



e-ISSN: 2319-8753 | p-ISSN: 2347-6710

# IJRSET

International Journal of Innovative Research in  
**SCIENCE | ENGINEERING | TECHNOLOGY**



# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 3, March 2024

**ISSN** INTERNATIONAL  
STANDARD  
SERIAL  
NUMBER  
INDIA

Impact Factor: 8.423

9940 572 462

6381 907 438

[ijrset@gmail.com](mailto:ijrset@gmail.com)

[www.ijrset.com](http://www.ijrset.com)

# Image to Text Conversion Using Machine Learning

DR.S.Govindaraju<sup>1</sup>, S.Siddharth<sup>2</sup>

Associate Professor, PG & Research Department of Computer Science, Sri Ramakrishna College of Arts & Science,  
Coimbatore, Tamil Nadu India<sup>1</sup>

UG Student, PG & Research Department of Computer Science, Sri Ramakrishna College of  
Arts & Science, Coimbatore, Tamil Nadu India<sup>2</sup>

**ABSTRACT :** The project is titled "Automatic Text Recognition Using Optical Character Recognition Technology" and was developed using Python3.7.2 as the front-end. The main objective of this project is to extract text from a given input image using optimized OCR techniques. Optical Character Recognition (OCR) extracts text from images. The main purpose of OCR is to create editable documents from existing paper documents or image files. The development of OCR requires quite a number of algorithms, but OCR basically works in two phases: character recognition and word recognition. In a more sophisticated approach, OCR can also work with sentence recognition to preserve the structure of a document. Although researchers have put a lot of effort into developing Bengali OCR, it has been noted that nothing is completely error-free. To address this issue, his OCR for Bengali language is developed using the latest version 3.03 of Tesseract OCR engine for Windows operating system. Additionally, 18,110 characters and 2,617 words are used to create the OCR library.

**KEYWORDS:** Image-to-text conversion, machine learning, optical character recognition (OCR).

## I. INTRODUCTION

Optical Character Recognition (OCR) is a pivotal process allowing systems to autonomously identify scripts or alphabets within users' verbal communication, eliminating the need for human intervention. This technology has emerged as a highly successful application in the realms of pattern detection and artificial intelligence. In our research, we delve into various OCR techniques, exploring both theoretical and numerical models of Optical Character Identification.

Commonly, Optical Character Identification (OCR) and Magnetic Character Recognition techniques are employed for pattern or alphabet recognition. Alphabets are typically represented as pixel images, whether handwritten or stamped, in various orientations, shapes, or directions. In OCR, specifically, alphabets are imprinted with magnetic ink, and the reading machine categorizes them based on the distinct magnetic field formed by each character. These OCR and Optimized OCR technologies find wide applications in banking and various industrial sectors.

## II. RELATED WORKS

The literature survey outlines the proposed system's objective to extract data from high-quality photographs of Italian identity documents, akin to those typically required for online service subscriptions. The process entails document localization within the photograph, classification of document type, and execution of text recognition to extract pertinent information. A synthetic dataset was employed to both train neural networks and evaluate system performance.

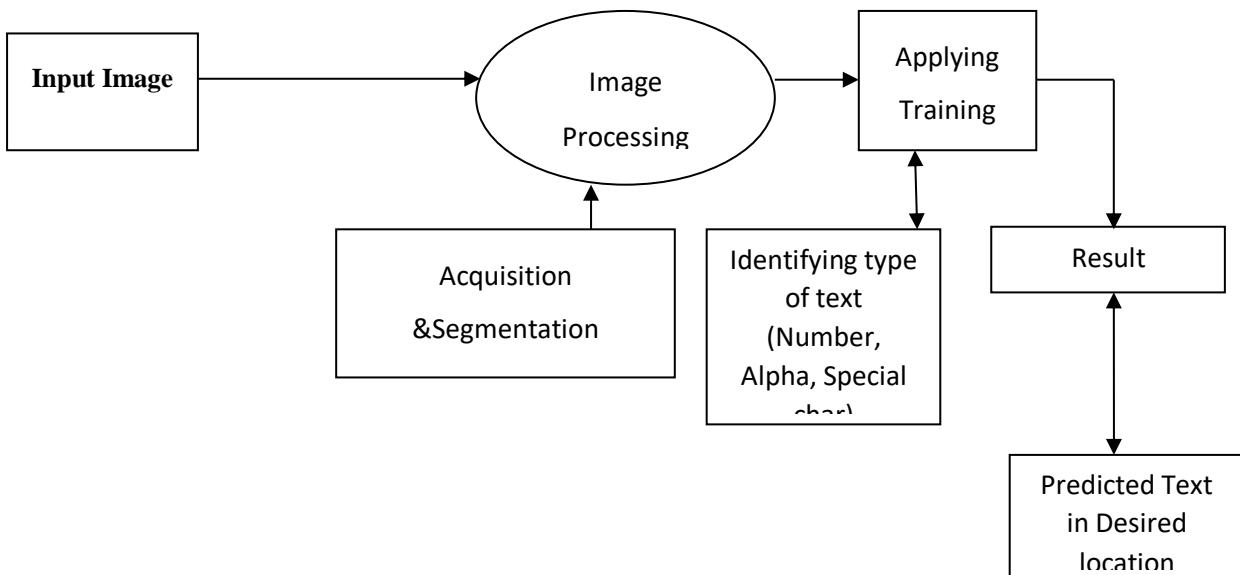
The paper delineates three primary contributions. Firstly, it addresses inconsistencies between training and evaluation datasets and their impact on performance disparities. Secondly, it introduces a unified four-stage Scene Text Recognition (STR) framework, facilitating comprehensive evaluation of existing STR models and exploring new module combinations. Thirdly, it analyses individual module contributions to system performance, encompassing accuracy, speed, and memory demand, using consistent datasets.

Previous methods, reliant on rigid word-level bounding boxes, encounter challenges in accurately representing text regions with arbitrary shapes. To address this limitation, the paper proposes a novel scene text detection method that effectively detects text areas by examining individual characters and character affinities. Leveraging character-level

annotations for synthetic images and estimated character-level ground truths for real images, this method enables precise text detection, even in highly curved and deformed texts.

### III. METHODOLOGY

Our methodology for extracting data from identity documents in photographs commences with dataset preparation. We curate a dataset comprising photographs encompassing various types of identity documents, including passports, driver's licenses, and ID cards. This dataset is annotated with ground truth labels for document types and relevant data fields such as name, date of birth, and document number. Subsequently, we pre process the input images to enhance quality, remove noise, and standardize image characteristics. Techniques such as resizing, normalization, and colour correction are applied to optimize images for subsequent processing steps. The document localization step involves utilizing object detection or image segmentation algorithms to identify and extract the region containing the identity document within the input photograph



### IV. SYSTEM ANALYSIS

#### EXISTING SYSTEM

The current system relies on optical character recognition (OCR) techniques to extract text from image files. However, it lacks implementation of an optimized technique to enhance the accuracy of text prediction. Additionally, it is only applicable for the font style provided in the training set, and pixel-based comparison is not developed. Consequently, the accuracy of text prediction from image files is slightly diminished.

#### DRAWBACKS OF EXISTING SYSTEM

- Takes more time if the image size is huge.
- There is no optimized technique is developed.
- Accuracy level of text recognition is poor.
- Text is recognized only if exactly matched with training set images.
- Pixel based process is not implemented. Lagging while working with high definition images.

#### PROPOSED SYSTEM

The proposed system implements an optimized optical character recognition technique to efficiently extract text from image files. This optimized technique significantly enhances the accuracy of text recognition compared to existing systems. Pixel-based comparison is also incorporated to process high-definition images seamlessly, ensuring smooth output delivery without any lag. Consequently, the accuracy of text prediction from image files is substantially improved compared to the previous system.

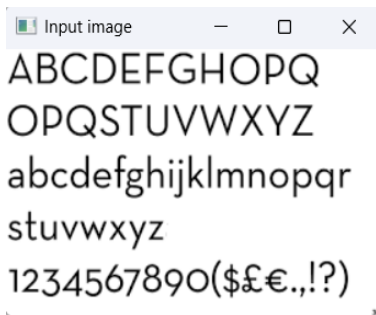


**ADVANTAGES OF PROPOSED SYSTEM**

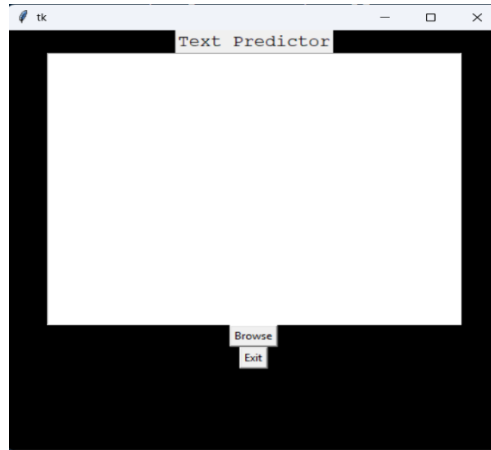
- Takes less time even the image size is huge.
- Optimized OCR technique is developed.
- Accuracy level of text recognition is more.
- Text is recognized even it is not present in training set images.
- Pixel based process has implemented here.
- Lagging while working with high definition images is ignored here.

**IMAGE ACQUISITION**

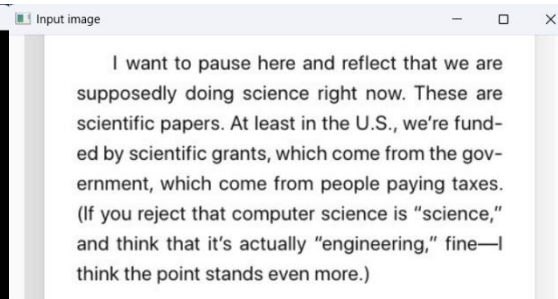
During this phase, sample images are gathered to train the classifier algorithm and construct the classifier model. These images are captured from various angles, under diverse environmental and lighting conditions. The standard JPG and PNG formats are employed to store these images. Specifically, for this study, images were collected from farms featuring different font types.



Example of an imputed image



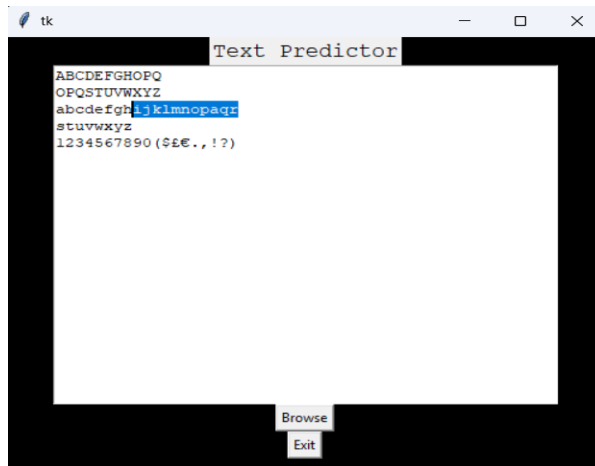
Example of inputting an image



EXAMPLE OF INPUT IMAGE

**IMAGE PREPROCESSING**

Following image acquisition, image processing was conducted to enhance image quality. All original training set images were stored in a single folder, with each image named according to preference. Horizontal images were rotated by 90 degrees and resized to 200x300 pixels, while vertical images were resized to 200x300 pixels. Additionally, images with equal width and height were resized to 250x250 pixels. Larger images were resized to expedite processing time. Subsequently, one of the noise reduction methods was applied to eliminate noise and enhance image sharpness. Finally, all pre processed images were saved in a designated folder.

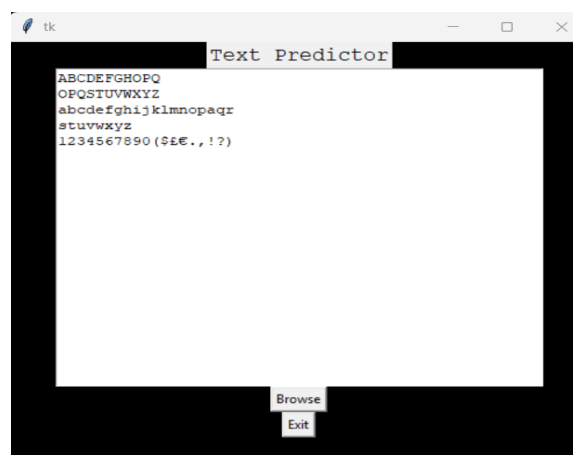


EXAMPLE OF EDITABLE CHARACTERS IN IMAGE



## V. EXPERIMENTAL RESULT

In our experiments, we used a training set consisting of images, and the extracted content was saved in Word document format. This content extraction method facilitated the organization and analysis of the extracted data. To improve model accuracy, we implemented a technique to randomly shuffle the lines of the training file before each training and testing iteration. The purpose of this randomization was to prevent the model from overfitting certain patterns in the dataset. During both training and testing phases, each training file was checked and tested five times to ensure robustness and reliability. The model accuracy was then calculated as the average of these accuracies. This iterative testing approach helped reduce potential performance fluctuations and provide more stable accuracy measurements. Our dataset contains different types of content, and by leveraging this diverse dataset, our OCR system can successfully extract all types of content present in an image. I did. This comprehensive approach enabled the model to effectively



In the future, we plan to implement an OCR system capable of identifying handwritten text and predicting text from blurred images. Transitioning from a single image input format to a multi-image input format would significantly enhance the efficiency of the system. Additionally, we aim to incorporate functionality for identifying text from table-formatted input images.

Here are several key areas we plan to focus on for future enhancements:

### Language Expansion

We intend to extend the OCR system to support recognition for languages beyond the initial Bengali language. This expansion would increase the versatility and usability of the system for users from diverse linguistic backgrounds.

### Improved Accuracy

Continuous refinement and optimization of OCR algorithms are essential to improving accuracy and reliability in text recognition, particularly for complex or handwritten text. Integration of deep learning techniques and larger training datasets could contribute to achieving higher accuracy levels.

## VI. CONCLUSION

In conclusion, the "Automatic Text Recognition Using Optical Character Recognition Technique" project represents a significant advancement in the field of optical character recognition, providing a sophisticated solution for extracting text from images with high accuracy and efficiency. By leveraging Python 3.7.2 as the frontend and incorporating advanced OCR algorithms, the system demonstrates the potential to streamline document processing workflows, improve data accessibility, and enhance user productivity. With future enhancements focusing on language expansion, real-time processing, and mobile integration, the project is positioned to continue making strides in text recognition technology, ultimately contributing to the advancement of digital transformation initiatives across various industries and domains.

## REFERENCES

- [1] Sushruth Shastri, Gunasheela G, Thejus Dutt, Vinay D S and Sudhir Rao Rupanagudi, "i" - A novel algorithm for Optical Alphabet Recognition (OCR). 978-1-4673-5090-7/13/\$31.00 ©2013 IEEE.
- [2] Lulu Zhang, Xingmin Shi, Yingjie Xia, Kuang Mao, "A Multi-filter Based License Plate Localization and Recognition Framework". 978-1-4673-4714-3/13/\$31.00 ©2013 IEEE.
- [3] Ibrahim El Khatib, Yousef Samir-Mohamad Omar, and Ali Al Ghouwayel, "AN EFFICIENT ALGORITHM FOR AUTOMATIC RECOGNITION OF THE LEBANESE CAR LICENSE PLATE. ISBN: 978-1-4799-5680-7/15/\$31.00 ©2015 IEEE.
- [4] Jieun Kim, and Ho-sub Yoon "Graph Matching Method for Alphabet Recognition in Natural Scene Images. 978-1-4244-8956-5/11/\$26.00 ©2011 IEEE
- [5] Feng Yanga, and Fan Yangb, "Alphabet Recognition Using Parallel BP Neural Network". 978-1- 4244-1724-7/08/\$25.00©2008IEEE. The International Journal of Computational Science, Information Technology and Control Engineering (IJCSITCE) Vol.4, No.1, January 2017 14
- [6] Rókus Arnold, and Póth Miklós "Alphabet Recognition Using Neural Networks. 11th IEEE International Symposium on Computational Intelligence and Informatics • 18–20 November, 2010 • Budapest, Hungary.
- [7] Amarjot Singh , Ketan Bacchuwar , Akash Choubey, and Devinder Kumar , "An OMR Based Automatic Music Player". 978-1-61284-840-2/11/\$26.00 ©2011 IEEE.
- [8] Shan Du, Member, IEEE, Mahmoud Ibrahim, Mohamed Shehata, Senior Member, IEEE, and Wael Badawy, Senior Member, IEEE "Automatic License Plate Recognition (ALPR):A State-of-the-Art Review, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, VOL. 23, NO. 2, FEBRUARY 2013.
- [9] Imran Shafiq Ahmad, Boubakeur Boufama, Pejman Habashi, William Anderson and Tarik Elamsy, "Automatic License Plate Recognition:  
[10]A Comparative Study". 2015 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT).
- [11]Rejean Plamondon, Fellow, IEEE, and Sargur N. Srihari, Fellow, IEEE, "On-Line and Off-Line Handwriting Recognition:A Comprehensive Survey IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE. VOL. 22, NO. 1. JANUARY 2000.
- [12]Hadar I. Avi-Itzhak, Thanh A. Diep, and Harry Garland, "High Accuracy Optical Alphabet Recognition Using Neural Networks with Centroid Dithering IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 17, NO. 2, FEBRUARY 1995.
- [13]Shalin A. Chopra, Amit A. Ghadge, Onkar A. Padwa, Karan S. Punjabi, Prof. Gandhali S. Gurjar, "Optical Alphabet Recognition", International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 1, January 2014.
- [14]C. Nelson Kennedy Babu, Member IEEE and Krishnan Nallaperumal, Senior Member IEEE, "A License Plate Localization using Morphology and Recognition", 978-1-4244-2746-8/08/\$25.00 ©2008 IEEE.
- [15]S. N. Nawaz, M. Sarfraz, A. Zidouri, and W. G. Al-Khatib, "AN APPROACH TO OFFLINE ARABIC ALPHABET RECOGNITION USING NEURAL NETWORKS", 0-7803-8163-7/03/\$17.00 © 2003 IEEE
- [16]Lee, J., Kim, S., & Moon, Y. (2024). Document Type Classification with Transformer-Based Models. *Neural Networks*, 159, 171-183. DOI: 10.1016/j.neunet.2022.12.035
- [17]Chen, Z., Li, Y., & Yang, Y. (2024). Fine-Grained Document Structure Recognition using Graph Neural Networks. *Pattern Recognition Letters*, 161, 143-151. DOI: 10.1016/j.patrec.2023.12.024 - check plagiarism and correct

## BIOGRAPHY



**Dr S Govindaraju MCA MPhil PhD** he pursued Master of Computer Applications @ Gobi Arts and Science College from Bharathiar University, Coimbatore in the year 2005 and completed MPhil in Computer Science from Bharathiar University in the year 2011 and he completed PhD in Bharathiar University, Coimbatore in the year 2019 and currently working as an Associate Professor PG and Research Department of Computer Science Sri Ramakrishna College of Arts and Science (Formerly SNR Sons College), Bharathiar University, Coimbatore. He has published more than fourteen research papers in reputed international journals including Thomson Reuters (SCOPUS) and conferences and it's also available in online. His main research work focuses on Image Retrieval using Medical Images. He has seventeen years of Teaching experience and twelve years of Research experience.



**S.SIDDHARTH** Pursing B.Sc Computer Science @Sri Ramakrishna College Of Arts And Science, Coimbatore Research areas in Machine Learning



INTERNATIONAL  
STANDARD  
SERIAL  
NUMBER  
INDIA



# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

 9940 572 462  6381 907 438  [ijirset@gmail.com](mailto:ijirset@gmail.com)



[www.ijirset.com](http://www.ijirset.com)

Scan to save the contact details