International Journal of Research in Engineering and Innovation (IJREI)

Journal home page: https://ijrei.com/

RESEARCH PAPER

Automated Document Data Extraction Using OCR and RPA

Abhishek Soam, Shobhit Saini, Taran Malik, Aman Chaudhary, Punit Mittal

Department of Information Technology, Meerut Institute of Engineering and Technology, Meerut, India

Article Information

Received: 15 April 2026
Revised: 01 May 2026
Accepted: 03 May 2026
Available online: 04 May 2026

Keywords:

Optical Character Recognition
Robotic Process Automation
Convolutional Neural Networks
Image Acquisition
RPA-Based Automation

Abstract

In the current scenario, the process of document processing is mostly tedious, time-consuming, and error- prone. Document data extraction is one such process in the current organizational scenario. It is mostly tedious and time consuming while working with huge amounts of scanned documents and images. This paper describes the design and development of the document data extraction system using the combined technology of Optical Character Recognition (OCR) and Robotic Process Automation (RPA). It can be effectively used to improve the efficiency and accuracy of data processing activities. The OCR part of the proposed document data extraction system is based on deep learning technology. It can extract data from different types of documents such as invoices, forms, and book pages. Preprocessing techniques can be applied to the data to improve its readability. In the RPA part, the data is automatically transferred to the required application such as Notepad and databases. Image preprocessing techniques can be applied to the document images to improve the efficiency of the document data extraction using the OCR part. Using the combined technology of OCR and RPA, the unstructured visual data can be easily converted into structured digital data. From the experimental results, it is evident that the proposed document data extraction system is highly efficient in terms of time consumption and achieves high accuracy in data extraction.

1. Introduction

Document data extraction refers to the transformation of unstructured visual information into a structured, machine-readable format. Across industries such as banking, healthcare, and education, a vast number of documents are processed daily, creating a strong need for efficient and reliable extraction methods. However, conventional approaches primarily based on rule-based and template-driven systems often fail to perform effectively in real-world scenarios due to variability in document formats and quality. Optical Character Recognition (OCR) plays a critical role in this domain by enabling the extraction of textual content from images and scanned documents [1]. Traditional OCR techniques, which rely on predefined rules and templates, are limited in handling complex or noisy data. In contrast, advanced approaches leveraging deep learning, particularly Convolutional Neural Networks (CNNs), offer significantly improved performance and adaptability. Despite these advancements, OCR systems are generally restricted to text extraction and do not inherently support further data processing or workflow automation [2].

Robotic Process Automation (RPA), on the other hand, is widely used to automate repetitive, rule-based tasks but is typically limited to structured, digital data. It cannot directly process unstructured visual inputs. As a result, combining OCR with RPA has emerged as a practical solution for document-driven automation workflows [3].

To address these limitations, a hybrid framework integrating OCR-based text extraction with RPA-driven automation is proposed. This system operates through a unified pipeline that enhances both efficiency and accuracy. Initially, image preprocessing techniques are applied to improve input quality.

Corresponding author: Abhishek Soam
Email Address: abhishek.soam.csit.2022@miet.ac.in
https://doi.org/10.36037/IJREI.2026.10204

Subsequently, deep learning-based OCR extracts textual information from diverse document types. The extracted data is then cleaned and structured to remove noise and ensure usability. Finally, RPA scripts automate data entry into applications such as spreadsheets or databases. By seamlessly integrating these components, the proposed approach reduces human intervention, minimizes errors, and significantly improves processing efficiency in large-scale document handling systems [4].

This work proposes a hybrid framework that combines image preprocessing, Optical Character Recognition (OCR), and Robotic Process Automation (RPA) to automate document data extraction and processing. The goal is to transform unstructured visual data into structured, machine-readable formats that can be directly utilized in digital systems [5]. Each stage of the system plays a critical role in ensuring accuracy, efficiency, and scalability.

1.1 Image Acquisition

Image acquisition is the first and most fundamental stage of the system. In this phase, document images are collected from multiple sources such as scanned documents, book pages, invoices, receipts, and structured or semi-structured forms. These inputs may vary significantly in quality, resolution, lighting conditions, orientation, and format, which directly impacts downstream processing.

The acquired image is mathematically represented as:

$$I = \{p_1, p_2, p_3, \ldots\ldots\ldots\ldots\ldots\ldots p_n\}$$

Where each p_i denotes an individual pixel value, and n represents the total number of pixels in the image. This representation highlights that a document image is essentially a collection of pixel intensities, which must be processed to extract meaningful textual information.

A key challenge at this stage is handling diverse input formats and ensuring that images are properly captured without distortion. Poor-quality images can significantly degrade OCR performance, making preprocessing an essential next step.

1.2 Image Preprocessing

Image preprocessing aims to enhance the quality of the acquired image to improve the accuracy of OCR. Real-world document images often contain noise, uneven illumination, skewness, and background artifacts that hinder text recognition.

The preprocessing stage is represented as:

$$I_{processed} = f(I)$$

where $f(\cdot)$ denotes a series of image enhancement operations.

Key preprocessing techniques include grayscale conversion, noise removal, thresholding, skew correction, and contrast enhancement. Grayscale conversion transforms a colored image into a single-channel intensity image, reducing computational complexity and allowing the system to focus on textual features rather than color variations. Noise removal, or denoising, eliminates unwanted disturbances such as salt-and-pepper noise and blur through filtering techniques like Gaussian or median filters, thereby improving image quality. Thresholding, also known as binarization, converts the grayscale image into a binary format where text appears distinctly against the background, facilitating better character recognition. Skew correction addresses issues of misalignment by properly orienting tilted or rotated documents, ensuring that text lines are horizontally aligned. Additionally, contrast enhancement improves the visibility of faint or low-contrast text by adjusting intensity levels, making characters more distinguishable [6]. Collectively, these preprocessing steps significantly enhance text clarity, minimize distortions, and ensure that the OCR system receives a clean, high-quality input for accurate extraction.

1.3 Text Processing and Cleaning

After OCR extracts text from the processed image, the output often contains errors such as misrecognized characters, extra symbols, spacing issues, and formatting inconsistencies. Therefore, a text processing and cleaning phase is necessary to refine the extracted content.

The cleaning process is expressed as:

$$T_{clean} = g(T)$$

where T is the raw OCR output and $g(\cdot)$ represents the text cleaning function.

Text processing and cleaning involve several essential steps to refine the raw OCR output into accurate and usable data. Unwanted characters such as special symbols, noise, or irrelevant elements introduced during OCR are first removed to improve clarity. This is followed by the correction of OCR errors, where common misrecognized characters are fixed using dictionaries, pattern matching, or rule-based techniques. Sentence formatting is then applied to adjust spacing, punctuation, and capitalization, ensuring the text is readable and grammatically consistent. Subsequently, data structuring organizes the cleaned text into predefined formats such as tables, key-value pairs, or structured fields like names, dates, and amounts, making it suitable for further processing. Normalization is also performed to standardize formats for dates, numerical values, and currency, ensuring consistency across the dataset. This stage is critical because even highly advanced OCR systems can produce imperfect outputs, and clean, well-structured data is essential for reliable automation and accurate analysis.

1.4 RPA-Based Automation

Once the text is cleaned and structured, Robotic Process Automation (RPA) is used to automate repetitive and rule-based tasks. RPA tools mimic human interactions with

software systems, enabling seamless data entry and processing without manual intervention.

The automation process is represented as:

$$Output=RPA(T_{clean})$$

where T_{clean} is the processed data and RPA (·) represents the automation workflow.

1.5 Final Output

The final stage of the system involves storing the processed and structured data in user-friendly and accessible formats. The output can be saved in multiple forms depending on application requirements, such as:
The final output of the system is stored in multiple formats to ensure flexibility and usability across different applications. Text files created using software such as Notepad are suitable for simple storage, quick access, and basic documentation of extracted information. For more advanced handling, Microsoft Excel spreadsheets enable structured tabular representation, support calculations, and allow further data analysis through formulas and visualization tools. In addition, databases such as MySQL or Oracle Database provide robust solutions for large-scale data storage, efficient querying, and seamless integration with enterprise-level applications.

2. Methodology

The methodology for automating document data extraction proposed in this research combines image preprocessing, OCR and RPA under one umbrella. Thus, with the amalgamation of these elements, the system quickly digitizes unstructured document images into structured data/records such as PDF format and digitizes further processing steps. The preprocessing step enhances image quality, which in turn increases the accuracy of text extraction performed by OCR. I would also implement deep learning-based OCR models, which enable the system to handle a diverse array of document formats and real-world variations. The methodology flow chart is shown in Fig. 1.
RPA enables automation of repetitive processes including data entry and interaction within applications increasing efficiency and reducing human error. Then the complete pipeline allows a constant passage from input image to final output, which makes the system robust and scalable to practical usage.
Experimental results demonstrate that the integrated approach achieves superior efficiency and accuracy compared to standalone OCR or manual processing methods.
In conclusion, the presented OCR + RPA-based approach provides a practical approach for automating document processing tasks. It has a great potential for real-world usage in many industrial domains like banking, healthcare and administrative systems where extremely large number of documents need to be processed very quickly and accurately. Detection and recognition of handwritten text, support for more languages, and cloud service interaction with scalability are some possible future developments.

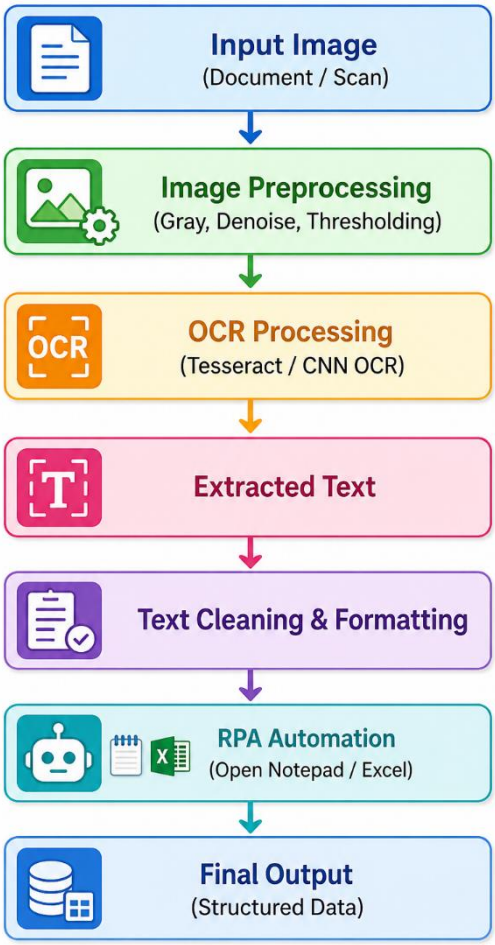

Figure 1: Flowchart for Methodology

2.1 Experimental analysis

Multiple document images are utilized to evaluate the performance of the OCR + RPA-based document data extraction system, including book pages, printed documents, and scanned forms. These inputs represent diverse real-world scenarios, containing variations in format and quality such as low-resolution images, noisy scans, and standard printed text. The system processes these documents through preprocessing, OCR-based text extraction, and RPA-driven automation to assess its robustness and adaptability. The performance of the proposed system is measured in terms of accuracy, processing time, and overall efficiency, and the results are compared with traditional OCR approaches and manual data entry methods to demonstrate improvements.
To ensure comprehensive evaluation, the system is tested using multiple performance metrics. Accuracy measures how effectively the system extracts correct textual information from documents.

$$\text{Accuracy}= \frac{Correctly\ Extracted\ Words}{Total\ Words}$$

Processing time evaluates the total time required to complete both OCR extraction and RPA automation processes,

reflecting the system's efficiency in handling large volumes of data.

Time=OCR Time+Automation Time

Error rate quantifies the proportion of incorrectly extracted words relative to the total number of words, indicating the reliability of the system.

$$\text{Error Rate} = \frac{Incorrect\ Words}{Total\ Words}$$

Together, these metrics provide a comprehensive assessment of system performance, highlighting its effectiveness in accurately extracting, processing, and automating document data under varying conditions.

2.2 Accuracy Comparison

The comparative analysis of different document data extraction methods clearly demonstrates the effectiveness of the proposed system. Manual data entry achieves an accuracy of 80%, which reflects reasonable reliability but is highly dependent on human effort and is prone to fatigue-related errors. Basic OCR using Tesseract OCR records a slightly lower accuracy of 78%, mainly due to its limitations in handling noisy, low-quality, or complex document formats. When preprocessing techniques are incorporated with OCR, the accuracy improves to 85%, indicating that image enhancement steps such as denoising and binarization significantly contribute to better text recognition. However, the highest accuracy of 91% is achieved by the proposed system, which integrates OCR with deep learning (CNN) and RPA-based automation. This improvement highlights the ability of advanced models to handle diverse document conditions and the role of automation in reducing human-induced errors. The comparative Accuracy Analysis of Document Data Extraction Methods is shown in Table 1.

Table 1: Comparative Accuracy Analysis of Document Data Extraction Methods

Method	Accuracy
Manual Data Entry	80%
Basic OCR (Tesseract)	78%
OCR + Preprocessing	85%
Proposed System (OCR + RPA + CNN)	91%

Fig. 2 represents a comparative analysis of accuracy achieved by different document data extraction methods. Manual data entry shows an accuracy of 80%, indicating moderate reliability but with dependence on human effort and the possibility of errors. The basic OCR method achieves an accuracy of 78%, which is slightly lower due to its limitations in handling noisy or complex document formats. When preprocessing techniques are applied along with OCR, the accuracy improves to 85%, demonstrating the positive impact of image enhancement methods on text recognition. The proposed system, which integrates OCR with preprocessing, Convolutional Neural Networks, and Robotic Process Automation, achieves the highest accuracy of 91%. This significant improvement highlights the effectiveness of combining advanced learning techniques with automation to enhance both precision and consistency. Overall, the figure clearly shows that the proposed system outperforms traditional and semi-automated approaches in terms of accuracy.

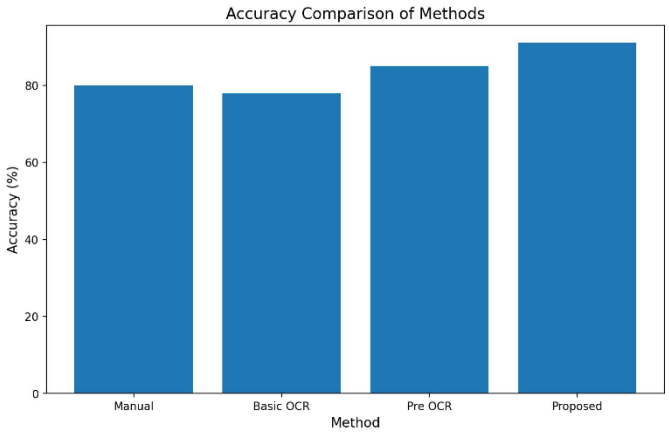

Figure 2: Accuracy comparison of document data extraction methods

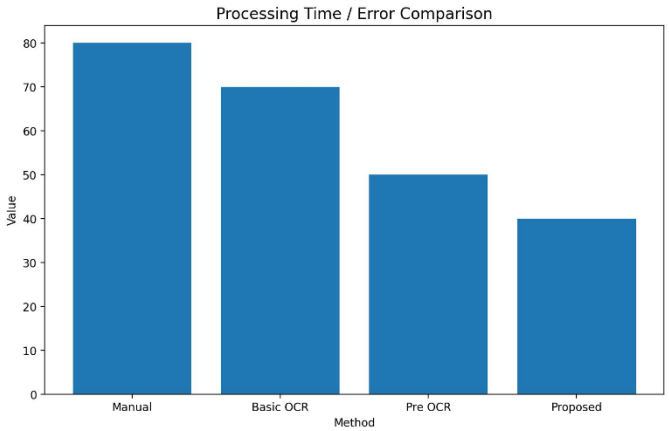

Figure 3: Comparative Analysis of Processing Time Across Document Data Extraction Methods

Fig. 3 illustrates the comparison of processing time required by different document data extraction methods. Manual data entry exhibits the highest processing time with a value of 80, reflecting the slow and labor-intensive nature of human-based operations. Basic OCR reduces the processing time to 70, showing some improvement due to automation, but it still requires additional corrections and validation. When preprocessing techniques are integrated with OCR, the processing time further decreases to 50, indicating that improved input quality helps in faster and more efficient text extraction. The proposed system demonstrates the lowest processing time with a value of 40, highlighting the efficiency of combining OCR with advanced techniques and RPA-based automation. This significant reduction in processing time indicates that the proposed approach not only enhances speed but also ensures streamlined and automated workflow.

The observations from the study clearly highlight the effectiveness of integrating preprocessing, advanced OCR techniques, and automation. Image preprocessing plays a crucial role in improving OCR accuracy by enhancing image quality and reducing noise, which directly impacts text recognition performance. CNN-based OCR demonstrates superior performance, particularly when dealing with real-world images such as low-resolution photos or noisy scans, as it can learn complex patterns and variations more effectively than traditional methods. Additionally, the use of Robotic Process Automation significantly reduces repetitive manual tasks by automating data entry and workflow execution, thereby minimizing human effort and errors.

3. Conclusion

Automated document data extraction is a challenging task due to different formats of documents, changing image quality, noise, and complex layouts in real-life scenarios. Manual data entry was slow, labour-intensive and highly prone to human error, while dedicated Optical Character Recognition (OCR) systems often lack reliable performance when it comes to processing low-quality or unstructured data. To address these limitations, in this paper we present a hybrid architecture that merges image preprocessing techniques with deep learning based
optical character recognition models and RPA (Robotic Process Automation) for achieving efficient and accurate document processing.

The OCR module successfully extracted text from many types of documents, provided preprocessing was used (like converting to grayscale, reducing noise and adaptive thresholding). These preprocessing procedures significantly improved the quality of input photos, which achieved better accuracy in word recognition. The CNNbased OCR adopted in the system, also made it capable of learning relevant patterns and variances captured in both typefaces as well as layouts for real-world document scans unlike text lines from purely controlled settings. Conversely, the RPA module played a pivotal role in mechanizing routine and rules-based processes such as data input, application engagement, and output generation. RPA emulated human operations to ensure seamless integration of extracted text into applications such as Notepad, Excel and databases. It involved no manual intervention, significantly reduced operational time whilst considerably decreasing errors. What this did is that it led to an uninterrupted stream where visual unstructured data was converted into structured and useable digital information—the combination of OCR RPA.

Results from the experimental study indicated that the proposed system outperforms traditional manual processing and standalone OCR systems in accuracy, response time, and reliability. As a result, the combination of preprocessing, smart text extraction, and automation enhanced overall efficiency as well as made systems more robust to noisy and complex inputs. The results show how actual document processing challenges are solved through integrated data extraction and automation technology.

References

[1] B. Shi, X. Bai and C. Yao,"AnEnd-to-End Trainable Neural Network for Image-Based Sequence Recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 11, pp. 2298–2304, 2017.

[2] Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In Proceedings of the 23rd International Conference on Machine Learning (ICML '06) (pp. 369–376). Association for Computing Machinery. https://doi.org/10.1145/1143844.1143891

[3] Singh, A., & Garg, S. K. (2023). Comparative study of optical character recognition using different techniques on scanned handwritten images. In Micro-Electronics and Telecommunication Engineering. Springer Nature Singapore.

[4] R. Girshick, "Fast R-CNN," 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 2015, pp. 1440-1448, doi: 10.1109/ICCV.2015.169.

[5] Afrin, S., Roksana, S., & Akram, R. (2025). AI-enhanced robotic process automation: A review of intelligent automation innovations. IEEE Access, 13, 1–17. https://doi.org/10.1109/ACCESS.2024.3513279

[6] Tupsakhare, P. (2025). Intelligent automation: Integrating AI and RPA for smarter processes. International Journal on Science and Technology, 16(1). https://doi.org/10.71097/IJSAT.v16.i1.2833

Cite this article as: Abhishek Soam, Shobhit Saini, Taran Malik, Aman Chaudhary, Punit Mittal, Automated Document Data Extraction Using OCR and RPA, International Journal of Research in Engineering and Innovation Vol-10, Issue-2 (2026), 56-60. *https://doi.org/10.36037/IJREI.2026.10204*