ISSN No:-2456-2165

Leveraging Artificial Intelligence for Simplified Invoice Automation: Paddle OCR-based Text Extraction from Invoices

Jaya Krishna Manipatruni[1], R Gnana Sree[2], Ranjitha Padakanti[3], SreePriya Naroju[4], Bharani Kumar Depuru[5]
[1,2]Research Associate, [3]Mentor, [4]Team Leader, Research and Development, [5]Director
Innodatatics, Hyderabad, India

*Corresponding Author: Bharani Kumar Depuru
OCR ID: 0009-0003-4338-8914

Abstract:- In this study, we delve into the utilization of PaddleOCR, a readily available tool for optical character recognition (OCR), in extracting text from invoices. It is of utmost importance to accurately extract data from invoices, including information about vendors, invoice dates, item descriptions, quantities and prices to effectively manage finances. We achieved this by leveraging the powerful deep learning models and pre-trained weights provided by PaddleOCR to process invoice images and extract the necessary textual details.

Our investigation commences with a comprehensive analysis of the PaddleOCR framework, exploring its capabilities and potential for customization. We explore various techniques aimed at enhancing image quality and improving OCR accuracy. The PaddleOCR framework offers advanced functionalities such as text detection, recognition and layout analysis that we seamlessly incorporate into our workflow to accommodate diverse invoice layouts and formats.

To train our OCR model effectively, we curate a meticulously crafted dataset comprising real world invoice images with varying characteristics. With this dataset in hand, we fine tune the PaddleOCR model with a specific focus on enhancing its performance in extracting text from invoices.

Upon training the model successfully, we evaluate its performance using an independent test dataset while measuring key metrics like Character Error Rate (CER) and Word Error Rate (WER).

Our research strongly confirms the efficacy of the PaddleOCR powered system in precisely extracting text from invoices that have different layouts and formats. Additionally, we conduct a comparison between our methodology and other OCR techniques, emphasizing the benefits of PaddleOCR's advanced deep learning framework.

Furthermore, we seamlessly integrate the invoice text extraction pipeline into a comprehensive automated system for invoice processing. This integrated system streamlines the extraction, parsing, and organization of invoice data, leading to more efficient financial workflows. We also consider the potential applications of this technology, including invoice digitization, data analytics, and process automation, all of which contribute to significant improvements in operational efficiency and reduced manual labour in organizations.

In summary, this research demonstrates the successful use of PaddleOCR for text extraction from invoices. Our developed pipeline excels in accuracy and adaptability across various invoice layouts, paving the way for increased automation in financial management and document processing.

Keywords:- *PaddleOCR, Optical Character Recognition (OCR), Data Extraction, Data Learning Models, Text Detection, Character Error Rate (CER), Word Error Rate (WER), Invoice Digitization, Data Analytics.*

I. INTRODUCTION

The primary goal of this research study and the development of an automatic text extraction system is to improve the accuracy and efficiency of extracting information from store bills while also reducing the time required for this process. The approach for this research study is based on the Cross Industry Standard Process for Machine Learning with Quality Assurance, CRISP-ML(Q) methodology, which is conveniently accessible as an open-source framework on the 360DigiTMG website (ak.1) [Fig. 1].

ISSN No:-2456-2165

Fig. 1: CRISP-ML(Q) - approach for quality assurance for each of the six phases.
(Source: Mind Map - 360DigiTMG)

In the world of financial transactions, paper bills have long served as the traditional repository for vital information. These documents offer reliability but come with the time-consuming and labour-intensive task of manual data retrieval, especially when searching for specific transactions. In contrast, modern computing systems excel at such tasks, raising the question of how to automate the extraction and storage of information from paper bills.

In today's digital era, the need to bridge the gap between the convenience of digital storage and the persistence of paper-based transactions has grown. Fortunately, optical character recognition (OCR) technology has evolved to a point where it can reliably and accurately convert printed text from images into machine-readable text, faithfully preserving the original order and spacing of words [8].

Our solution embraces the transformative potential of OCR technology, specifically focusing on the use of PaddleOCR, to automate the extraction and storage of information from paper bills, with a particular emphasis on supermarket invoices. By doing so, we not only streamline the process but also eliminate the need for manual data entry. This approach significantly reduces the time required to retrieve essential information from bill images, enhancing efficiency and accuracy in managing financial records.

In this article, we explore the immense potential of OCR technology, especially when applied to the context of supermarket invoices. We delve into how this technology, with the use of PaddleOCR, can revolutionize the way we handle financial transactions, making them more efficient, accurate, and accessible in our increasingly digital world.

Optical Character Recognition (OCR) has emerged as a game-changing technology in the field of artificial intelligence. OCR empowers computers to understand and interpret text from images. While humans effortlessly read characters within images, OCR's significance lies in equipping computers with similar capabilities.

Within the scope of machine learning and data analysis, OCR plays a crucial role in extracting information from different sources, with a specific focus on documents in image form, including typed, handwritten, or printed content. The real challenge lies in creating intelligent models that can accurately recognize and understand text in these images, even when dealing with complex documents. In today's fast-paced business environment, automating tasks like data extraction from invoices and payments is essential for organizations of all sizes.

Invoices are important documents in business deals. They contain important information like details about the customer, the seller, what was bought, how much it cost, taxes, and when the payment is due. However, dealing with invoices manually, especially when there are a lot of them, takes a long time and can lead to mistakes. Invoices can come in different forms like paper, emails, PDFs, pictures, or electronic data, which makes it harder to get the information from them. Also, invoices can have different kinds of information, like tables, words, pictures, or logos.

ISSN No:-2456-2165

Python is a helpful tool for automatically getting information from invoices. This study talks about how Python can be really useful for getting both structured (like tables) and unstructured (like words and pictures) information from invoices, especially PDFs. It also mentions using machine learning to help with this.

Getting data from invoices can be tricky because there's a mix of structured and unstructured information. Structured data follows clear patterns, like tables, while unstructured data doesn't, like words and images. This is why automation is so important for this task.

Now we would discuss PaddleOCR and its applications. It's a powerful system for reading text from images and documents. PaddleOCR has different pre-made models that balance accuracy and speed. These models can find text, figure out its direction, and read it. PaddleOCR is based on a special kind of neural network called CRNN, which is good at understanding sequences of information.

II. METHODS AND TECHNIQUES

A. *OpenCV:*

OpenCV, short for Open-Source Computer Vision Library, is a powerful open-source software library and collection of programming functions designed primarily for computer vision and image processing tasks. It provides developers with a comprehensive toolbox to create applications that can interpret visual information from images and videos [3].

OpenCV was originally developed by Intel in 1999 and has since become one of the most widely used computer vision libraries globally. It supports various programming languages, including C++, Python, and Java, making it accessible to a broad range of developers [3].

The library offers a vast array of features, including image and video capture, image manipulation, object recognition, machine learning, and deep learning integration. OpenCV is used in a diverse range of applications, from robotics and autonomous vehicles to medical imaging, augmented reality, facial recognition, and more.

Its versatility and extensive documentation make OpenCV an invaluable resource for researchers, engineers, and developers seeking to harness the power of computer vision in their projects and applications.

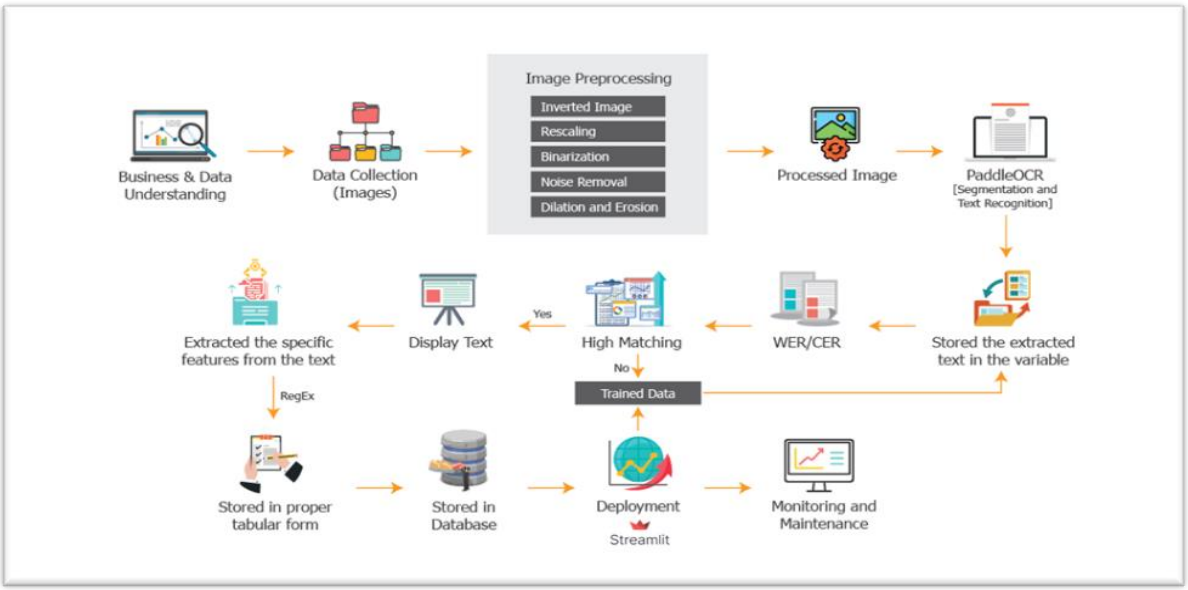
B. *PaddleOCR:*

PaddleOCR is an open-source optical character recognition (OCR) toolkit developed by PaddlePaddle, an AI framework by Baidu. PaddleOCR is designed to accurately and efficiently recognize text in images and scanned documents, making it a valuable tool for a wide range of applications, including document digitization, text extraction, and data analysis [4].

PaddleOCR leverages deep learning models and pre-trained weights to achieve high accuracy in text recognition tasks. It provides support for multiple languages and fonts, making it versatile for use in various global contexts. PaddleOCR's architecture includes text detection, text recognition, and layout analysis components, enabling it to handle diverse document layouts and formats.

One of the standout features of PaddleOCR is its adaptability and ease of customization. Users can fine-tune the OCR models to suit their specific requirements, making it suitable for both general and domain-specific applications.

With its robust capabilities and open-source nature, PaddleOCR has gained popularity in fields such as document management, data extraction, and automation. It offers a comprehensive and efficient solution for extracting valuable information from images and documents, contributing to improved efficiency and productivity in various industries.



ISSN No:-2456-2165

Fig. 2: Workflow of Methodology - Framework and training process of Text Extraction using PaddleOCR model.
(Source: - *Open-Source ML Workflow Tool- 360DigiTMG*)

C. Workflow of Methodology:

➢ *Step 1: Upload Invoices*
- Users can upload invoices in either image or PDF format (ak.2) [Fig.2]
- If a PDF is uploaded, it is converted into an image format for further processing.

➢ *Step 2: Basic Pre-processing*
- The uploaded image is subjected to basic pre-processing steps, including:

✓ Image resizing: Ensuring uniform dimensions for consistent OCR results.
✓ Binarization: Converting the image to binary format to enhance text clarity.
✓ Normalization: Adjusting image properties for uniformity.
✓ Noise Removal: Eliminating unwanted artifacts and enhancing text quality.

➢ *Step 3: Paddle OCR Engine*
- The pre-processed image is passed through the Paddle OCR engine.
- Paddle OCR performs text detection, recognition, and layout analysis.
- The result is a comprehensive extraction of text from the invoice.

➢ *Step 4: Information Extraction*
- Using regular expressions and text analysis techniques, the system identifies and extracts specific information [Fig.2], including:

✓ Invoice Number
✓ Bill Date
✓ Product Name
✓ Unit Price
✓ Quantity
✓ Total Amount

➢ *Step 5: Storage in Database*
- The extracted information is stored in a structured tabular form within a database.
- This database serves as a centralized repository for all invoice data.
- Storing the information in a database allows for easy retrieval, search, and analysis of invoice details.

Overall, this workflow enables the seamless extraction and structured storage of critical information from uploaded invoices, ensuring accuracy and efficiency in financial data management. The database serves as a valuable resource for managing and analysing invoice data efficiently.

D. Evaluation Metrics

➢ *Accuracy*

Accuracy is a critical metric for evaluating the performance of the PaddleOCR in our project. It quantifies the effectiveness of the OCR engine in accurately extracting text from invoices. The accuracy is determined by considering the number of correctly extracted words and characters by PaddleOCR in relation to the total number of words and characters present in the original invoice.

In our context, higher accuracy signifies the efficiency of our pre-processing techniques and PaddleOCR in extracting data accurately from invoices. An accuracy of x% indicates that out of 100 characters present in the original invoice, PaddleOCR successfully extracted x characters. Consequently, the remaining (100 - x) % represents any errors or discrepancies introduced by PaddleOCR, which may include misclassified characters or the inability to extract certain characters accurately.

In essence, accuracy serves as a crucial measure of the OCR system's performance and its ability to reliably extract textual information from invoices, a fundamental aspect of our project's success.

Word Error Rate (WER) and Character Error Rate (CER) are commonly used metrics to evaluate the accuracy of optical character recognition (OCR) systems like PaddleOCR. They measure the dissimilarity between the recognized text and the ground truth (i.e., the actual text in the document). Lower error rates indicate higher accuracy [7].

E. Character Error Rate (CER):
- CER measures the accuracy at the character level. It calculates the percentage of incorrectly recognized characters in the recognized text compared to the ground truth.

Formula: CER=S+D+INCER=NS+D+I

Where:
- SS represents the number of substitutions (incorrectly recognized characters).
- DD represents the number of deletions (characters present in the ground truth but not recognized).
- II represents the number of insertions (characters recognized but not present in the ground truth).
- NN is the total number of characters in the ground truth.
- Lower CER values indicate higher accuracy, with a CER of 0% indicating a perfect match between recognized and ground truth text.

F. Word Error Rate (WER):
- WER measures accuracy at the word level. It calculates the percentage of incorrectly recognized words in the recognized text compared to the ground truth.

Formula: WER=S+D+INWER=NS+D+I

Where:
- SS represents the number of substitutions (incorrectly recognized words).

- DD represents the number of deletions (words present in the ground truth but not recognized).
- II represents the number of insertions (words recognized but not present in the ground truth).
- NN is the total number of words in the ground truth.

Similar to CER, lower WER values indicate higher accuracy, with a WER of 0% indicating perfect word-level recognition.

Both CER and WER provide a quantitative assessment of the OCR system's performance, allowing you to gauge its accuracy by comparing recognized text with the actual content of the document.

Table 1: Quantitative assessment of the OCR system's performance

	WER	CER
invoice	2.0992%	2.0055%

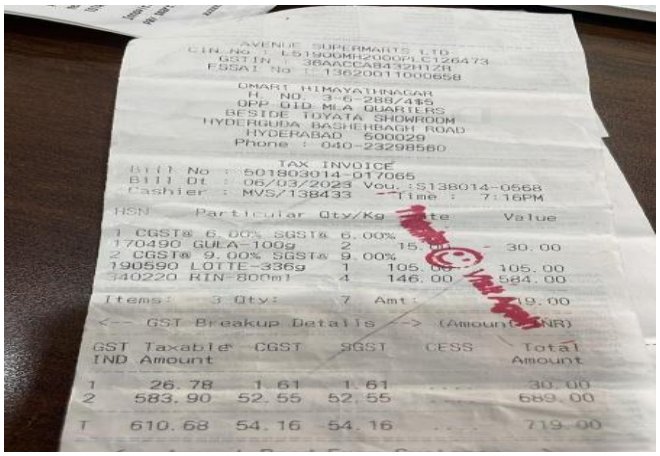


Fig. 3: Invoice from supermarket - Store bill containing items bought by customer

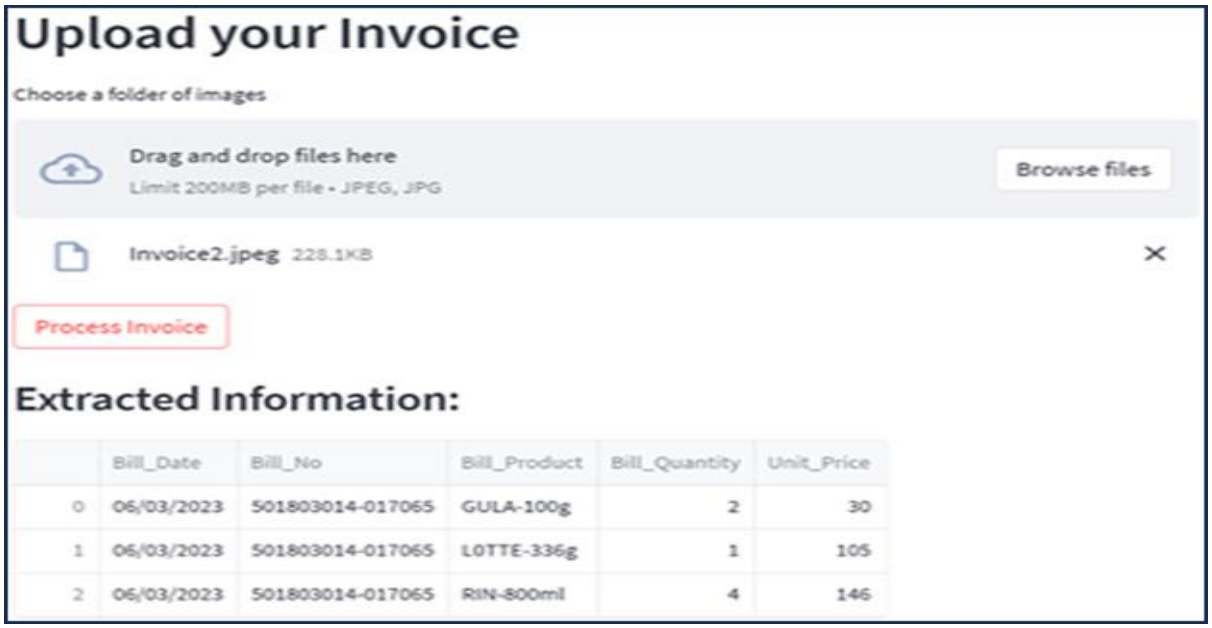


Fig. 4: Deployment Output - Extracted information from invoice

III. RESULT AND CONCLUSION

The application of advanced technologies like Optical Character Recognition (OCR) in our project has demonstrated the remarkable potential to accurately extract information from bills and invoices. This innovation proves invaluable in efficiently managing and tracking financial transactions, especially those involving cash payments. It simplifies the storage and accessibility of financial records, making them easily searchable and reducing the tediousness of manual record-keeping.

With the successful implementation of this system, its practicality extends to a wide range of scenarios, including retail establishments and households. It addresses the common challenge of tracking and organizing bills, offering an additional feature of budget tracking, which is a vital component in financial management.

As we look ahead, the proposed system has the potential for further enhancements, such as the verification of financial documents. This advanced functionality could involve the use of complex regular expressions to process the extracted text, particularly numeric data, and identify relationships between various financial elements. Additionally, the system can be extended to detect and accurately store tables, a common feature in financial documents.

Recognizing the challenges posed by older bills with fading or smudged text, future developments could explore innovative solutions such as Augmented Reality (AR) and Virtual Reality (VR) for restoring partial or degraded text, ensuring the preservation of critical information.

Moreover, the conversion of text images into machine-readable text opens doors to accessibility solutions, benefiting visually impaired individuals. The technology can facilitate text-to-speech conversion, making it simple for those with visual impairments to access and comprehend textual content, extending its potential even to the recitation of entire books.

In conclusion, our project showcases the transformative power of OCR technology, particularly with PaddleOCR, in streamlining financial data management and document processing. It offers a robust, accurate, and adaptable solution for text extraction from invoices, contributing to enhanced automation, efficiency, and accessibility in financial and document-related domains.

DECLARATIONS

➢ *Acknowledgments:*

We acknowledge that with the consent from 360DigiTMG, we have used the CRISP-ML(Q) methodology (ak.1) and the ML Workflow which are available as open-source in the official website of 360DigiTMG (ak.2).

➢ *Funding and Financial Declarations:*

- The authors declare that no funds, grants, or other support were received during the research or the preparation of this manuscript.
- The authors declare that they have no relevant financial or non-financial interests to disclose.

➢ *Data Availability Statement:*

The datasets used, generated and/or analysed during this study are not publicly available due to internal Data Privacy Policy but are available from the corresponding author on reasonable request.

COMPLIANCE WITH ETHICAL STANDARDS

- **Disclosure of potential conflicts of interest:** The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.
- **Research involving Human Participants and/or Animals:** It is declared by all the authors that there was no involvement of any human and/or animal trial or test in this research.
- **Informed consent:** As there were no human subject involved in this research hence a informed consent is not applicable to the best of the authors' understanding.
- **Conflict of Interest Statement:** The authors declare that there are no conflicts of interest that could influence the results or interpretation of this manuscript. This research was conducted in an impartial and unbiased manner, and there are no financial, personal, or professional relationships that might be perceived as having influenced the content or conclusions presented in this work.

REFERENCES

[1.] M. A. Ozdil and F. T. Y. Vural, "Optical character recognition without segmentation," Proceedings of the Fourth International Conference on Document Analysis and Recognition, Ulm, Germany, 1997, pp. 483-486 vol.2, doi: 10.1109/ICDAR.1997.620545.

[2.] Weiliang Liu, Xueguang Yuan, Yangan Zhang, Mengya Liu, Zhenyu Xiao, Jianlan Wu, "An End to End Method for Taxi Receipt Automatic Recognition Based on Neural Network", Automation Control Conference (ITNEC) 2020 IEEE 4th Information Technology Networking Electronic and, vol. 1, pp. 314-318, 2020

[3.] Vedant Kumar, Pratyush Kaware, Pradhuman Singh, Dr. Reena Sonkusare Electronics & Telecommunications Department Sardar Patel Institute of Technology Mumbai, India" Extraction of information from bill receipts using optical character recognition", https://ieeexplore.ieee.org/document/9215246/authors #authors

[4.] Sanyam, "Optical Character Recognition using PaddleOCR" 'https://learnopencv.com/optical-character-recognition-using-paddleocr/

[5.] Manoharan, S. (2019)," Smart Image Processing Algorithm For Text Recognition, Information Extraction And Vocalization For The Visually Challenged",Journal of Innovative Image Processing (JIIP), 1(01), 31 38.

[6.] Shakya, Subarna. "Virtual Restoration Of Damaged Archeological Artifacts Obtained From Expeditions Using 3D Visualization." Journal of Innovative Image Processing (JIIP) 1, no. 02 (2019): 102-110.

[7.] Kenneth Leung, "Evaluate OCR Output Quality with Character Error Rate (CER) and Word Error Rate (WER), Jun 24, 2021, https://towardsdatascience.com/evaluating-ocr-output-quality-with-character-error-rate-cer-and-word-error-rate-wer-853175297510 .

[8.] J. Liang, D. Doermann, and H. Li, "Camera-based analysis of text and documents: a survey," International Journal of Document Analysis and Recognition (IJDAR), vol. 7, pp. 84–104, 2005.