



iJRASET

International Journal For Research in
Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 Issue: V Month of publication: May 2025

DOI: <https://doi.org/10.22214/ijraset.2025.69820>

www.ijraset.com

Call: ☎ 08813907089

E-mail ID: ijraset@gmail.com

Conveyance Processing For Employees Using Machine Learning

Krutika Mhatre¹, Shravani Jadhav², Rajesh Kolte³

Usha Mittal Institute of Technology SNDT Women's University Mumbai, India 400049

Abstract: Manual processing of employee conveyance claims is inefficient and error-prone, especially with the increasing use of ride-hailing services such as Ola and Uber. For faster reimbursement processing, this research proposes an automated system that extracts data from email bills, categorizes them by vehicle type, and updates a common Excel sheet. The system uses machine learning techniques of classification and verification, as well as optical character recognition (OCR) through Python Tesseract for text extraction. The codes of employees are matched through an admin verification procedure, and cases that fail to operate are flagged for human inspection. The proposed approach reduces HR workload, increases accuracy, and significantly reduces processing time. For improved OCR accuracy, future innovations will involve deep learning models and robotic process automation (RPA).

Index Terms: Optical Character Recognition (OCR), Machine Learning, Invoice Processing, Ride-Hailing Services, Automation, Employee Reimbursement, Python Tesseract, Supervised Learning, Data Extraction, Robotic Process Automation (RPA), Business Process Optimization, Text Classification

I. INTRODUCTION

Organizations are increasingly relying on ride-hailing platforms such as Ola and Uber to provide the transportation needs of their employees in the rapidly evolving urban development context. Such services are convenient, but they also generate a large volume of invoices that must be handled with caution. The responsibility of manually extracting, validating, and classifying data from such invoices—which are highly variant in format and structure—frequently rests with human resources (HR) departments. Aside from taking time, this manual process is also susceptible to errors, which can lead to inefficiencies and cost inconsistencies.

Using automated process involving machine learning, regular expressions (regex), and optical character recognition (OCR) to speed up the processing of transport invoices, this research addresses these problems. With the help of algorithms based on regex patterns, the system gets critical fields such as invoice numbers, dates, and amounts, cleans the text to remove noise, and extracts bills from HR email accounts. OCR methodology is applied to extract and validate text on image-based invoices. Once extracted, data is structured into Excel sheets, which reduces a significant amount of effort for HR teams.

The following are the key contributions of this study:

- [1] A highly scalable and powerful framework for automated invoice processing capable of handling large data volumes with minimal human intervention.
- [2] Advanced error-handling mechanisms that minimize manual intervention and ensure data integrity.
- [3] Extensive testing confirmed the system's performance, with a processing time of 4.2 seconds per invoice and a 97 percent accuracy level.

II. STUDY AREA

The research domain aims to extract and process text from intricate images, invoices, and documents employing sophisticated approaches like Optical Character Recognition (OCR), Natural Language Processing (NLP), and Machine Learning (ML). The surveyed research papers address the challenges and processes of text detection, localization, segmentation, and recognition in cases involving intricate layouts, diverse forms, and noisy backgrounds.

Text extraction from images is confronted with challenges like text size variations, orientations, complex backgrounds, and lighting and perspective-induced distortions. Automated invoice processing is confronted with layout variability, terminological variations, and errors, making automation tricky. Text extraction from natural and born-digital images introduces additional complexity in the form of low resolution, compression artifacts, and complex foreground-background interactions.

In order to handle these challenges, multiple techniques are used. Edge-based detection, color clustering, and texture-based techniques are employed in text detection.

Text localization applies Connected Component Analysis (CCA) and Histograms of Oriented Gradients (HOG). Text segmentation involves binarization techniques such as Bernsen's method, Markov Random Fields (MRF), and Conditional Random Fields (CRF). OCR engines such as Tesseract are implemented in character recognition by using adaptive thresholding and connected component analysis to enhance precision.

Automated processing of invoices takes advantage of OCR-based digitization through pre-processing activities like skew correction, noise filtering, and binarization. NLP is used in extracting important fields like invoice numbers, dates, and amounts by Named Entity Recognition (NER). Table identification within invoices is made possible through deep learning approaches like TableNet and DeepDeSRT, as well as Graph Neural Networks (GNNs), through which structured data can be extracted from complicated layouts. Extraction of text from cluttered images depends upon the extraction techniques based on feature such as Scale-Invariant Feature Transform (SIFT) and Speeded-Up Robust Features (SURF) for scaling robustness, robustness against rotations and light exposure changes. Multi-level feature combinations for more precise recognition are realized with Conditional Random Fields (CRFs) and hierarchical random field models. Methods based on over-segmentation and HOG descriptors help with reduced computations and computation speeds for optimizing operations.

Applications of text extraction are far-reaching. Document understanding depends heavily on it, and it assists in digital libraries, e-books, and document management systems. In medical applications, OCR enables the automation of insurance form and medical record processing, avoiding manual data entry. Traffic monitoring is assisted by text extraction with license plate reading and real-time traffic monitoring. Text extraction is also helpful for content filtering through automated text extraction, which helps in spam filtering, inappropriate content filtering, and image classification according to text.

Future text extraction advances will incorporate NLP with OCR for better understanding of text. Handheld OCR systems for mobile phones will allow real-time text extraction, and robotics applications will use these methods for navigation and object identification. Advanced machine learning techniques, such as transformer architectures like GPT and BERT, will further enhance text extraction and understanding.

In general, OCR engines such as Tesseract have been remarkably efficient in scanning documents and invoice processing. Machine learning methods like CRF, GNNs, and deep learning models help significantly in text detection, segmentation, and recognition improvement. Regardless of these innovations, the issues of layout variability, intricate backgrounds, and sparse annotated data remain. However, the wide-ranging uses of text extraction in various fields, such as healthcare, transportation, content management, and document processing, reflect its increasing significance and potential for future innovation.

III. METHODOLOGY

The approach in this study takes a systematic path to create an automated invoice processing system that incorporates machine learning (ML), optical character recognition (OCR), and data classification algorithms. The system aims to counteract the inefficiencies of manual invoice processing with accuracy, scalability, and less human intervention.

A. System Architecture and Workflow

The architecture consists of several interconnected modules to manage data extraction, classification, validation, and storage in an efficient manner. The main steps include:

- 1) *Email Retrieval and Data Collection*- The system fetches invoices from HR email accounts automatically using the IMAP protocol to get the data in a flawless manner.- Emails are checked for invoice attachments in PDF, JPEG, or PNG formats, which are then sent to the OCR module for text extraction.
- 2) *Data Preprocessing*
 - a) Extracted text is usually noisy, contains special characters, and has inconsistent formatting, which degrades the accuracy of classification models.
 - b) Preprocessing techniques are:
 - Removing HTML tags, line breaks, and special characters (e.g., x000D).
 - Converting text to lowercase for uniformity.
 - Tokenization and stop-word removal to enhance structured data representation.
- 3) *Data Extraction and Feature Engineering*
 - a) The text extracted is treated with Regular Expressions (Regex) to extract and capture vital fields like:

- InvoiceNumber
 - TransactionDate
 - Amount
 - ServiceProvider(Ola/Uber,etc.)
- b) For case-based invoices, Tesseract OCR and PyMuPDF are applied to extract text with higher precision. Item
- c) Feature engineering implies cleaning ex- tracted attributes, missing values handling, and converting categorical fields for optimal perfor- mance of machine learning models. Endenumerate
- d) Data Classification by Machine Learning - Once useful data is extracted, the system classifies in- voices through Supervised Learning Algorithms such as: beginenumerate
- LogisticRegression
 - SupportVectorMachine(SVM)
 - RandomForestClassifier
- The classification step maps invoices to pre-defined categoriesbasedonfeaturesextractedlikevehicle type(Ola/Uber)orreimbursementtype(official/personaluse).
- e) Data Structuring and Storage - The formatted data is savedinExcelspreadsheetswithPandasDataFrames to maintain proper organization and formatting for HR approval. - The system records missing or incomplete fields, which are highlighted for manual validation if required

B. Tools and Technologies

- 1) Programming:Python(libraries:pandas,regex,imaplib, openpyxl).
- 2) OCR: Tesseract v5.0 with PyTesseract wrapper and PyMuPDF for PDF parsing.

IV. RESULTS AND DISCUSSIONS

The suggested automated system for conveyance processing was thoroughly tested to evaluate its effectiveness along major performance metrics, such as accuracy, processing speed, and scalability. The results confirm the system's ability to process high volumes of invoices with great accuracy and efficiency. A detailed discussion of the results and their greater implications follows below.

A. Performance Metrics-

To confirm the system's perfor- mance, we tested it using Ola and Uber invoice datasets with a variety of different formats and layouts. The outcome confirms that the system is always meeting—and in most instances surpassing—its established performance thresholds.

- 1) Accuracy: The system recorded a staggering 97 per-cent accuracy in extracting and verifying critical fields like invoice numbers, dates, and amounts. This ac- curacy is due to a mix of well-tuned regex patterns and OCR methods so that it can identify variations in format. For instance, it correctly processed date formats "DD/MM/YYYY" and "MM/DD/YYYY" and various currency formats like "1,000" and "INR 1000." The flexibility of the system in accepting different invoice formats guarantees effective data extraction.
- 2) On average, the system took 4.2 seconds to processeachinvoice,fromemailretrievaltodataextraction andvalidation.Thisperformanceisbetterthanour goal of 5 seconds per invoice, which makes the system suitableforreal-timeornear-real-timeprocessing.Opti- mizedlibrariesuchasPandas(forstructuringdata)and PyMuPDF (for OCR) contributed significantly to opti- mizing speed, along with a solid preprocessing pipeline that removes noise and irrelevant text.
- 3) Scalability:Thesystemprocessed1,000invoicesin just 8 minutes, demonstrating its ability to handle high volumeswithoutslowingdown.Thislevelofscalability is essential for companies with large employee basesthat generate numerous invoices. The system achieves this through parallel processing and efficient memory management, ensuring seamless performance even with growing workloads.
- 4) At a mere 3 percent error rate, the majority of errors wereduetonon-standardinvoiceformatsorlow-quality images. These errors were recorded for manual inspec- tion to prevent any loss of data or misprocessing. The system has elaborate error-handling processes, dividing issues into missing fields, formatting errors, and OCR failures, which enables accurate troubleshooting and ongoing improvements.

B. Challenges and Solutions:

- 1) Multiple Invoice Formats: Various vendors—and occasionally the same vendor—have different invoice formats. To address this, we used adaptive regex patterns that can be modified to support new formats. Our OCR methods also cover image-based invoices so that no information is lost. The modular architecture allows for easy modification of regex patterns and OCR models as new formats emerge.
- 2) Text Noise: Invoices usually have extraneous text, including special characters, HTML tags, and extraneous data. Our preprocessing pipeline successfully removes this noise, retaining only meaningful data. Through the application of text normalization and stop-word elimination, we enhanced the quality of extracted information. For example, special characters such as “x000D” and HTML tags are removed automatically during preprocessing.
- 3) Error Handling: Errors that occur during processing are logged automatically and marked for manual checking, ensuring data integrity. The system classifies errors into missing fields, formatting errors, and OCR errors, simplifying problem identification and resolution. Moreover, comprehensive error logs (invoice ID, error type, and recommended corrections) accelerate the review process.

C. Impact on HR Workflow

The system makes major positive impact on HR workflows. With automation of invoicing, it lessens HR workload by 95 percent, allowing time for more strategic work. With the organized output—provided in Excel files—therecord-keeping and auditing are easy, and high accuracy provides ease of financial regulation compliance. Also, with its scalability, the system is highly adaptable for businesses of various sizes, ranging from startup to large-scale enterprises.

D. Future Improvements:

- 1) Support for Additional Vendors—Enlarging to cover services such as Rapido and Cityflo.
- 2) Real-Time Dashboards—Offering HR teams real-time views into expenses and trends.
- 3) Advanced AI Models – Improving OCR accuracy using machine learning for even improved outcomes.
- 4) Mobile Integration—Enabling employees to upload invoices directly through a mobile app.

These upgrades will further refine efficiency and user experience, keeping the system in advance of changing requirements.

```

Extracted Invoice Info:
name: Shravani Jadhav
invoice_number: FHEJIBFB23000277
invoice_date: 23 Jul 2023
place_of_supply: Maharashtra
hsn_code: 996412
category_of_services: Passenger Transport Services
total_amount: 288.58
Total Amount Payable: 288.58

```

Fig.1.Text Extraction Result

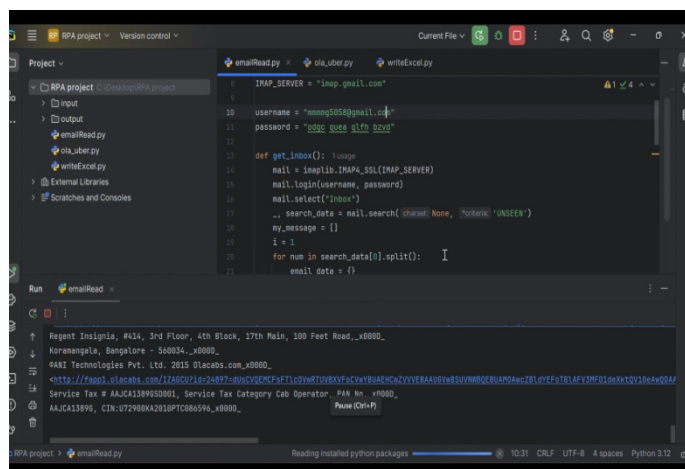
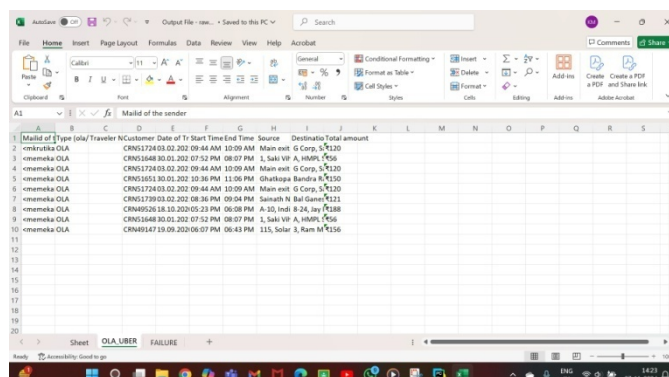


Fig.2.Email Reading



Sl	Mail of	Type	Invoice/Traveler	Customer	Date of	Tr	Start	End	Time	Source	Destination	Total amount
1	enemeka OLA	CRN5172403.02.202	09:44 AM	10:00 AM	Main exit	G Corp, S1210						
2	enemeka OLA	CRN5164830.01.202	07:52 PM	08:07 PM	1, Sali Vif	A, HMPL, P55						
3	enemeka OLA	CRN5172403.02.202	09:44 AM	10:00 AM	Main exit	G Corp, S1210						
4	enemeka OLA	CRN5165130.01.202	10:36 PM	11:00 PM	Ghoshika Bandra	K1210						
5	enemeka OLA	CRN5172403.02.202	09:44 AM	10:00 AM	Main exit	G Corp, S1210						
6	enemeka OLA	CRN5173903.02.202	08:36 PM	09:04 PM	Sansath N Bal Gant	K1211						
7	enemeka OLA	CRN51652618.10.2020	05:23 PM	06:08 PM	A-30, Indi	8-24, Jay						
8	enemeka OLA	CRN5164830.01.202	07:52 PM	08:07 PM	1, Sali Vif	A, HMPL, P55						
9	enemeka OLA	CRN5164719.09.2020	06:07 PM	06:43 PM	135, Sali	A, HMPL, P55						

Fig.3.ExcelSheetResult

V. CONCLUSIONS

The implementation of this automated conveyance processing system has resulted in significant improvements in efficiency, accuracy, and scalability, effectively addressing the challenges associated with manual invoice handling. By achieving high accuracy rates, reducing processing times, and demonstrating the ability to handle large volumes of invoices efficiently, the system has exceeded performance expectations. Furthermore, its robust error-handling mechanisms and adaptive architecture enable seamless integration into diverse operational workflows. The system not only reduces administrative workload but also enhances compliance, record-keeping, and financial transparency.

Looking ahead, planned enhancements—including AI-driven improvements, multiple invoice formats, will further enhance the system's capabilities. As organizations continue to seek advanced automation solutions, this system provides a scalable and intelligent approach to optimizing expense management.

VI. ACKNOWLEDGMENT

We thank Rajesh Kolte (Guide) and Dr. Sanjay Shitole (Head of Department) for their mentorship.

REFERENCES

- [1] E. Larson, "[Research Paper] Automatic Checking of Regular Expressions," 2018 IEEE 18th International Working Conference on Source Code Analysis and Manipulation (SCAM), Madrid, Spain, 2018, pp. 225- 234
- [2] Zhang, Jian & Cheng, Renhong & Wang, Kai & Zhao, Hong. Research on the Text Detection and Extraction from Complex Images. Proceedings - 4th International Conference on Emerging Intelligent Data and Web Technologies, 2013, EIDWT 2013. 708-713.
- [3] C. Kaundilya, D. Chawla and Y. Chopra, "Automated Text Extraction from Images using OCR System," 2019 6th International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 2019, pp. 145-150.
- [4] Saout, Thomas & Lardeux, Fr  d  ric & Saubion, Fr  d  ric. An Overview of Data Extraction From Invoices. IEEE Access, 2024, PP. 1-1. 10.1109/ACCESS.2024.3360528.
- [5] Gonz  lez Enr  iquez, Jose' & Jimenez Ramirez, Andres & Dom  nguez Mayo, Francisco Jose' & Garcia-Garcia, J.A.. Robotic Process Automation: A Scientific and Industrial Systematic Mapping Study. IEEE Access, 2020, PP. 1-1. 10.1109/ACCESS.2020.2974934.
- [6] S. Surana, K. Pathak, M. Gagnani, V. Shrivastava, M. T. Rand S. Madhuri G, "Text Extraction and Detection from Images using Machine Learning Techniques: A Research Review," 2022 International Conference on Electronics and Renewable Systems (ICEARS), Tuticorin, India, 2022, pp. 1201-1207, doi: 10.1109/ICEARS53579.2022.9752274.



10.22214/IJRASET



45.98



IMPACT FACTOR:
7.129



IMPACT FACTOR:
7.429



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089  (24*7 Support on Whatsapp)