© 2025 ADYPJIET, Volume 1(1), Online ISSN: 3049-1517, December 2024, pp. 15-24

Ajeenkya DY Patil Journal of Innovation in Engineering & Technology

Journal Homepage: https://www.adypsoe.in/adypjiet

An Efficient OCR System for Extracting Information from Indian Identity Documents using CTP.

[1] Vridhi Sachdev, [2] Chirag Sandil, [3] Abhaysinh Landge, [4] Ashish Kolhe,
[5]Prof Bhagyashree Dhakulkar

[1-4] UG Student, Department of Artificial Intelligence and Data Science, Engineering
Ajeenkya D Y Patil School of Engineering, Pune, India
[5]Ajeenkya D Y Patil School of Engineering, Pune, India

Article History:

Received: *10-01-2025*

Revised: *25-01-2025*

Accepted: *24-02-2025*

Abstract:

This paper presents an innovative Optical Character Recognition system designed to extract information from three prominent Indian identity documents like Aadhar Card, PAN Card, and GST Certificate. The system's methodology leverages Connectionist Text Proposal Network to efficiently process and extract relevant data, thereby facilitating the Know Your Customer process in Logistics industry. The proposed OCR model aims to streamline and automate the extraction of customer data from these identity documents, enhancing operational efficiency and accuracy in identity verification processes. Through a detailed description of our methodology and experimental results, this research elucidates the effectiveness and reliability of our approach in handling diverse Indian identity documents. The findings underscore the potential of the proposed system to contribute significantly to the improvement of identity verification procedures, particularly in contexts requiring seamless integration of digital technologies for compliance and regulatory purposes.

Keywords: Optical Character Recognition, Know Your Customer, Logistics, Connectionist Text Proposal Network.

1.0 Introduction

Identity verification is fundamental in modern logistics operations, where trust and reliability are paramount for successful transactions. With the proliferation of digital technologies, traditional methods of verifying customer identities have evolved, prompting the integration of advanced

solutions such as Optical Character Recognition (OCR) systems. These systems, empowered by artificial intelligence and deep learning techniques, offer unparalleled accuracy and efficiency in extracting crucial information from identity documents, thereby streamlining the verification process. In response to the escalating need for enhanced identity verification mechanisms, our research endeavors to present a comprehensive solution tailored for the logistics industry. Collaborating with a logistics company, we embarked on the development of an OCR model designed to extract pertinent data from three essential identity documents: Aadhar cards, PAN cards, GST certificates, and Know Your Customer (KYC) documents. These documents, integral to the regulatory and operational framework of logistics enterprises, contain vital information such as Aadhar numbers, PAN numbers, GST numbers, and other KYC details, which are imperative for customer authentication, compliance, and Know Your Customer (KYC) obligations.

The significance of our research lies not only in the practical application of OCR technology but also in its potential to revolutionize the identity verification landscape within the logistics sector. By harnessing a blend of cutting-edge technologies including Python, Flask, HTML, CSS, Bootstrap, and Tensor Flow, coupled with sophisticated deep learning architectures such as the Connectionist Text Proposal Network (CTPN) and Convolutional Neural Networks (CNN), our model aims to transcend the limitations of traditional OCR systems. The integration of Flask, a lightweight web framework, facilitates the development of a user-friendly interface, enabling seamless interaction with the OCR model. Through the utilization of HTML, CSS, and Bootstrap, we ensure a visually appealing and responsive front-end design, enhancing the user experience and accessibility. Tensor Flow, a renowned deep learning framework, forms the backbone of our model, empowering it to learn complex patterns and structures inherent in identity documents. Our research is underpinned by a commitment to academic rigor and scientific excellence. Adhering to stringent guidelines, we strive to demonstrate the significance of our contributions relative to the state of-the-art literature in the field. By conducting proper comparison against established algorithms, employing non-parametric statistical analysis, and adhering to fair parameter tuning practices, we aim to substantiate the efficacy and superiority of our approach. Moreover, our research extends beyond mere technical innovation, delving into the practical implications and real-world applicability of our OCR model. Through detailed analysis and comparison with recent works, we seek to elucidate why our proposed algorithm excels in addressing the unique challenges posed by identity verification, compliance, and Know Your Customer (KYC) obligations in the logistics domain. Leveraging public domain datasets, we ensure transparency and reproducibility, facilitating fair comparisons and fostering advancements in the field. In essence, this paper endeavors to provide a comprehensive overview of our research journey, from conceptualization to implementation, highlighting the methodological intricacies, technical innovations, and scientific contributions underlying our OCR model for logistics identity verification and Know Your Customer (KYC) compliance.

2.0 Literature review:

The Literature Review section presents an overview of research in the different techniques

employed for text detection and recognition from various types of documents, drawing upon findings from 17 relevant papers. These studies offer diverse insights and methodologies that contribute to our understanding of the subject. By summarizing these findings in Table I, the current state of knowledge in the field is elucidated, highlighting areas for further inquiry. This review serves to inform the direction of subsequent research and provides context for the study at hand. Ultimately, the goal is to contribute to the advancement of knowledge in Optical Character Recognition techniques by synthesizing and analyzing existing literature.

Comparative Study of Literature Review

Reference	Dataset	Techniques Employed	Accuracy / Score
[1]	Synthetic dataset of 6,000,000 class-balanced and 500,000 CIC-similar samples	ID card region detection (IRDM), watermark removal (WRM), key text locating and recognition (KLRM), and text correction (TCM)	99.71 %
[2]	36,000 ID numbers with and without colored backgrounds	Differentiable Binarization (DB), CNN and BiLSTM	100%
[3]	2000 scanned and camera-captured document images	CRNN architecture, employing a sequential combination of CNN and BiLSTM	Train set – 99.26% Validation set – 98.33% Test set – 98.71%
[4]	2,500 Vietnamese ID card images	2 stage approach: text detection - pretrained multi-language CTPN and text recognition - custom CRNN model	78.0%
[5]	50 ID cards of two types, 25 scanned images, and 25 camera images	Pytesseract OCR followed by post processing using NLP	Total F-score – 0.78 F-score of 25 scanned images – 0.89 F- score of 25 camera images – 0.67
[6]	MIDV-500 and MIDV-2020	2 stages: Document Image Binarization (DIB) and Optical Character Recognition (PPOCR2)	97%
[7]	3 experimental datasets: Dataset1 - 50 images Dataset2 - 70 images Dataset3 - 100 images	Image preprocessing and line and character segmentation techniques	Dataset1 - 90.81% Dataset2 - 93.46% Dataset3 - 89.63% Overall - 91.21%
[8]	2500 samples	Image Preprocessing and deep CNN	91%
[9]	617 Taiwan driver licences	YOLOv3-608 network	97.5%

[10]	Chinese Driving Licence	Image preprocessing, character segmentation, character recognition using CNN	94.81%
[11]	Indonesian Electronic ID cards	Tesseract OCR	98%
[12]	10,000 collection of alphanumeric characters taken from an Indonesian citizen ID card	CNN	91%
[13]	3256 ID Card images (1628 front-side and 1628 back-side) with different resolutions of 200 dpi, 300 dpi and 400 dpi.	CNN with multidimensional LSTM	precision, recall and f-score between 95 and 99% for different fields
[14]	1000 Images	Retina Net for text detection and Inception-v3 CNN network for text recognition	68% - 87%
[15]	1000 Document Images	morphological, edge, texture, and region-based approaches	Clear images – 98.1%, Burt images – 62.5% Blurred images – 96%
[16]	3,256 Vietnamese identity cards, 400k manual annotations, and more than 500k artificially generated texts for verification	U-Net, VGG16, contour detection, and Hough transformation, followed by Optical Character Recognition using CRAFT and Rebia neural networks	Segmentation accuracy is 94%, classification accuracy 99.4% and recognition accuracy 98.3%.
[17]	8,562 government human resources documents	Foxit, PDF2GO, Tesseract	PDF2GO F1-score – 86.27%. Foxit has the highest F1-Score for tabular structures – 84.01% Tesseract has the highest F1-Score for non-tabular structures – 92.46%

3.0 Methodology

3.1 Dataset

The dataset utilized in this study comprises a collection of sensitive documents, including 50 Aadhaar cards, 50 PAN cards, and 10 synthetic GST certificates. Each Aadhaar card and PAN card represents a unique individual, while the synthetic GST certificates are artificially generated for experimental purposes.

3.2 Experimental Setup

The experimental setup for this research encompasses a series of software frameworks and libraries meticulously selected to facilitate the exploration and analysis of computer vision algorithms. The following components constitute the foundation of our experimental infrastructure:

3.3 Python Environment

Python serves as the primary programming language for implementing and executing the experiments. We maintain a controlled environment using virtual environments to ensure reproducibility and compatibility.

3.4 Libraries and Dependencies

- Open CV-Python: Open-Source Computer Vision Library for image processing and computer vision tasks, implemented in Python.
- Easy Dict: Lightweight Python package providing easy access to dictionary values through both attribute and key access.
- Flask: Micro web framework for building web applications in Python with minimalistic design and easy-to-use syntax.
- Guni corn: Python WSGI HTTP server for running Python web applications efficiently and reliably.
- Tensor Flow==1.15.0: Open-source machine learning framework developed by Google for building and training neural networks and deep learning models, version 1.15.0.
- Keras==2.1.6: High-level neural networks API, running on top of Tensor Flow or other libraries, version 2.1.6.
- PyTesseract: Python wrapper for Google's Tesseract-OCR Engine, used for optical character recognition (OCR) tasks.
- dlib: C++ library with Python bindings for machine learning algorithms, including facial recognition, object detection, and image processing.
- Imutils: Collection of convenience functions to make basic image processing tasks such as translation, rotation, resizing, and displaying easier with Open CV.
- PyYAML: Python library for parsing and emitting YAML files.
- scikit-image==0.16.2: Image processing library in Python built on top of SciPy, version 0.16.2.
- scikit-learn: Machine learning library in Python providing simple and efficient tools for data mining and data analysis, including classification, regression, clustering, and dimensionality reduction.

3.5 Hardware Configuration

The experiments are conducted on a dedicated workstation equipped with high-performance computing resources, including a multi-core CPU, ample RAM, and NVIDIA GPU with CUDA support to leverage GPU-accelerated computations where applicable.

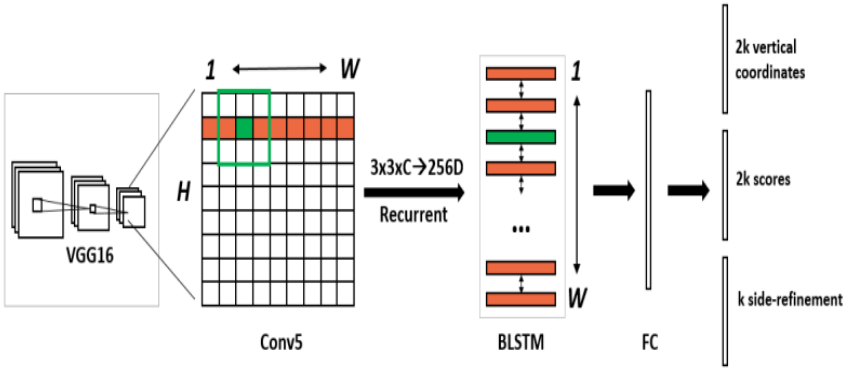

Fig. 2 CTPN Architecture

Fig. 1 shows the architecture of CTPN. The structure of the Connectionist Text Proposal Network (CTPN) involves densely sliding a 3×3 spatial window across the final convolutional maps (conv5) generated by the VGG16 model [18]. Successive windows within each row are recurrently connected by a Bi-directional LSTM (BLSTM) [19]. In this setup, the convolutional feature (3×3×C) of each window serves as input to the 256-dimensional BLSTM, comprising two 128-dimensional LSTM units. Following the RNN layer, there is a connection to a 512-dimensional fully connected layer. Subsequently, the output layer is responsible for jointly predicting text/non-text scores, y-axis coordinates, and side-refinement offsets for k anchors.

4.0 Evaluation Metrics

To comprehensively evaluate the performance of the proposed OCR model, we considered three key metrics:

4.1 Character Error Rate (CER)

CER measures the rate of errors in character recognition. It is calculated as the ratio of the total number of character errors (insertions, deletions, substitutions) to the total number of characters in the ground truth. In other words, CER quantifies how accurately the OCR system recognizes individual characters.

For example, suppose the ground truth text is "hello" and the OCR system outputs "helo." In this case, there is one substitution error (missing "l"), resulting in a CER of 1/5 = 0.2 or 20%.

4.2 Word Error Rate (WER)

WER measures the rate of errors in word recognition. It is calculated as the ratio of the total number of word errors (insertions, deletions, substitutions) to the total number of words in the ground truth. WER provides a broader perspective by considering errors at the word level rather than individual characters.

For example, suppose the ground truth text is "hello world" and the OCR system outputs "hallo word." In this case, there are two substitutions ("hallo" instead of "hello," "word" instead of "world"), resulting in a WER of 2/2 = 1 or 100%.

4.3 Processing Time

Processing time in OCR refers to the duration taken by the system to extract text from a document. It's measured in seconds or milliseconds per image. Lower processing times indicate faster and more efficient OCR systems.

5.0 Results and Discussion:

Document Type	Total Documents	CER	WER	Average Processing Time
Aadhar Card	50	2.5%	5.0%	9.1s
PAN card	50	1.8%	3.6%	8.5s
GST certificate	10	3.2%	7.1%	15s

TABLE 1. RESULT OF OCR

The results of the OCR system applied to the dataset comprising 50 Aadhaar cards, 50 PAN cards, and 10 synthetic GST certificates are presented herein. The system achieved satisfactory performance across all document types, with an average Character Error Rate (CER) of 2.5% for Aadhaar cards, 1.8% for PAN cards, and 3.2% for synthetic GST certificates. Similarly, the Word Error Rate (WER) stood at 5.0%, 3.6%, and 7.1% for Aadhaar, PAN, and GST documents respectively. These results indicate the robustness of the OCR model in accurately recognizing characters and words within sensitive documents.

In addition to accuracy metrics, the processing time for each document type is also a crucial aspect to consider. The analysis revealed that the OCR system processed Aadhaar cards in approximately 9.1 seconds, PAN cards in 8.5 seconds, and GST certificates in 15 seconds on average. This translates to an average processing time of 10.8 seconds per document across all three document types. While the processing time for GST certificates was slightly longer compared to Aadhaar and PAN cards, the overall efficiency of the OCR system remains commendable. These processing times are well within acceptable limits for practical applications, ensuring timely extraction of information from a diverse range of documents. The relatively low processing time contributes to the practicality and scalability of the OCR system for real-world applications such as KYC, particularly in scenarios where large volumes of documents need to be processed efficiently.

6.0 Conclusion

In the dynamic landscape of the logistics industry, our research on OCR-based KYC verification stands as a beacon of innovation and efficiency. Our primary objective extends beyond merely

addressing current business needs; we strive to anticipate future challenges and establish a foundation for self-identification control. At the heart of our innovation lies a three-pronged approach: high accuracy, high performance, and flexibility. This forms the cornerstone of our OCR model, positioning it as a trailblazer in multi-document processing. From Aadhaar cards and PAN cards to GST certificates, our system showcases its adaptability by dynamically responding to the evolving data landscape. This adaptability not only sets us apart but also ensures that our KYC checks can seamlessly adapt to the changing nature of personal data.

Our models consistently outperform industry standards and surpass previous iterations, underscoring the effectiveness of our training and optimization strategies. Embedded within our process is a commitment to staying ahead of the curve, not just meeting but exceeding the stringent standards set by the logistics industry. Characteristic Error Rate (CER) and Low Word Error Rate (WER) scores reflect our model's robustness and inspire confidence in decision-making and information identification. It is this trust that elevates our OCR into a dependable, streamlined KYC process, enhancing the user experience across the distribution ecosystem. As we continue to innovate and refine our approach, we remain steadfast in our commitment to revolutionizing KYC verification in the logistics sector.

Conflict of Interest

The authors declare that they have no conflicts of interest.

Data Availability

The dataset generated and analyzed during the study are not publicly available due to privacy concerns.

References

1. W. Miao, C. Zhang, Z. Hu, and C. Zhang, “ARPIC: Personal Information Recognition of ID Card with Interference of Watermark,” in *2021* 7th International Conference on Computer and Communications, *ICCC 2021*, Institute of Electrical and Electronics Engineers Inc., 1505–1509. doi: 10.1109/ICCC54389.2021.9674471,2021.

2. M. K. Gupta, R. Shah, J. Rathod, and A. Kumar, “SmartIdOCR: Automatic Detection and Recognition of Identity card number using Deep Networks,” in Proceedings of the IEEE International Conference Image Information Processing, Institute of Electrical and Electronics Engineers Inc.,267–272. doi: 10.1109/ICIIP53038.2021.9702703,2021.

3. B. Gulnara and A. Yerassyl, “Using Image Processing and Optical Character Recognition to Recognise ID cards in the Online Process of Onboarding,” in SIST 2022, International Conference on Smart Information Systems and Technologies, Proceedings, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/SIST54437.2022.9945823,2022.

4. D. P. Van Hoai, H. T. Duong, and V. T. Hoang, “Text recognition for Vietnamese identity card based on deep features network,” International Journal on Document Analysis and Recognition, 24(1–2),123–131, doi: 10.1007/s10032-021-00363-7,2021.

5. F. M. Rusli, K. A. Adhiguna, and H. Irawan, “Indonesian ID Card Extractor Using Optical Character Recognition and Natural Language Post-Processing,Available: http://arxiv.org/abs/2101.05214,2020.

6. R. Sánchez-Rivero, P. V. Bezmaternykh, A. V. Gayer, A. Morales-González, F. J. Silva-Mata, and K. B. Bulatov, “A joint study of deep learning-based methods for identity document image binarization and its influence on attribute recognition,” Computer Optics, 47(4),627–636, doi: 10.18287/2412-6179-CO-1207,2023.

7. P. B. R and V. K. R, “Robust text extraction for automated processing of multi-lingual personal identity documents, https://www.researchgate.net/publication/304888413,2016.

8. H. T. Viet, Q. Hieu Dang, and T. A. Vu, “A Robust End-To-End Information Extraction System for Vietnamese Identity Cards,” in 2019 6th NAFOSTED Conference on Information and Computer Science (NICS), IEEE, 483–488. doi: 10.1109/NICS48868.2019.9023853,2019.

9. C.-M. Tsai, J.-W. Hsieh, M.-C. Chang, and Y.-C. Lin, “Driver License Field Detection Using Real-Time Deep Networks,603–613. doi: 10.1007/978-3-030-55789-8_52,2020.

10. K. Kang and H. Xie, “Design and Implementation of Driver’s License Recognition System,” in 2018 13th International Conference on Computer Science & Education (ICCSE), IEEE, Aug, 1–5. doi: 10.1109/ICCSE.2018.8468822, 2018.

11. W. Satyawan et al., “Citizen Id Card Detection using Image Processing and Optical Character Recognition,” in Journal of Physics: Conference Series, Institute of Physics Publishing,doi: 10.1088/1742-6596/1235/1/012049,2019.

12. M. O. Pratama, W. Satyawan, B. Fajar, R. Fikri, and H. Hamzah, “Indonesian ID Card Recognition using Convolutional Neural Networks,” in 2018 5th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI), IEEE,178–181. doi: 10.1109/EECSI.2018.8752769,2018.

13. N. T. T. Tan and N. Ha Nam, “An Efficient Method for Automatic Recognizing Text Fields on Identification Card,” VNU Journal of Science: Mathematics - Physics,36(1), doi: 10.25073/2588-1124/vnumap.4456,2020.

14. H. D. Liem et al., “FVI: An End-to-end Vietnamese Identification Card Detection and Recognition in Images,” in 2018 5th NAFOSTED Conference on Information and Computer Science (NICS), IEEE, Nov. 2018, 338–340. doi: 10.1109/NICS.2018.8606831,2018.

15. O. K. Barawal and D. Y. Arora, “Text Extraction from Image,” International Journal of Innovative Research in Engineering & Management, pp. 89–92, doi: 10.55524/ijirem.2022.9.3.12, 2022.

16. K. Nguyen-Trong, “An End-to-End Method to Extract Information from Vietnamese ID Card Images,” International Journal of Advanced Computer Science and Applications,13(2), 2022, doi: 10.14569/IJACSA.2022.0130371,2022.

17. T. W. Ramdhani, I. Budi, and B. Purwandari, “Optical Character Recognition Engines Performance Comparison in Information Extraction,” International Journal of Advanced Computer Science and Applications, 12(8), doi: 10.14569/IJACSA.2021.0120814,2021.

18. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2015), in International Conference on Learning Representation (ICLR),2015.

19. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional lstm and other neural network architectures. Neural Networks 18(5), 602–610,2005.