as improving OCR accuracy can be achieved by training a custom OCR model; using a combination of image transformation techniques can improve OCR accuracy; creating a cleanup heuristic for extracted tables can aid in the improvement of results; contextual understanding of tabular data (via metadata or prior knowledge of the data being extracted) can aid in the improvement of the post-processing pipeline.

In the presence of complex layouts with large white spaces, the deterministic approach fails to detect tables. The results show that the approach in [7] outperforms others, with correct detections increasing from 44 percent to 60.5 percent.

Moving ahead on the lines of Faster R-CNN, TableBank is a new image-based table detection and recognition dataset built with novel weak supervision from online Word and Latex documents presented in [15]. Existing research for image-based table detection and recognition typically fine-tunes pre-trained models on out-of-domain data with a few thousand human-labeled examples, making real-world applications difficult to generalize. We built several strong baselines using state-of-the-art models with deep neural networks on TableBank, which contains a lot of high-quality labeled tables. According to their findings, models perform well in the same domain. The inference can be drawn that the visual appearance of tables from various types of documents varies. Therefore, with small- scale training, we cannot simply rely on transfer learning techniques to obtain good table detection model data.

Enhancing and trying to obtain the best out of Faster R- CNN, [26] proposes a novel and improved algorithm based on the Faster R-CNN framework and the Faster R-CNN algorithm with skip pooling and contextual information fusion. Based on Faster R-CNN, this algorithm can improve detection performance under special conditions. The enhancement is divided into three parts where the first part adds a context information feature extraction model after the convolutional layer's conv5 3; the second part adds skip pooling so that the former can fully obtain the object's contextual information, especially when the object is occluded and deformed; and the third part replaces the region proposal network (RPN) with a more efficient guided anchor RPN (GA-RPN), which can maintain the recall rate while improving detection performance.

The researches of [4] proposed a proposal generation method that generates more proposals with higher intersection over union (IoU) with ground truth boxes than greedy search approaches that can better envelop entire objects. [29] proposed a fast-matching algorithm that can handle noisy image exemplar localizations and robustly matches region proposals with massive exemplars in terms of appearance and spatial context. One-stage algorithms are faster than two-stage algorithms in terms of detection speed, but the popular version of the former is better in terms of detection accuracy.

After analyzing the different approaches for extraction and detection, one comes to think that the process varies when dealing with different types of input documents. The project's use case consists of all types of documents therefore there is first a need to classify the documents into different categories. The work in [24] serves this purpose. The paper suggests an automated approach to categorize invoices as handwritten, machine printed, or receipts the proposed method is based on feature extraction with the deep convolutional neural network AlexNet. The features are classified using a variety of machine learning algorithms, such as Random Forests and K-nearest neighbors (KNN), as well as Naive Bayes. In the experiments, various cross-validation approaches are used to ensure the effectiveness of the proposed solution. The KNN achieved the best classification result of 98.4 percent (total accuracy). Such near-perfect performance allows the proposed method to be used in practice as a preprocess for OCR systems or as a standalone application.

With the aim of obtaining maximum accuracy, since the accuracy factor is of real importance in aircraft maintenance documents, review was done on papers dealing with other sectors requiring such a high degree of accuracy. The paper in [19] is on the topic Information Extraction from Text Intensive and Visually Rich Banking Documents, which also has almost no scope for error. Deep learning algorithms resulted in a 10% improvement on the IE subtasks, according to the experiments. The addition of word positional features resulted in a 3-percentage point improvement in some specific information fields. Similarly, their auxiliary learning experiments yielded a 2-percentage point improvement in some information fields associated with the specific transaction type detected by our auxiliary task. The integration of the information extraction system into a real-world banking environment reduced cycle time significantly. The impact of using different neural word representations (such as FastText, ELMo, and BERT) on IE subtasks (specifically, named entity recognition and relation extraction stages), word positional features on document images, as well as auxiliary learning with other tasks is being studied. The article proposes a new graph factorization-based relation extraction algorithm to solve the complex relation extraction problem where the relationships within documents are n-ary, nested, document-level, and previously indeterminate in quantity.

In response to the need for automatic layout analysis of Chinese academic papers in [28] and to address the problem of incomplete analysis of existing layout elements, this paper proposes a method of recognizing layout elements in Chinese academic papers based on Mask R-CNN. This method has been empirically proven to work with an accuracy of up to 89.3 percent for effective recognition and precise location of nine layout elements in academic papers, including headers, titles at various levels, main body, figures, tables, formulas, and references. As a result, the model can better meet the needs of practical application scenarios and has significant application value as a foundation for document information extraction, layout reconstruction, quality evaluation, and other applications. To begin, using data acquisition and manual

annotation, a layout image data set of Chinese academic papers was created. Following that, the RPN network was improved using a weighted anchor box generation mechanism based on the traditional Mask R-CNN model architecture. Finally, a model for recognizing layout elements in Chinese academic paper layout images was created.

III. OPTICAL CHARACTER RECOGNITION

By leveraging automated data extraction and storing capabilities, optical character recognition (OCR) technology is an effective business process that saves time, money, and other resources.

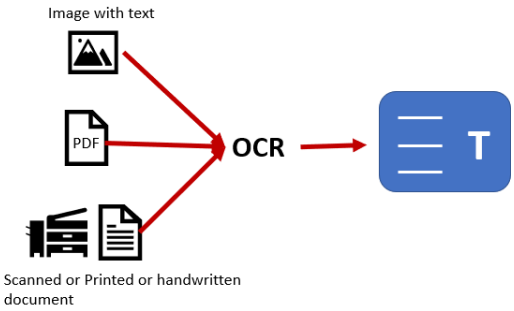

Fig. 1 Optical Character Recognition

Text recognition is another term for optical character recognition (OCR). OCR software extracts and re-purposes data from scanned papers, camera photos, and image-only PDF files as shown in Figure 1. OCR software extracts letters from images, converts them to words, and then sentences, allowing access to and alteration of the original material. It also eliminates the necessity for data entering by hand.

OCR systems turn physical, printed documents into machine-readable text using a mix of hardware and software. Text is typically copied or read by hardware, such as an optical scanner or dedicated circuit board, and then advanced processing is handled by software.

OCR software can use artificial intelligence (AI) to accomplish more complex methods of intelligent character recognition (ICR), such as distinguishing languages or handwriting styles. OCR is most typically used to convert hard copy legal or historical documents into PDF documents that users may edit, format, and search as if they were generated with a word processor.

A. Working of optical character recognition

The idea behind OCR is straightforward. However, due to a variety of factors such as font variety and letter formation methods, its implementation can be quite difficult. When non-digital handwriting samples are used as input instead of typed writing, for example, an OCR implementation can become exponentially more complex.

The entire OCR process consists of a series of steps with three main goals: image pre-processing, character recognition, and post-processing the specific output. Here, we will demonstrate how optical character recognition works and explain the basic steps of OCR technologies

1. Document Scanning

The first step in OCR is to connect to a scanner and scan the document. Because scanning the document standardizes the inputs, it reduces the number of variables to account for when developing OCR software. Furthermore, by ensuring perfect alignment and sizing of the specific document, this step specifically improves the efficiency of the entire process.

2. Image Refinement

The optical character recognition software improves the elements of the document that must be captured in this step. Any imperfections, such as dust particles, are removed, and edges and pixels are smoothed to produce a clean and clear text. This step makes it easier for the programme to capture and clearly "see" the words as they are entered, without smudges or irregular dark areas.

3. Binarization

The refined picture document is then turned into a bi-level document image that only contains black and white colours, with black or dark areas designated as characters. Simultaneously, white or light areas are designated as backdrop. This stage tries to apply segmentation to the document in order to readily distinguish the foreground text from the background, allowing for optimal character recognition.

4. Character Recognition

The black areas are further processed in this step to identify letters or digits. An OCR typically focuses on one character or block of text at a time. Character recognition is accomplished through the use of one of two algorithms:

4.1. Recognition of patterns

The pattern recognition algorithm involves inserting text into the OCR software in various fonts and formats. After that, the modified software is used to compare and recognize the characters in the scanned document.

4.2. Feature detection

Through the feature detection algorithm, OCR software applies rules considering the features of a certain letter or number to identify characters in the scanned document. Examples of features include the number of angled lines, crossed lines, or curves used for comparing and identifying characters.

Simple OCR software compares the pixels of every scanned letter with an existing database to identify the closest match. However, sophisticated forms of OCR divide every character into its components, such as curves and corners, to compare and match physical features with corresponding letters.

5. Verifying the Precision

Following successful character recognition, the results are cross-referenced using the internal dictionaries of the OCR software to ensure accuracy. Measuring OCR accuracy is accomplished by comparing the output of an OCR analysis to the contents of the original version.

There are two common methods for determining the accuracy of OCR software:

- The number of characters correctly detected at the character level.
- The number of words correctly recognized at the word level.

In most cases, 98-99 percent accuracy at the page level is considered acceptable. This means that the OCR software should correctly identify 980-990 characters on a page of around 1,000 characters.

B. Advantages of optical character recognition

The fundamental advantage of optical character recognition (OCR) technology is that it streamlines data entry by allowing for simple text searches, modification, and storing. OCR enables organizations and people to keep files on their PCs, laptops, and other devices, guaranteeing that all paperwork is always available.

The following are some of the advantages of using OCR technology:

1. Increased efficacy and efficiency

   Simple manual document processing costs about \$6-8 per document on average. For more complex documents, the average cost per document can range from $40 to $50. More than 70% of businesses would fail within three weeks if they lost all of their paper-based records in a fire or flood.

   Intelligent document processing, according to industry experts, ushers in a slew of significant changes.

   - Reduce the risk of errors by at least 52%.
   - Reduce the cost of manual document processing by 35%.
   - Reduce the amount of time spent on document-related tasks by 17%.
   - Reduce document processing times by 50\% to 70%.
   - Reduce operating costs by 30% YOY
   - Cut document verification time by 85 percent.
   - Reduce the entire financial aid application process from 6 weeks to a few days.
   - Attain a 99 percent accuracy rate.

2. Increased efficacy and efficiency

   The impressive accuracy rate of an intelligent document processing solution makes it the ideal solution for handling any compliance-related document or those containing sensitive information such as personally identifiable information (PII) or health records. Because IDP eliminates the need for humans to open, review, or handle any of the data included documents, it reduces the risk of sensitive information being exposed to third parties. Furthermore, IDP can help to streamline and improve the accuracy of regulatory reporting.

3. Enhanced data quality and usability

   On average, 80% of an organization's data is "dark data" - meaning it's locked in emails, text, PDFs and scanned documents. Using RPA and AI-based tools, IDP unlocks the value of dark data by transforming into high quality, structured data that is primed for analysis.

   As the experts at Mckinsey explain, "by combining the data derived from paper documents with the wealth of digital data already available, a comprehensive data landscape can be established, significantly enhancing data evaluation and analytics possibilities."

4. Encourages and expands automation

   Intelligent document processing, in conjunction with workflow management tools, is a powerful enabler of end-to-end process automation. It facilitates the integration of various systems involved in automating complex business processes and achieving hyper automation.

   Furthermore, cognitive technologies such as RPA and AI require structured, high-quality data from which to "learn" and operate. Intelligent document processing optimizes data for RPA/AI consumption by transforming unstructured data found in documents into streams of cleaned, structured data.

C. Open-Source OCR Tools

1. Keras-OCR

The Keras CRNN implementation and the published CRAFT text detection model have been slightly polished and packaged. It offers a high-level API for building a text detection and OCR pipeline.

2. Tesseract

Tesseract was created between 1985 and 1994 at Hewlett-Packard Laboratories Bristol and Hewlett-Packard Co, Greeley Colorado, with some changes made in 1996 to port to Windows and some C++izing in 1998. HP released Tesseract as open source in 2005. Google worked on it from 2006 to November 2018.

Python-tesseract is a Python-based optical character recognition (OCR) tool. In other words, it will recognize and "read" text embedded in images. Python-tesseract is an OCR engine wrapper for Google's Tesseract. It can also be used as a standalone tesseract invocation script because it can read all image types supported by the Pillow and Leptonica imaging libraries, such as jpeg, png, gif, bmp, tiff, and others. Furthermore, when run as a script, Python-tesseract will print the recognized text rather than writing it to a file.

3. EasyOCR

EasyOCR is a Python module for text extraction from images. It is a general OCR that can read natural scene text as well as dense text in a document. It currently supports over 80 languages and is growing. Jaided AI, a company that specializes in Optical Character Recognition services, created and maintains the EasyOCR package. Python and the PyTorch library are used to implement EasyOCR. If you have a CUDA-capable GPU, the underlying PyTorch deep learning library can significantly improve text detection and OCR speed. EasyOCR currently only supports OCRing typed text. They also intend to release a handwriting recognition model later.

D. Conclusion

OCR Prediction is not only affected by the model, but also by a variety of other factors such as image clarity, greyscale, hyperparameter, weightage, and so on. Tesseract does well with high-resolution images. Dilation, erosion, and OTSU binarization are morphological operations that can help improve pytesseract performance. EasyOCR is a lightweight model that performs well for receipt or PDF conversion. It produces more accurate results with organized texts such as pdf files, receipts, and bills. Keras-OCR is a tool for image recognition. Keras-ocr produces good results when text is contained within an image and its fonts and colours are disorganized.

Though there are no hard and fast rules, we can consider the three points listed above when selecting an OCR tool.

IV. TABLE EXTRACTION TOOLS

A. Introduction

Most of the useful data comes in the form of tables which are locked away if the table is in a PDF or a scanned image. Tables are used to organize data that is too detailed or complicated to be adequately described in the text, allowing the reader to see the results quickly. They can be used to highlight trends or patterns in data and to make a manuscript more readable by removing numerical data.

Table extraction in its own is a challenge to address which separate set of libraries have been developed apart from the general OCR engines. Maintaining a table structure and recreating it after text extraction while dealing with continuously growing number of table structures. We have discussed and analyzed few such Python libraries.

1. pdf-table-extract

This library examines a PDF page for well-defined table cells and extracts the text in each cell. JSON, XML, and CSV lists of cell locations, shapes, and contents, as well as CSV and HTML versions of the tables, are among the outputs. This utility was originally designed to read the tables in ST Micro's datasheets and is intended to be the first step in automatically processing data in tables from a PDF file. Numpy and poppler are required by the script (pdftoppm and pdftotext)

2. Tabula

Tabula or tabula-py is a simple Python wrapper for tabula-java that can read PDF tables. You can read PDF tables and convert them to DataFrames in Pandas. tabula-py can also convert PDF files to CSV/TSV/JSON files.

3. Pdfplumber

This library is used to search a PDF for information about each text character, rectangle, and line, as well as table extraction and visual debugging It works best with machine-generated PDFs as opposed to scanned PDFs. It's based on pdfminer.six.

4. PDFTables

The Sensible Code Company created PDFTables. When converting a PDF, it employs an algorithm that examines the structure of the PDF. It recognizes the spacing between items to identify the rows and columns, similar to how your eye does when scanning a page.

5. Camelot

The previous libraries either produce good results or fail miserably. There is no middle ground. This is ineffective because everything in the real world is hazy, including PDF table extraction. As a result, ad-hoc table extraction scripts for each type of PDF table are created. Camelot was designed to provide users with complete control over table extraction. If the default settings do not produce the desired results, you can tweak them to get the job done.

B. Comparative Analysis

The purpose of this research is to compare the previously mentioned open-source libraries and tools as shown in Figure 2. The extraction was tested on different layouts as mentioned in the first column of the table and feedback is given on the performance of each library on that structure.

Camelot is clearly superior to other open-source alternatives available as it gives accurate results in almost all cases mainly due to availability of customization in Camelot.

V. COMMERCIAL OCR TOOLS

1. Google Doc AI

Google Document AI creates pretrained models for extracting information from documents using computer vision, optical character recognition (OCR), and natural language processing (NLP). Google DocAI offers a wide range of parsers for various industries. Google's Lending DocAI and Procurement DocAI can assist organizations in processing large volumes of documents while minimizing processing time. DocAI also includes generic parsers such as OCR and form parsers that can be used to structure data and easily extract values. These parsers are housed in a unified dashboard, from which they can be tested by directly uploading a document into the console.

2. Amazon Textract

Amazon Textract is a machine learning (ML) service that extracts text, handwriting, and data from scanned documents automatically. To identify, understand, and extract data from forms and tables, it goes beyond simple optical character recognition (OCR). Textract reads and processes any type of document using machine learning, accurately extracting text, handwriting, tables, and other data with no manual effort. Whether you're automating

Document Characteristics / Tools	Tabula	Camelot	pdfplumber	pdftables	pdf-table-extract (pte)
Header text is vertical, columns span multiple cells.	doesn't output all the header text	puts vertical headers in reverse order	messes up header text	output unusable, merged columns	gets the table out as is
Columns spans multiple cells	moves some headers on the top-right to the left	gets them in the correct cells	gets the table out as is	output unusable, merged columns	gives extra columns
The table is rotated counter-clockwise.	output is unusable	gets the table out as is	output unusable	output unusable	doesn't account for table rotation
There are two tables on a single page.	output is unusable	gets the tables out as they are	doesn't identify two tables and output is unusable	doesn't combine multi-line rows	output unusable
There are two tables on a single page.	output is unusable	gets the tables out as they are	doesn't identify two tables and output is unusable	output unusable, merged columns	detects one table and merges first row with header
Two columns don't have any values	output is unusable	gets the tables out as they are	output unusable, merged columns	output unusable, merged columns	output unusable

Fig. 2 Open-Source Tools Comparative Analysis

	Google Doc AI	Amazon Textract	Nanonets	Docparser
	doc.ai	Amazon Textract	Nanonets	docparser
Text recognition	easily recognize text from even unstructured documents	supports grouping text through NLP, provide an accuracy of about 90+%	fetches the data with 95%+ accuracy.	accuracy of 90%+
Input formats	PDF, GIF and TIFF data formats.	JPEG, PNG, PDF, and TIFF formats	DOC, JPEG, PDF, and XLSX/XLS.	DOC, JPEG, and PDF.
Image quality	operate on any image quality	with moderate to HD image quality.	can easily handle handwritten text, low-resolution images, images with varying fonts and sizes, shadowy text, blurred images	moderate to high-level HD images.
Data extraction technology	cloud-based processing with AI integration	AI & machine learning (ML) service for extracting handwritten text	technologies such as AI and ML with 95+% precision.	zonal OCR
Template dependency	95+% of text recognition accuracy	does not require unnecessary templates	does not require any template	problems handling unknown templates.
New document training	allows developers to train and deploy using inputs such as target schema	not possible to train the software for a new document type	can train the doctype as per the requirements	custom PDF parser
Key-value pair and table extraction	software can train itself for a new document type	90%+ success rate	supports key-value structure and table extraction	can handle key-value pair
Use-case specification	consists of innumerable benefits where you can access the data from scanned documents using data capturing techniques through NLP and computer vision.	amazing option in financial services, the public sector, and life sciences.	most effective with financial and accounting documents	most effective in cases involving purchase orders, invoices, and bank statements
Pre-trained APIs	The software allows functionalities like parsers, solutions and tools through unified API	various pre-trained APIs	free version of pre-trained APIs to build their own custom deep learning models.	user can work on Rest APIs

Fig. 3 Commercial Tools Comparative Analysis

loan processing or extracting information from invoices and receipts, you can quickly automate document processing and act on the information extracted. Textract can extract data in minutes rather than hours or days. You can also use Amazon Augmented AI to add human reviews to your models and validate sensitive data.

3. Nanonets

Nanonets have been able to commercialize an OCR pipeline by utilizing it not only for character recognition but also for object detection and classification. It includes intelligent structured field extraction that is automated. It works well with a variety of languages. Text in the wild performs well, Train on your own data to make it work for your use-case, Continuous learning, and no need for an in-house development team.

4. Docparser

Docparser is a cloud-based document data extraction solution that assists businesses of all sizes in extracting data from PDFs, Word documents, and image files. Docparser can extract

data fields such as shipping address, purchase order number, and date to put them in a tabular format and move information to where it belongs by automating the document-based workflow.

Capability Check

The analysis in Figure 3 is based on various capability checks of the previously mentioned commercial OCRs in the market.

The areas of study include text recognition performance, input formats supported, the image qualities it can deal with, the data extraction technology behind them, how dependent it is on the meta data extracted or the template, capability to be altered based on new document training, accuracy on key value pairs and table detection, the use cases it is built for and how do the APIs function.

Most of them excel in a particular area, whereas fail in the others. Amazon Textract, on the other hand, seems to do an acceptable job in all the fields as in can deal with moderate to HD quality documents, it does not require unnecessary templates and has a 90%+ success rate in the most difficult area which is key value pairs and table extraction.

REFERENCES

[1] Yasser Alginahi. "Preprocessing techniques in character recognition". In: Character recognition 1 (2010), pp. 1–19.

[2] Saumya Banthia, Anantha Sharma, and Ravi Mangipudi. "TableZa–A classical Computer Vision approach to Tabular Extraction". In: arXiv preprint arXiv:2105.09137 (2021).

[3] Dipali Baviskar, Swati Ahirrao, and Ketan Kotecha. "Multi-layout Unstructured Invoice Documents Dataset: A dataset for Template-free Invoice Processing and its Evaluation using AI Approaches". In: IEEE Access 9 (2021), pp. 101494–101512.

[4] Gong Cheng et al. "High-quality proposals for weakly supervised object detection". In: IEEE Transactions on Image Processing 29 (2020), pp. 5794–5804.

[5] Gong Cheng et al. "Learning rotation-invariant and fisher discriminative convolutional neural networks for object detection". In: IEEE Transactions on Image Processing 28.1 (2018), pp. 265–278.

[6] Deng-Ping Fan et al. "Camouflaged Object Detection". In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog- nition (CVPR). June 2020.

[7] Azka Gilani et al. "Table detection using deep learning". In: 2017 14th IAPR international conference on document analysis and recognition (ICDAR). Vol. 1. IEEE. 2017, pp. 771–776.

[8] Abdeslam El Harraj and Naoufal Raissouni. "OCR accuracy improvement on document images through a novel pre-processing approach". In: arXiv preprint arXiv:1509.03456 (2015).

[9] Adam Jatowt et al. "Deep statistical analysis of OCR errors for effective post-OCR processing". In: 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL). IEEE. 2019, pp. 29–38.

[10] Dong-Kyo Jeong et al. "Mask-RCNN based object segmentation and distance measurement for Robot grasping". In: 2019 19th International Conference on Control, Automation and Systems (ICCAS). IEEE. 2019, pp. 671–674.

[11] Gong Jian et al. "Ship target detection based on infrared polarization image". In: Spectroscopy and Spectral Analysis 40.2 (2020), pp. 586–594.

[12] Alan Jiju, Shaun Tuscano, and Chetana Badgujar. "OCR text extraction". In: International Journal of Engineering and Management Research 11.2 (2021), pp. 83– 86.

[13] Venkata Naga Sai Rakesh Kamisetty et al. "Digitization of Data from Invoice using OCR". In: 2022 6th International Conference on Computing Methodologies and Communication (ICCMC). IEEE. 2022, pp. 1–10.

[14] Aditya Kekare et al. "Techniques for Detecting and Extracting Tabular Data from PDFs and Scanned Documents: A Survey". In: Tabula 7.01 (2020).

[15] Minghao Li et al. "Tablebank: Table benchmark for image-based table detection and recognition". In: Proceedings of The 12th language resources and evaluation conference. 2020, pp. 1918–1925.

[16] Weihong Ma et al. "Joint layout analysis, character detection and recognition for historical document digitization". In: 2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR). IEEE. 2020, pp. 31–36.

[17] Mujibur Rahman Majumder et al. "Offline optical character recognition (OCR) method: An effective method for scanned documents". In: 2019 22nd International Conference on Computer and Information Technology (ICCIT). IEEE. 2019, pp. 1– 5.

[18] Xiaoqing Zhang Na Lin Lirong Feng. "Aircraft Detection in Remote Sensing Image based on Optimized Faster-RCNN". In: Remote Sensing Technology and Application 36.2, 275 (2021), p. 275. doi: 10.11873/j.issn.1004-0323.2021.2.0275. url: http://www.rsta.ac.cn/EN/abstract/article 3339.shtml.

[19] Berke Oral et al. "Information extraction from text intensive and visually rich banking documents". In: Information Processing & Management 57.6 (2020), p. 102361.

[20] Martha O Perez-Arriaga, Trilce Estrada, and Soraya Abad-Mota. "TAO: system for table detection and extraction from PDF documents". In: The Twenty-Ninth International Flairs Conference. 2016.

[21] AS Revathi and Nishi A Modi. "Comparative Analysis of Text Extraction from Color Images using Tesseract and OpenCV". In: 2021 8th International Conference on Computing for Sustainable Global Development (INDIACom). IEEE. 2021, pp. 931–936.

[22] Binwen W Shuaishuai L Yu Y. "An Automatic Crack Detection Method for Structure Test Based on Improved Mask RCNN". In: Journal of Vibration, Measurement Diagnosis. IEEE. 2021, pp. 487–494.

[23] Harneet Singh and Anmol Sachan. "A proposed approach for character recognition using document analysis with ocr". In: 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS). IEEE. 2018, pp. 190–195.

[24] Ahmad S Tarawneh et al. "Invoice classification using deep features and machine learning techniques". In: 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT). IEEE. 2019, pp. 855–859.

[25] Jiaqi Wang et al. "Region proposal by guided anchoring". In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019, pp. 2965– 2974.

[26] Yi Xiao et al. "Object detection based on faster R-CNN algorithm with skip pooling and fusion of contextual information". In: Sensors 20.19 (2020), p. 5490.

[27] Ting XU et al. "Intelligent Document Processing: Automate Business with Fluid Workflow". In: Konica Minolta technology report 18 (2021), pp. 89–94.

[28] Ziyi Yang and Ning Li. "Identification of Layout elements in Chinese academic papers based on Mask R-CNN". In: 2022 2nd International Conference on Consumer Electronics and Computer Engineering (ICCECE). IEEE. 2022, pp. 250–255.

[29] Yu Zhang et al. "Exploring weakly labeled images for video object segmentation with submodular proposal selection". In: IEEE Transactions on Image Processing 27.9 (2018), pp. 4245–4259.

[30] Zhiguo ZHOU et al. "Object Detection and Tracking of Unmanned Surface Vehicles Based on Spatial-temporal Information Fusion". In: 43.6 (2021), pp. 1698–1705.