INVOICE DATA EXTRACTION USING OCR TECHNIQUE

Dr. V. Krishna Vijaya[1], Associate Professor, Department of IT,
KKR & KSR Institute of Technology and Sciences, Vinjanamapdu, Guntur Dt., Andhra Pradesh.

Darsi Geethika[2], Maddu Lahari[3], Bheemana Pujitha[4], Gorantla Thribhuvaneswari[52,3,4,5] UG Students, Department of IT,
KKR & KSR Institute of Technology and Sciences, Vinjanampadu, Guntur Dt., Andhra Pradesh.

[1] vkvijaya13@gmail.com, [2] dgeethika27@gmail.com, [3] madduakanksha@gmail.com,
[4] bheemanapujitha@gmail.com, [5] thribhuvaneswarigorantla@gmail.com

Abstract

Traditional invoice processing involves manual entry of data, leading to human errors, delays,and increased operational costs. The lack of automation results in inefficiencies, hindering organizations from promptly accessing critical financial information. This research addresses the pressing need for a reliable OCR-based solution to automate invoice data extraction, ultimately improving accuracy, reducing processing time, and enhancing overall business productivity. The project aims to automate invoice data extraction through Optical Character Recognition (OCR) techniques. Leveraging advanced image processing and machine learning, the system will analyze scanned or photographed invoices, extracting relevant information such as vendor details, itemized costs, and dates.This automation streamlines manual data entry processes, enhancing accuracy and efficiency in managing financial records. OCR invoicing is the process of training a template-based OCR model for specific invoice layouts, setting up input paths for these invoices, extracting data, and integrating the extracted data with a structured database.

Keywords: Invoice, OCR, YOLO algorithm, Data Extraction, Image Processing, Database Integration.

1. Introduction

The invention of invoice data extraction using OCR (Optical Character Recognition) technique is rooted in the fields of computer vision, and document processing. This innovation aims to enhance efficiency in accounting and data management by accurately extracting data like vendor information, dates, amounts, and item details from scanned or digital invoices. The background involves the need for efficient and accurate digitization of invoice data from physical documents to streamline accounting processes, reduce manual effort, and minimize errors. OCR technology enables machines to "read" text from images or scanned documents, making it an ideal solution for extracting data from invoices. This innovation

streamlines the tedious process of manual data entry, improving efficiency and accuracy in invoice processing for businesses. It simplifies tasks for businesses by reducing manual data entry, minimizing errors, and speeding up invoice processing. This innovation revolutionizes invoice management by automating data extraction, improving accuracy, and saving time for businesses across various industries.

Invoice data extraction using OCR (Optical Character Recognition) technology is a game-changer for businesses. It means that instead of people having to type in all the details from invoices, a computer can read the information automatically from scanned or photographed invoices.

This saves a ton of time and reduces mistakes. Now, businesses can quickly get important details like invoice numbers, dates, and what was bought without having to do it all manually. This not only makes things faster and more accurate but also lets companies focus on more important tasks, boosting productivity.

The remainder of the paper will be presented below.The section 2 Literature review.Section 3 contains the proposed methodology. Section 4 contains the Implementation part. Section 5 contains the conclusion. Section 6 contains limitations. Section 7 the future work is proposed and discussions are followed by the section 8 acknowledgements.

2. Literature Review

Automated document processing, especially for invoices, receipts, and other business documents, is rapidly advancing [1,3,11, 12, 13, 19]. Two main approaches are driving this progress: machine learning and deep convolutional neural networks (CNNs).

Machine learning techniques are used to classify scanned documents and extract information [1, 3, 11, 12, 13, 19]. For example, some researchers achieved an accuracy rate of 85% for known invoice structures using a case-based reasoning method [11].

Deep convolutional neural networks (CNNs) offer another promising approach for document processing [5, 18]. These networks excel at document classification and retrieval tasks [5, 18].

Researchers are also exploring transfer learning, where pre-trained models can improve information extraction [6, 14]. This approach involves leveraging existing knowledge from one task to enhance performance on a new task.

The development and evaluation of these methods depend on well-defined datasets [2, 4, 9]. Some papers highlight specific datasets used to train and evaluate information extraction systems [2, 4, 9]. These datasets include FUNSD for understanding noisy scanned forms [2], Kleister for extracting information from

complex documents [4], and a dataset for scanned receipt processing [9].

3. Proposed Methodology

The proposed method for our Invoice data Extraction using OCR consists of several key components:

1.Pre-processing :

Image Ingestion:

Here the users can enter the invoices into the system. This can be done by either scanning physical copies or uploading digital files.

Fig 1: Invoice image that will be uploadedFormat Standardization:

We will convert the image into a consistent format, such as grayscale, to improve the accuracy of Optical Character Recognition (OCR).

Noise Reduction:

In this we applied techniques like filtering to remove background noise or smudgesfrom the image.

Fig 2: Invoice image with noise

2.Optical Character Recognition (OCR):

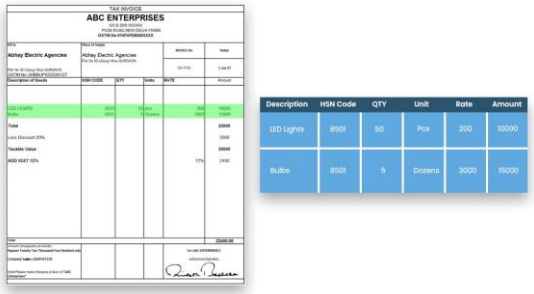


Fig 3: Scanning the text data using OCR

We used the Optical Character Recognition (OCR) engine to extract the text from the invoice image and convert it into a format that computers can understand.

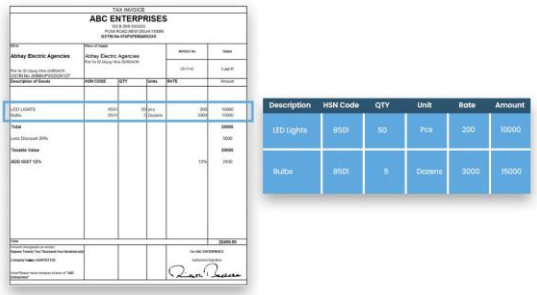


Fig 4: Recognition of the text

Text Segmentation:

Then the extracted text is segmented into meaningful sections, such as lines, paragraphs, or tables.

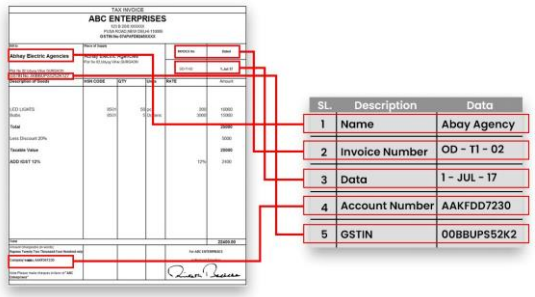


Fig 5: Recognition text is segmented into sections.

Model Architecture:

We used machine learning algorithms to classify the extracted text data.

Here's a high-level model architecture:

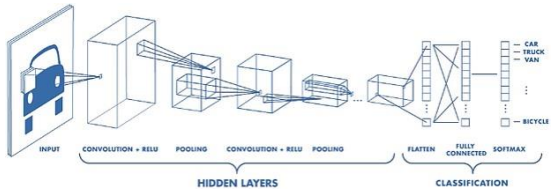


Fig 6: Model Architecture

Input Layer:

In this the system receives the pre-processed text data.

Hidden Layers:

Then the system analyzes the pre-processed text data to find patterns and how the information connects.Here we use techniques like convolutional neural networks (CNNs) to analyse.

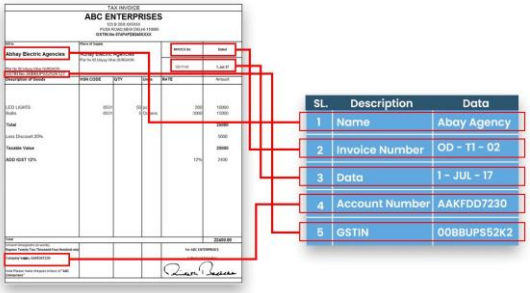
YOLO Model Output Layer:

We classify the extracted text into specific fields like vendor name, invoice number, date, quantity, description, price.

Post-processing and Validation:

Then the Extracted data undergoes a final check for accuracy and completeness.we used techniques like fuzzy matching to identify and rectify the potential errors.

3. Output:



Once everything is verified, then we export the validated data into a structured format like CSV or database, for further processing to the user.

Fig 7: Invoice image data is stored in the form of table

4. Implementation

The implementation of the our project involves several components, including a Streamlit used for building a user-friendly web interface where users can upload images, an OCR algorithms used for extracting the text, and a You Only Look Once (YOLO) model for Invoice data extraction using the RoboFlow dataset.Below is an overview of the implementation steps:

1. Environment Setup:

We installed necessary libraries like Ultralytics (depending on chosen YOLO version), OpenCV for image processing, EasyOCR, pandas, Streamlit and bcrypt.

We downloaded a pre-trained YOLO modellike YOLOv8.

2. Data Preparation:

Obtain a Invoice image dataset with annotations for different image types like class labels as (billing_addres, invoice_date,invoice_number,products ,shipping_address,subtotal,tax,tax_pe rcentage,total,& etc).

We implemented functions to read images and corresponding annotations.Apply data augmentation techniques (flipping rotation, cropping, noise reduction) to increase dataset size and improve modelgeneralization.

Dataset:

We generated training datasets based on Roboflow dataset.Visited the Roboflow website and explored their datasets. Choosed a dataset that aligns with our project or task. We selected a dataset containing images labeled with different class labels bounding boxes for various objects and downloaded the dataset in a format suitable for our framework, such as the YOLO v8 version, PyTorch, or others.

Training Details:

In our training process, we opted for a total of 150 epochs. This choice was made after considering the complexity of the task, the size of the dataset, and computational resources. While training, the model learns from the dataset multiple times, allowing it to capture patterns and present them in the data.

We set the learning rate to 0.89 for our training process. This choice was based on initial experimentation and hyperparameter tuning. A higher learning rate can lead to faster convergence, but it must be carefully chosen to prevent the model from overshooting optimal parameters. Throughout training, the learning rate remains constant to ensure stable optimization.

For our object detection task, we employed a combination of loss functionsto effectively train the model.

Bounding Box Regression Loss: This loss measures between predicted bounding box coordinates and the ground coordinates. Minimizing this loss encourages the model to accurately predict the locations of objects within the image.

Classification Loss:This loss encourages the model to correctly classify objects present in the image.

3. YOLO Model Customization:

If using a pre-trained model, modify the final layers to predict the desired number

of class labels in the image (e.g:billing_addres,invoice_date,invoice_nu
mber,products,shipping_address,subtotal, tax,tax_percentage,total,& etc).

Define the loss function by combining localization loss(e.g,IoU loss) and classification loss(e.g, cross entropy) suitable for bounding box prediction and class probabilities.

4. Design and User Interface(UI):

Design a form with fields for username/email and password .Include options like'Remember Me' (optional).Implement a "Login" button to submit the login credentials.

5. Uploading Image:

We used a python code to upload the image of invoice images and click the submit option. After,the detection and classification of the class labels and return the labels are stored in the database in the form of ".csv".

6. Deployment:

Depending on the application, the trained models can be deployed for real-time invoice data extraction on the new images.This might involve saving the model in a lightweight format and integrating it with a user interface for image processing and result visualization with high accuracy and speed

Results

The results of the project will be as follows:

Fig.1: upload page

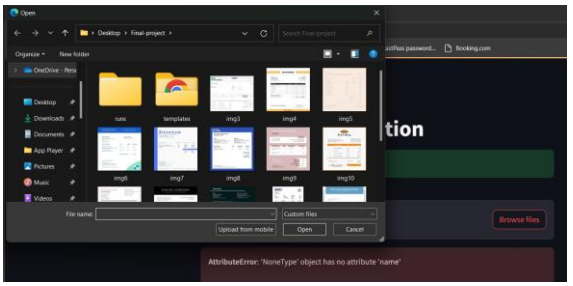


Fig.2: choosing the image

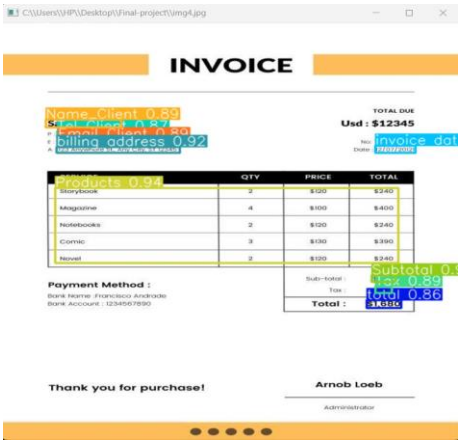


Fig.3: Data extraction

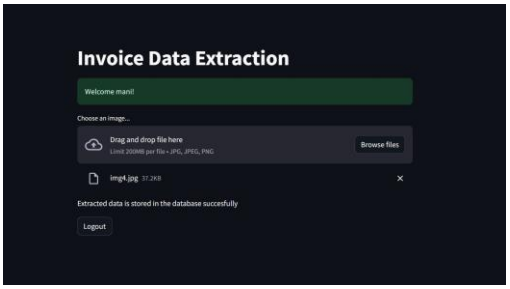
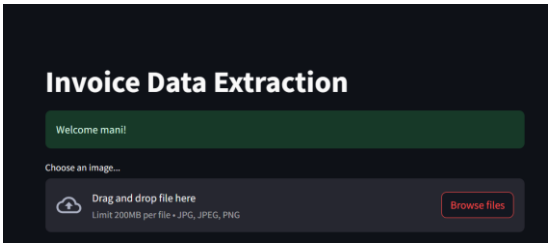


Fig.4: Extracted data is stored in database



5. Conclusion

OCR technology offers a compelling solution for automating invoice data extraction. It significantly boosts efficiency and speed compared to manual data entry, reducing processing time and resource allocation. Furthermore, OCR minimizes human error, leading to improved data accuracy. Additionally, OCR scales well, making it suitable for businesses processing high volumes of invoices. However, it's important to acknowledge limitations. OCR can struggle with invoices in various formats and may have difficulty recognizing handwritten text. To overcome these limitations and achieve maximum accuracy and flexibility, consider combining OCR with Machine Learning techniques.

6. Limitations:

The Limitations for the project may be asfollows:

1. Poor Image Quality:

Low resolution or blurry images hinder accuracy.

2. Color:

For black and white documents, capture images in grayscale or monochrome mode to improve OCR accuracy. For color documents,maintain color fidelity without oversaturation.

3. Format:

Save images in standard formats like JPEG or PNG to ensure compatibility with OCR software. Avoid compression that may degrade image quality.

7.Future Work 1.Adaptability to Variations:

Teach OCR to handle various invoice formats and styles effortlessly, whether it's handwritten or printed.

2. Faster Processing:

Develop OCR to extract data quicker, saving time and boosting efficiency in invoice processing.

3. Integration with AI:

Integrate OCR with AI algorithms for better contextual understanding and more precise data extraction.

4. Mobile Compatibility:

Make OCR technology available on mobile devices for on-the-go invoice scanning and data extraction.

5. User-friendly Interface:

Design a simple and intuitive interface for easy interaction with the OCR tool, making it accessible to users of all skill levels.

8.Acknowledgement

We would like to express our sincere gratitude to everyone who contributed to the successful completion of this project. First and foremost, we extend our heartfelt appreciation to the developers and contributors of the open-source tools and libraries that formed the backbone of our implementation.

We are deeply thankful to our project Guide Dr. V. Krishna Vijaya for their invaluable guidance, support, and

insightful feedback throughout the project's development. Their expertise and encouragement have been instrumental in shaping our approach and overcomingchallenges.

This project would not have been possible without the collective efforts and support of everyone mentioned above . We are truly grateful for their contributions and collaboration.

References

[1] Cristani, M., Bertolaso, A., Scannapieco, S., Tomazzoli; Future paradigms of automated processing of business documents; International Journal of Information Management; Volume 40, June 2018, Pages 67-75;
DOI:10.1016/j.ijinfomgt.2018.01.010

[2] Jaume, G., Ekenel, H.K., Thiran, J.P.: FUNSD: A Dataset For Understanding Noisy Scrutinized Documents.; In: International Conference on Document Analysis and Recognition Workshops (ICDARW 2019). Volume 2., IEEE (2019) 1–6;
DOI: 10.1109/ICDARW.2019.10029

[3] H. T. Ha · A. Horak;Information Extraction from Scrutinized Invoice Images using Text Analysis and Layout Features; arXiv;8 Aug 2022;
DOI:10.48550/arXiv.2208.04011

[4] Stanislawek, T., Gralinski, F., Wroblewska, A., Lipinski, D., Kaliska, A., Rosalska, P., Topolski, B., Biecek, P.: ”Kleister: Key information extraction datasets involving long documents with complex layouts”; Springer International; (2021)PP: 564–579;
DOI:10.1007/978-3-030-86549-8_36

[5] Harley, A.W., Ufkes, A., Derpanis, K.G.:t ;Evaluation of deep convolutional nets for document image classification and retrieval. In: 13th International Conference on Document Analysis and Recognition (ICDAR 2015), IEEE (2015)PP: 991–995;
DOI:10.1109/ICDAR.2015.7333910

[6] Arkhipov, M., Tromova, M., Kuratov, Y., Sorokin, A.: “Tuning Multilingual Transformers for Named Entity Recognition on Slavic Languages”;BSNL (2019)PP: 89-93;
DOI : 10.18653/v1/W19-3712

[7] Majumder, B.P., Potti, N., Tata, S., Wendt, J.B., Zhao,Q., Najork, M.: Representation learning for information extraction from form-like documents. In: proceedings of the 58th annual meeting of the Association for Computational Linguistics; (2020)PP: 6495-6504;
DOI: 10.18653/v1/2020.acl-main.580

[8] Lohani, D., Bela d, A., Bela d, Y.: An invoice reading system using a graph convolutional network. In: Asian Conference on Computer Vision, Springer (2018)PP: 144-158;
DOI:10.1007/978-3-030-21074-8_12

[9] Huang, Z., Chen, K., He, J., Bai, X., Karatzas, D., Lu, S., Jawahar, C.: Competition on Scrutinized Receipt OCR and Information Extraction. In: International Conference on Document Analysis and Recognition (ICDAR 2019), IEEE (2019)PP:1516-1520;
DOI : 10.1109/ICDAR.2019.00244

[10] Wu, X., Du, Z., Guo, Y.: A visual attention-based keyword extraction for document classification . In :Multimedia Tools and Applications 77(8) (2018)PP: 25355-25367;
DOI : 10.1007/s11042-018-5788-9

[11]Hamza, H., Belaid, Y., Belaïd, A.: A case-based logic approach for invoice

structure extraction. In: Document Analysis and Recognition, Ninth International Conference on. Volume 1., IEEE (2007) 327–331
DOI : 10.1109/ICDAR.4378726

[12]Bart, E., Sarkar, P.: Information extraction by finding repeated structure. In: Proceedings of the 9th International Workshop on Document Analysis Systems, ACM(2010) 175–182
DOI : 10.1006/ICDAR.26-789-09

[13]Lendak, I., Verma, C.: Invoice classification using deep features and machine literacy ways. In: IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT 2019), IEEE(2019) 855–859
DOI :10.1109/JEEIT.2019.8717504

[14]Lukasz Garncarek, Powalski, R., Stanis lawek, T., Topolski, B., Halama, P., Grali´nski, F.LAMBERT: LayoutAware (Language) Modeling using BERT for information extraction .
DOI : 10.1007/978-3-030-86549-8_34

[15].Krieger, F., Drews, P., Funk, B., Wobbe, T.: Information Extraction from Invoices: A Graph Neural NetworkApproach for Datasets with High Layout Variety. In: International Conference on Wirtschaftsinformatik.
DOI : :10.1007/978-3-030-86797-3_1

[16]Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation 9(8) (1997) 1735–1780
DOI : 110.1162/neco.1997.9.8.1735

[17] . Palm, R.B., Laws, F., Winther, O.: Attend, copy,parse end-to-end information extraction from documents.In: International Conference on Document Analysis andRecognition (ICDAR 2019), IEEE (2019) 329–336 .
DOI :0.1109/icdar.2019.00060

[18]Riba, P., Dutta, A., Goldmann, L., Forn´es, A., Ramos,O., Llad´os, J.: Data discovery in invoice documents by graph neural networks. In: International Conference On Document Analysis and Recognition (ICDAR 2019),IEEE (2019) 122–127.
DOI : 10.1109/ICDAR.2019.00028

[19]Tarawneh, A.S., Hassanat, A.B., Chetverikov, D.,Lendak, I., Verma, C.: Invoice categorization using deep features and machine literacy. In: IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT 2019), IEEE
(2019) 855–859
DOI : 10.1109/JEEIT.2019.8717504

[20].Esser, D., Schuster, D., Muthmann, K., Schill, A.: Few Exemplar Information Extraction for Business Documents. In: ICEIS (1). (2014) 293–298
DOI : 10.5220/0004946702930298