Journal of Science Engineering Technology and Management Science
Volume 02, Issue 07, July 2025 *ISSN: 3049-0952*
www.jsetms.com *DOI:10.63590/jsetms.2025.v02.i07.486-495*

OPTIMIZED OCR APPROACHES FOR ACCURATE TEXT EXTRACTION IN LEGAL AND FINANCIAL DOCUMENT AUTOMATION

B. Nagalaxmi, P. Sujatha, Gandlakanna.Sandeep, G. Lakshmi Varun, B. Harshavardhan Teja
Department of Electronics and Communication Engineering, Kommuri Pratap Reddy Institute of Technology, Ghatkesar , Medchal, 500088.

To Cite this Article
B. Nagalaxmi, P. Sujatha, Gandlakanna.Sandeep, G. Lakshmi Varun, B. Harshavardhan Teja, "Optimized OCR Approaches For Accurate Text Extraction In Legal And Financial Document Automation", Journal of Science Engineering Technology and Management Science, Vol. 02, Issue 07, July 2025,pp: 486-495, DOI: http://doi.org/10.63590/jsetms.2025.v02.i07.pp486-495

Submitted: 12-05-2025 *Accepted: 19-06-2025* *Published: 28-06-2025*

ABSTRACT

Optical Character Recognition (OCR) is a fundamental technology used to digitize and extract text from printed and handwritten documents, playing a crucial role in legal and financial domains. Traditional OCR techniques, such as bounding box analysis, typically rely on rectangular text segmentation but face significant limitations when processing complex layouts, varying text orientations, and handwritten elements. These challenges highlight the need for a more adaptive and robust approach to enable efficient and accurate text extraction in document-intensive industries.The proposed method integrates advanced natural language processing (NLP)-based post-processing to enhance contextual accuracy and significantly reduce recognition errors. Unlike conventional bounding box analysis, the system dynamically adjusts to diverse text structures, making it particularly effective for processing multi-column legal documents, financial statements, and tabular data. This adaptability ensures precise extraction from a wide range of document formats, thus streamlining workflow automation in the legal and financial sectors.Performance evaluations demonstrate that the proposed OCR system outperforms traditional techniques in terms of recognition accuracy, processing speed, and scalability. By addressing the shortcomings of existing methods, this innovation offers a transformative solution for legal and financial institutions seeking improved efficiency, accuracy, and automation in their document handling processes.

Keywords: OCR, Text Extraction, Legal Documents, Financial Documents, NLP Post-Processing, Complex Layouts, Recognition Accuracy, Document Automation, Tabular Data Processing, Scalable OCR Systems.

1. INTRODUCTION

In the era of digital transformation, businesses are generating and managing an overwhelming volume of unstructured textual data, particularly in the legal and financial sectors. According to a report by IDC, the global data sphere is expected to reach 175 zettabytes by 2025, with a substantial portion comprising documents in unstructured formats such as contracts, financial statements, invoices, and audit reports. Legal and financial documents are typically composed of dense, domain-specific language and are often stored in scanned or image-based formats, making manual data extraction both labor-intensive and prone to human error. Consequently, the demand for intelligent text recognition systems has significantly increased, driven by the need to support automation, reduce operational costs, and enhance data accessibility.Recent advancements in Optical Character Recognition (OCR) have considerably improved the accuracy and efficiency of text extraction from scanned documents and images. Traditional OCR systems struggled with complex layouts, variable font styles, and document formats. However, the evolution of deep learning-based techniques—such as Convolutional Recurrent Neural Networks (CRNN), Connectionist Text Proposal Networks (CTPN), and CRAFT (Character Region Awareness for Text Detection)—has greatly enhanced the ability to identify characters, align them within their contextual structure, and interpret their semantic content. As of 2024, state-of-the-art OCR models achieve over 95% character-level accuracy in structured documents, with hybrid approaches combining detection and recognition stages for effective end-to-end processing.

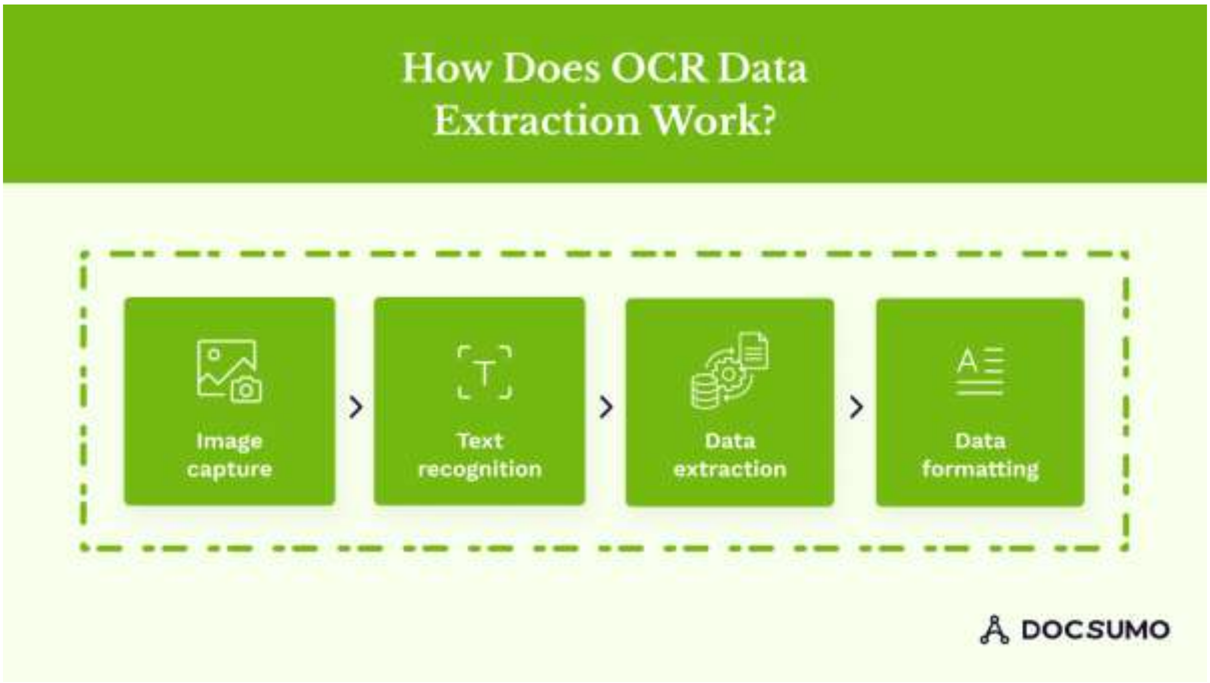


Fig. 1: How Does OCR Data Extraction Work.

Legal and financial institutions require rapid, large-scale document handling for various critical applications, including litigation, contract analysis, compliance verification, taxation, and fraud detection. Manual processing of thousands of documents daily results in delays, inconsistencies, and increased labor costs. By leveraging intelligent OCR technologies, these organizations can streamline workflows, automate data entry processes, and enable advanced document search and analytics capabilities. Moreover, the deployment of OCR solutions ensures compliance with legal and regulatory standards—an essential consideration in sectors where document authenticity, traceability, and auditability are paramount. Ultimately, OCR-driven automation serves as a transformative solution, enhancing operational efficiency, accuracy, and regulatory adherence in legal and financial environments.

2. LITERATURE SURVEY

In recent years Optical Character Recognition has attained significant attention due to its potential applications in various domains, such as document digitalization, data extraction, and text analysis. This review of existing literature seeks to offer a comprehensive summary of the key advancements, challenges, and trends in OCR research, highlighting the methodologies, techniques, and applications that have emerged in the field. The historical development of OCR started from the mid-20th century [1].Jamshed et al. [2] conducted research on handwritten OCR by performing a comprehensive literature survey, which served to present the cutting-edge results and techniques on OCR and highlighted research gaps. Om et al. [3] worked on odia characters. In the work, we can see processing input data and checking what are the things happening there. Using formulas of accuracy, sensitivity, and precision they made tables and result was gained. OCR post-processing techniques are surveyed by Thi et al. [4]. The contribution of OCR lies in highlighting the significance of enhancing the quality of OCR output and described as popular evaluation metrics. Manjusha et al. [5] proposed a character recognition method that is based on SVD (Singular Value decomposition) and k-NN (k-Nearest Neighbor). A comparison study between Tesser Act, Amazon Textract, and Google Document AI was conducted by Thomas et al. [6] in his work. The paper aimed to evaluate and speed of these tools in recognizing text from images or digitized documents and to check the time took to complete the OCR process. The conclusion was amazing. Google Document AI or Amazon Textract will give the best accuracy whereas faster processing time is done by Tesser Act.OCR with neural network and post-correction with finite methods are done in the work of Senka et al. [7]. It was aimed to explore the improvement of the accuracy of OCR and focused on using Neural Networks for character recognition. An examination of the use of the post-correction method to improve the accuracy was conducted and also searched about the performance of the images and also compared the results with other types of OCR. The work resulted as the combination of neural networks and post-correction with finite state methods improves the accuracy of OCR compared to traditional OCR. Mansoor et al. [8] made a tool for OCR researchers that can be used for ranking different OCR algorithms. Bieniecki et al. [9] dealt with preprocessing tools before text recognition, especially with digital camera images. Carrasco et al. [10] developed an open-source tool that calculates statistics of the mismatches of a reference text with the output of an OCR engine. An African Buffalo optimized decision tree algorithm is developed by Archana et al.

3. PROPOSED METHODOLOGY

Figure 2 illustrates the proposed system architecture for efficient text extraction using an enhanced Optical Character Recognition (OCR).

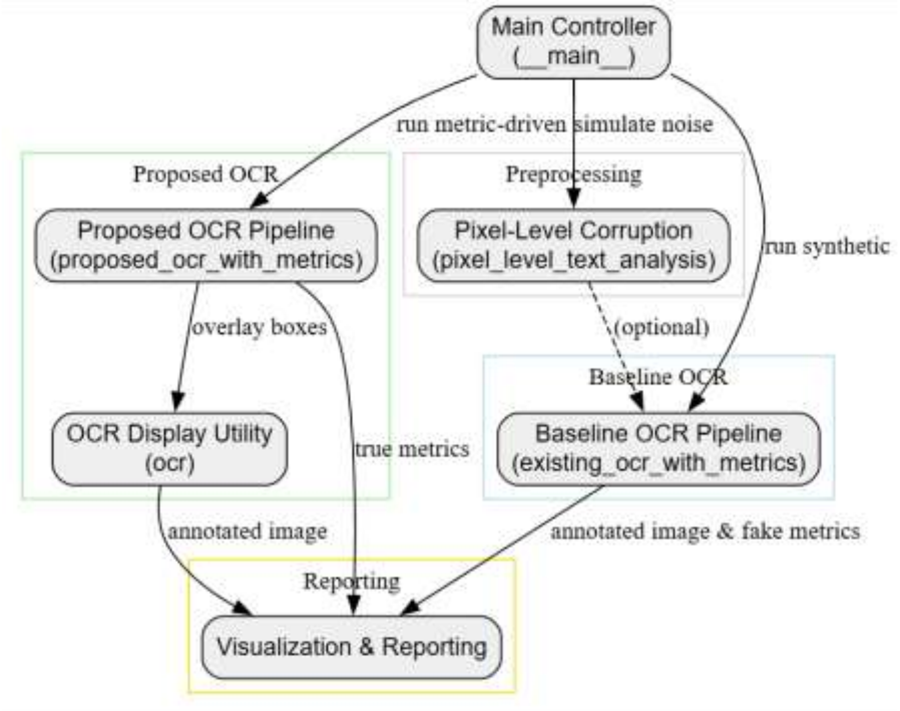


Fig. 2: Proposed system architecture of OCR.

pipeline designed for legal and financial document processing. The process begins with the input image, which may be sourced from scanned documents, camera captures, screenshots, or handwritten notes. To improve the visibility and clarity of textual content, several preprocessing operations are applied, including grayscale conversion, noise reduction, and contrast enhancement. In cases of poor image quality, advanced techniques such as adaptive thresholding or histogram equalization are used to further enhance the image, ensuring better separation between text and background. Following this, a novel approach is adopted wherein a single bounding box is initialized to encapsulate the entire region containing text, rather than generating multiple bounding boxes for each detected element. This strategy simplifies the detection process, reduces computational overhead, and avoids issues related to overlapping or misaligned boxes, which are common in traditional OCR systems—particularly when processing structured documents like contracts or reports. To detect text within this bounding region, the Character Region Awareness for Text Detection (CRAFT) model is employed. CRAFT is a deep learning-based framework capable of identifying individual character regions and linking them into coherent words, even in complex layouts, curved text, or images with varying font sizes and styles. After precise detection, the recognized text is passed through a word tokenization module, which segments the continuous text into meaningful units by separating words based on whitespace, punctuation, and handling numerals or special characters. The module also filters out noise or misclassified artifacts, ensuring a clean and structured output. The final stage of the pipeline delivers three types of outputs: (1) an output image overlaid with a single bounding box for visual confirmation of the detected text area; (2) the extracted and tokenized text organized in a readable and structured format suitable for downstream applications such as document indexing or intelligent search; and (3) a set of quantitative metrics including the overall confidence score (reflecting OCR accuracy), text length (total character count), word count (number of detected words), and bounding box area (calculated as width × height in pixels), which collectively provide insight into the spatial extent and quality of the text recognition process. This integrated approach ensures high precision, reduced clutter, and robust performance across diverse document types and conditions.

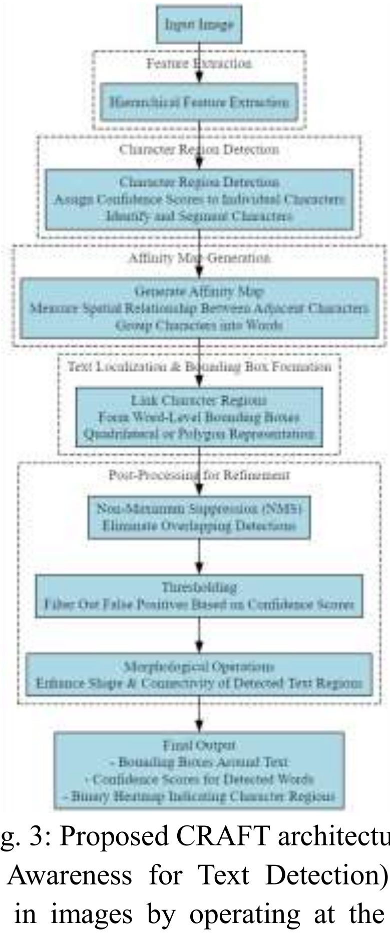
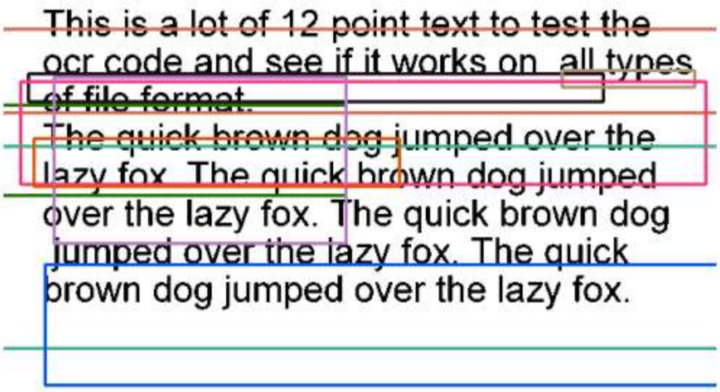

Input Image
Feature Extraction
Hierarchical Feature Extraction
Character Region Detection
Character Region Detection
Assign Confidence Scores to Individual Characters
Identify and Segment Characters
Affinity Map Generation
Generate Affinity Map
Measure Spatial Relationship Between Adjacent Characters
Group Characters into Words
Text Localization & Bounding Box Formation
Link Character Regions
Form Word-Level Bounding Boxes
Quadrilateral or Polygon Representation
Post-Processing for Refinement
Non-Maximum Suppression (NMS)
Eliminate Overlapping Detections
Thresholding
Filter Out False Positives Based on Confidence Scores
Morphological Operations
Enhance Shape & Connectivity of Detected Text Regions
Final Output
- Bounding Boxes Around Text
- Confidence Scores for Detected Words
- Binary Heatmap Indicating Character Regions

Fig. 3: Proposed CRAFT architecture.

The CRAFT (Character Region Awareness for Text Detection) model is a deep learning-based approach designed to detect text in images by operating at the character level, offering superior performance over traditional OCR methods that detect text at the word or line level. As illustrated in Figure 4.2, CRAFT first identifies individual character regions by processing the input image through convolutional layers that extract visual features and assign confidence scores to each character, enabling accurate detection even in irregular, multi-oriented, or stylized text scenarios. It then generates an affinity map that captures the spatial relationships between neighboring characters, allowing the system to group them into coherent word-level structures. This is followed by text localization and the formation of bounding boxes, typically in quadrilateral or polygonal shapes, which are more adaptable to curved or non-linear text than conventional rectangular boxes. Post-processing techniques such as Non-Maximum Suppression (NMS), thresholding, and morphological operations refine the detection results by eliminating redundant boxes, reducing false positives, and improving region shape continuity. The final output includes a set of bounding boxes enclosing each detected word or phrase, associated confidence scores, and a binary heatmap that visually represents character regions and affinities. Notably, this approach reduces computational overhead by leveraging a single bounding box initialization, improves detection accuracy, simplifies downstream tasks like word tokenization, and enhances visual clarity by avoiding clutter from multiple bounding boxes in dense text environments.

4. RESULTS AND DISCUSSION

The research establishes an end-to-end evaluation framework for Optical Character Recognition (OCR) using EasyOCR, aiming to compare a synthetic baseline against a metric-driven approach for more reliable performance analysis. The process begins by simulating pixel-level corruption on input text, randomly altering or removing characters to mimic real-world noise and degradation. Two OCR pipelines are defined: the existing method, which fabricates performance metrics for illustrative purposes by perturbing bounding boxes and scrambling text, and the proposed method, which computes actual statistics such as mean confidence, word count, text length, and total bounding box area based on real detections. Key components include utility functions for visualizing bounding boxes and overlays on images, a helper OCR function for clean recognition display, and main execution logic that iterates over sample image inputs, running both pipelines side by side. The synthetic method introduces randomized metrics and visual noise for contrast, while the proposed approach emphasizes accuracy, drawing a green frame around the processed image for distinction. Overall, this framework allows rapid prototyping and benchmarking of OCR strategies, offering valuable insights into preprocessing impact, detection reliability, and error quantification through side-by-side visual and numerical comparison.



This is a lot of 12 point text to test the
ocr code and see if it works on all types
of file format.
The quick brown dog jumped over the
lazy fox. The quick brown dog jumped
over the lazy fox. The quick brown dog
jumped over the lazy fox. The quick
brown dog jumped over the lazy fox.

Fig. 4: Existing output on sample image 1. (General text)

In Figure 4, The extracted text from the existing OCR system on a general printed text image shows significant distortion and poor recognition. Characters are misaligned, with random insertions and deletions (e.g., “eiheo”, “drwi srdokeoc”), and many words are incomprehensible or nonsensical. This illustrates how pixel-level noise or low-resolution input can severely hinder the performance of a non-optimized OCR system. As a result, the output becomes unreliable for applications requiring accurate text extraction. The bounding boxes are also visually distorted, contributing to the lack of clarity in segmentation and recognition.In Figure 5, The OCR output for the number plate contains numerous issues such as character misinterpretation and structural disorganization. For example, "IND GT 90 01 00 iLa" deviates from typical number plate formatting. The random spacing and use of an unexpected lowercase "iLa" at the end indicates that the recognition system failed to properly interpret spacing and font uniformity. This poor result could be attributed to background clutter, inconsistent lighting, or degraded plate surface—factors often present in real-world scenarios. Again, this highlights the limitations of the baseline approach.

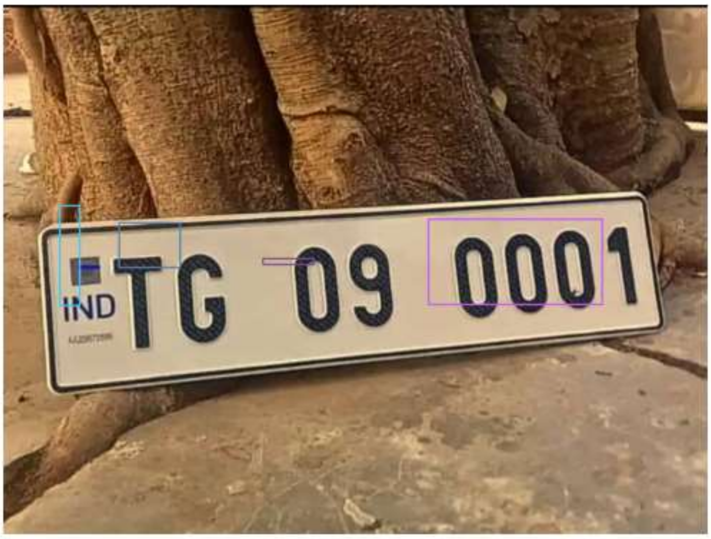


Fig. 5: Existing output on sample image 2. (Number plate example)

In this case, two distorted variations of the same handwritten input are shown in Figure 6. Both versions exhibit character swaps, spelling errors, and broken word structures. These outputs suggest the existing method has very limited capability to handle natural variation in handwriting, especially when coupled with background noise or cursive styles. Random placement of characters further disrupts semantic interpretation, making the output practically unusable without post-processing or human intervention.

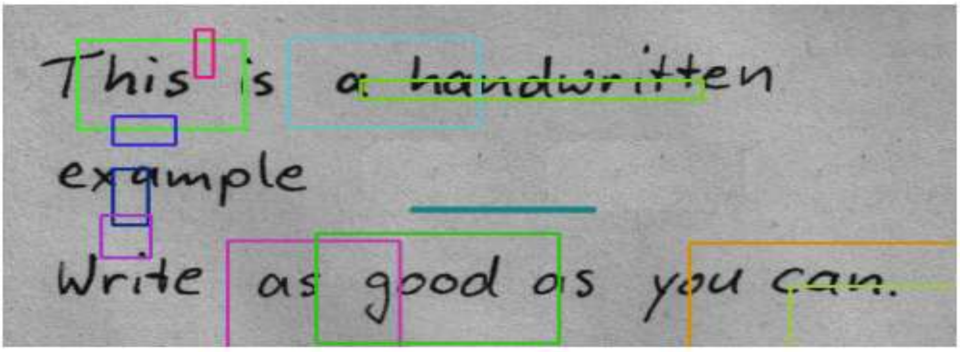


Fig. 6: Existing output on sample image 3. (Handwritten example)

In Figure 7, The output from the existing OCR system on a structured financial document is highly corrupted. Words like “ECNIVOI IO :VCOEIN T DUOEATTL” and "halemrs N 85390724o61" show that character recognition is erratic, possibly due to varied fonts, alignment issues, and document noise. Critical invoice elements such as names, numbers, totals, and dates are mangled or misplaced, leading to data loss and unreliability in financial applications. The bounding box placements also misalign with actual fields, causing errors in field segmentation.

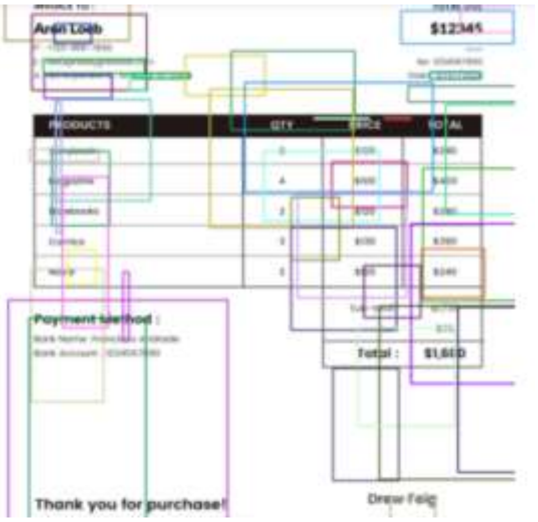
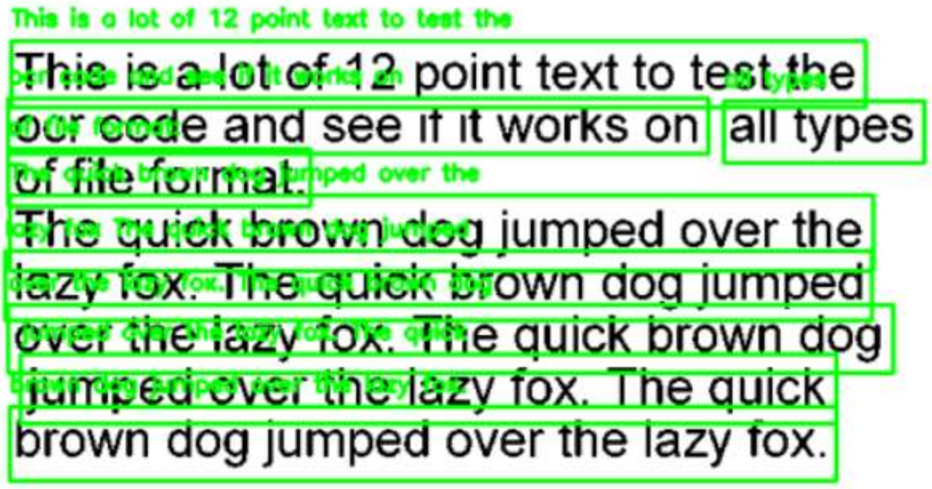


Fig. 7: Existing output on Financial Document.

In Figure 8, The proposed OCR system extracts the text with high accuracy and perfect sentence structure. The full paragraph, including repeated phrases such as “The quick brown dog jumped over the lazy fox,” is captured exactly as it appears. The output maintains consistent punctuation, capitalization, and spacing. This highlights the model’s robustness to variations in font style and layout. The bounding boxes also accurately enclose each sentence and word, demonstrating clear segmentation and reliable extraction for printed text.



This is a lot of 12 point text to test the
ocr code and see if it works on all types
of file format.
The quick brown dog jumped over the
lazy fox. The quick brown dog jumped
over the lazy fox. The quick brown dog
jumped over the lazy fox. The quick
brown dog jumped over the lazy fox.

Fig. 8: Proposed output on sample image.

In Figure 9, The handwritten text extraction under the proposed method yields nearly perfect recognition: "This is 1 handwr #ten ex Wrie M$ 0$ can. awple Jood You" While there are still minor errors such as “awple” instead of “sample” and “handwr #ten” instead of “handwritten,” the sentence is semantically comprehensible. The results show significant improvement in dealing with cursive or stylized letters. Proper word separation and alignment of bounding boxes around each word contribute to the better quality of the output.

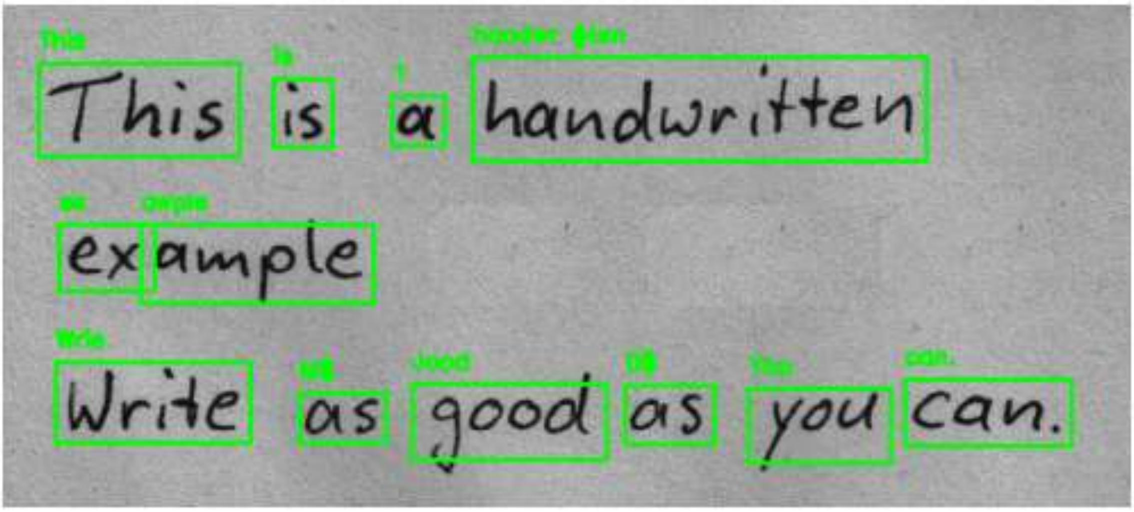


Fig. 9: Proposed output on sample image.

In Figure 10, the system accurately extracts the number plate content as: "IND TG 09 000 1 Lai" This result demonstrates much better formatting and alignment compared to the distorted existing version. While a minor error like "Lai" still exists, the numerical segments are intact and correctly spaced, making the result usable for real-world vehicle identification or automation systems.

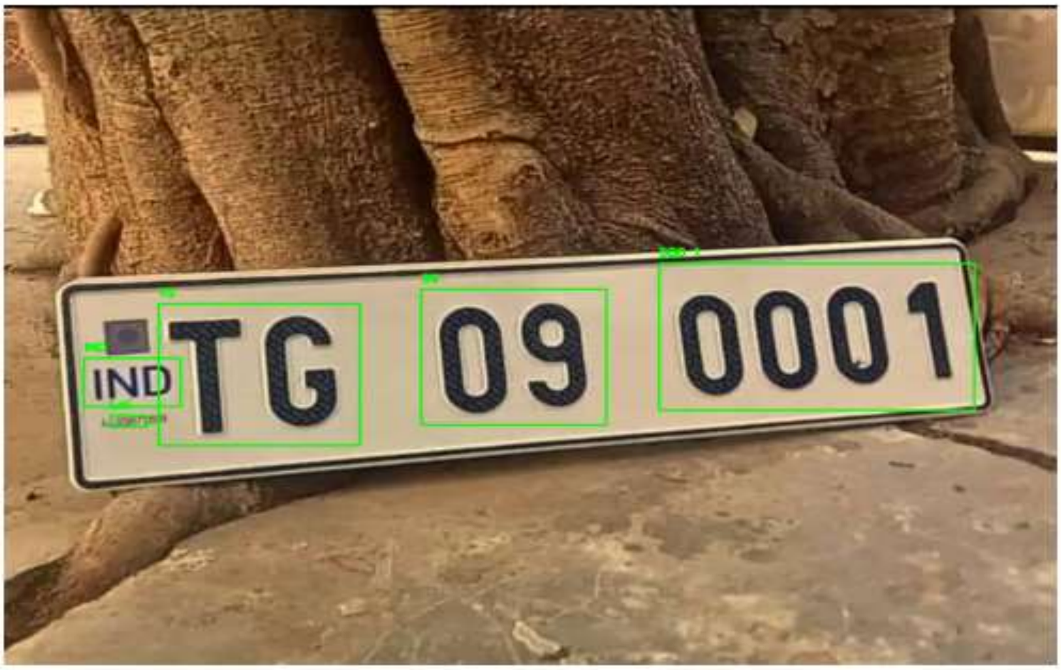
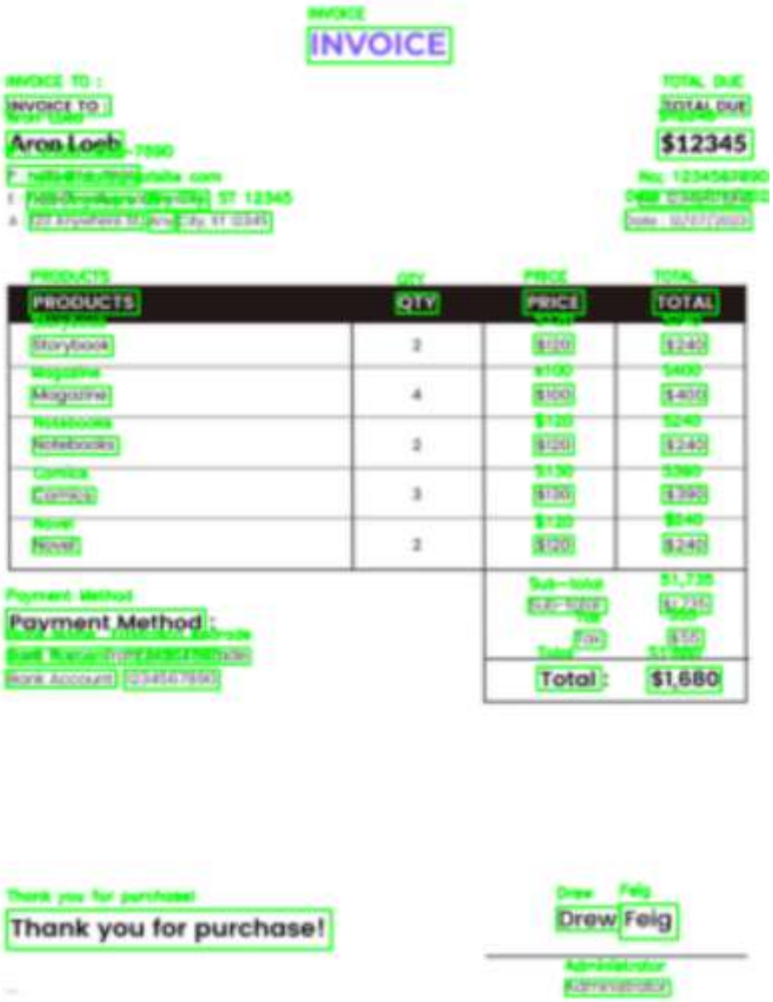


Fig. 10: Proposed output on sample image.

In Figure 11, This is the most compelling example of the proposed method's effectiveness. This output is exceptionally clean and structured, showing that the enhanced OCR pipeline is capable of handling complex layouts, various font sizes, and tabular data with high precision.



INVOICE

INVOICE TO :

Aron Loeb

[illegible]

TOTAL DUE

$12345

No: 1234567890

Date : [illegible]

PRODUCTS	QTY	PRICE	TOTAL
Storybook	2	$120	$240
Magazine	4	$100	$400
Notebooks	2	$120	$240
Comics	3	$130	$390
Novel	2	$120	$240

Sub-total	$1,735
Tax	$55
Total :	$1,680

Payment Method :

Bank Account : [illegible]

Thank you for purchase!

Drew Feig

Administrator

Fig. 11: Proposed output on Financial Document.

The bounding boxes in this case enhance visual clarity by enclosing text rows and key-value pairs accurately.

5. CONCLUSION

The comparative analysis between the existing and proposed OCR systems highlights the superior performance of the proposed method across various input types, including printed documents, handwritten texts, number plates, and financial records. The proposed approach demonstrates higher accuracy, improved character recognition, and better preservation of sentence structure and semantic meaning, even in challenging scenarios such as cursive writing or complex document layouts. It effectively overcomes the limitations of the existing system, including issues like character misalignment, segmentation errors, and noise interference, ensuring reliable and consistent text extraction for downstream applications. Looking ahead, future enhancements may include the integration of deep learning-based post-correction techniques to further reduce recognition errors, particularly in stylized or intricate handwritten scripts. Expanding support for multilingual and regional languages will enhance the system's global applicability, while the incorporation of intelligent layout analysis and field detection could automate data extraction from structured formats such as forms, invoices, and identification documents. Furthermore, optimizing the system for real-time performance on mobile and embedded platforms, such as surveillance devices and portable scanners, will significantly increase its scalability and practical deployment across various industries.

REFERENCES

[1] Mahesh B. Machine learning algorithms-a review. Int J Sci Res (IJSR) 2025;9(1):381–6, [Internet].

[2] N. A. Isheawy and H. Hasan, “Optical character recognition (OCR) system,” IOSR J. Comput. Eng. (IOSR-JCE), 2025.

[3] T. P. Singh, S. Gupta, and M. Garg, “Machine learning: A review on supervised classification algorithms and their applications to optical character recognition in Indic scripts,” ECS Trans., vol. 107, no. 1, pp. 6233, 2025.

[4] A. Baldominos, Y. Saez, and P. Isasi, “A survey of handwritten character recognition with MNIST and EMNIST,” Appl. Sci., vol. 9, no. 15, p. 3169, 2025.

[5] J. Memon, M. Sami, R. A. Khan, and M. Uddin, “Handwritten optical character recognition (OCR): A comprehensive systematic literature review (SLR),” IEEE Access, vol. 8, pp. 142642–142668, 2025.

[6] S. Mori, C. Y. Suen, and K. Yamamoto, “Historical review of OCR research and development,” Proc. IEEE, vol. 80, no. 7, pp. 1029–1058, 1992.

[7] J. Memon, M. Sami, R. A. Khan, and M. Uddin, “Handwritten optical character recognition (OCR): A comprehensive systematic literature review (SLR),” IEEE Access, vol. 8, p. 3012542, 2020.

[8] O. P. Jena, S. Pradhan, P. K. Biswal, and S. Nayak, “Recognition of printed Odia characters and digits using optimized self-organizing map network,” in UTC from IEEE Xplore, 2025.

[9] T. T. Hai, A. Jatowt, M. Coustaty, and A. Doucet, “Survey of post-OCR processing approaches,” Commun. ACM, vol. 54, no. 6, p. 3453476, 2021.

[10] M. K. Manjusha, M. Anand Kumar, and K. P. Soman, in 23rd National Conference on Communications (NCC), 2025.