



## OCR Image Reader Component

*Yash Korla<sup>\*1</sup>, Raj Motta<sup>\*2</sup>, Siddique Sayed<sup>\*3</sup>, Sanmit Gangurde<sup>\*4</sup>, Meena Talele<sup>\*5</sup>*

<sup>\*1,2,3,4</sup> Student, Department of Computer Engineering, Vivekanand Education Society's Polytechnic, Chembur, Mumbai, Maharashtra, India

<sup>\*5</sup> Lecturer, Department of Computer Engineering, Vivekanand Education Society's Polytechnic, Chembur, Mumbai, Maharashtra, India

### ABSTRACT

Optical Character Recognition (OCR) is the process of extracting text from a picture. the purpose of an OCR is to form editable documents from existing paper documents or image files. A significant number of algorithms are required to develop an OCR and it works in two phases character and word detection. just in case of a more sophisticated approach, an OCR also works on sentence detection to preserve a document's structure. it's been found that researchers put much effort into developing a Bengali OCR but none of them is error-free. By keeping this issue into consideration, the newest 3.03 version of Tesseract OCR engine for Windows package is employed to develop an OCR for the Bengali language. Moreover, 18,110 characters and 2,617 words are added to the OCR's library. during this research, 'Solaimanlipi' font and 200 input files were used to test the accuracy of OCR. it's found that for clean image files, the accuracy is measured in the percentage of correct characters and words.

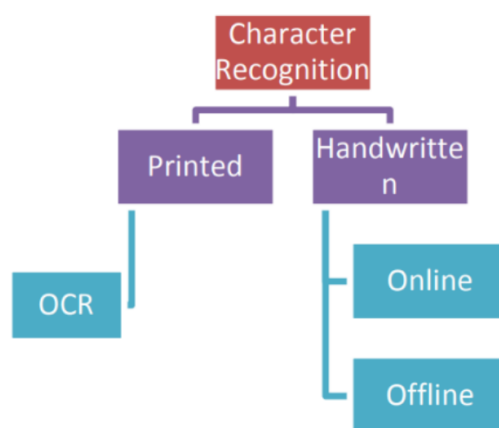
**Keywords:** OCR, Python, NumPy, PyTesseract, Jupyter, PostgreSQL

### 1. INTRODUCTION

Early OCR technologies were purely mechanical, dating back to 1914 when Emanuel Goldberg developed a machine that might read characters and convert them into standard telegraph code. In 1974, Ray Kurzweil and his colleagues started developing OCR systems, mainly that specialize in creating a "reading machine for the blind". The Tesseract library is because of which you are reading this tutorial now. These OCR engines still take a knowledgeable computer vision practitioner to control. Orientation and Script Detection (OSD) is the process of analyzing an input image for text meta-data, specifically the orientation and also the script/writing style. to get higher OCR accuracy, we may have to use OSD to see the text orientation, correct it, and then apply OCR. Any rules, heuristics, or assumptions an OCR system can make regarding a specific script or orthography will make the OCR engine more accurate. As hinted before, there are several meanings for OCR. It's most general meaning refers to extracting text from every possible image, be it a regular printed page from a book, or a random image with graffiti in it ("in the wild"). In between, you will find many other tasks, like reading license plates, no-robot captchas, street signs, etc. Although each of those options has its difficulties, clearly "in the wild" task is the hardest.

### 2. LITERATURE SURVEY

The problem of Character recognition dates back to the time when computers weren't even invented. The earliest OCR systems were mechanical devices and not computers, the major problem with these devices was their slow speed and extremely low accuracy. M. Sheppard's invention, a reading and robot GISMO will be considered the earliest work on modern OCR. GISMO's abilities are reading musical notations as well as reading words printed on a page one by one. The limitation of GISMO was the limited number of characters it could recognize. The machine also has the potential to copy a typewritten page. J. Rainbow devised a machine that could read uppercase typewritten English characters, one per minute, but the first OCR systems were criticized because of errors and slow recognition speed. Hence, the '60s and '70s did not focus on solving the problem of Character Recognition. The sole developments were done on government agencies and enormous corporations like banks, newspapers, airlines, etc. Due to the complexities and difficulties faced in character recognition, it was felt that there should be standardized OCR fonts that would make it easier to identify text via OCR. In 1970, ANSI and EMCA developed OCRA and OCRB, which helped us get acceptable character recognition rates. In the past thirty years, substantial research has been done on OCR. The research has led to the emergence of Document Image Analysis (DIA), Multi-Lingual, Handwritten, and Omni-Font OCRs. Even after all these efforts put into OCR, a machine's ability to read text was far below humans. Therefore, current OCR research is being done to improve the accuracy and speed of OCR for various style documents, printed/handwritten in unconstrained environments. OCR for complex languages like Sindhi and Urdu is still not available on any open-source platform or commercial software.



### 3.SYSTEM IMPLEMENTATION

#### A. EXPERIMENTAL SETUP

For this project, the programming language used is Python and is built from PyTesseract libraries using Visual Studio Code. For the database, we have used PostgreSQL to store container details and item information. Their short information is noted below:

- **NumPy:** NumPy, the abbreviation for Numerical Python, is a Python Package. It consists of multi-dimensional array objects and a collection of routines for processing those arrays. Using NumPy, Mathematical and Logical Operations can be performed on arrays.
- **Pandas:** Pandas is a Python Package that provides fast, flexible, and expressive data structures which are designed to make working with “relational” or “labeled” data easy. It aims to be the building block for doing practical and real-world data analysis in Python. Its broader goal is to become the most powerful and flexible open-source data analysis/manipulation tool available in any language.
- **Imutils:** Imutils are a series of functions that help us make basic image processing such as translation, resizing, rotating, skeletonization, and displaying Matplotlib images easier with OpenCV and both Python 2.7 and Python 3.
- **TensorFlow:** TensorFlow is an open-source platform for Machine Learning. It has a flexible and comprehensive ecosystem of tools, libraries, and community resources that helps researchers push the state of art in Machine Learning. TensorFlow helps the developers in building and deploying applications powered by Machine Learning.
- **PyTesseract:** PyTesseract, the abbreviation for Python Tesseract, is an Optical Character Recognition (OCR) tool for Python. It can identify and read the text embedded in an image. It is a wrapper for Google’s Tesseract-OCR Engine. It is useful as a standalone invocation script to Tesseract. It can read all types of images supported by the Pillow and Leptonica imaging libraries, which include jpeg, png, gif, BMP, tiff, etc. It can print the text instead of writing it in a file if we use PyTesseract as a script.
- **Matplotlib:** Matplotlib is a plotting library for Python and NumPy. It provides an Object-Oriented Application Programming Interface (API) for embedding plots into applications using general-purpose Graphical User Interfaces (GUI) toolkits like Tkinter, wxPython, Qt, or GTK. There is a procedural PyLab interface based on a state machine, designed to closely resemble that of MATLAB, though its use is discouraged. SciPy makes use of Matplotlib.
- **spaCy:** SpaCy is an Open-Source software library for Advanced Natural Language Processing (NLP), written in Python and Cython. The library is published under MIT License and its main developers are Matthew Honnibal and Ines Montani. They are the founders of the software company named Explosion.

#### B. PROJECT PROCEDURE AND FLOW

- We start by downloading and installing python on our laptops. We installed it with the following given configurations. We also installed Eclipse on our laptop.
- Then we started building the application by first building a demo application on the card reader to understand how an OCR works.
- Then we started building the main module in which we keep information about the container and store it in the database.
- Then we classify the information in the order of importance.
- Then we stored the container information in the database via PostgreSQL.

#### C. WHAT IS POSTGRE SQL?

Postgres SQL is the 4<sup>th</sup> most-popular database in the world. It is considered one of the most advanced databases in the world. After being in development for more than 20 years, PostGRE is managed by a highly organized, principled, and experienced open source community. Postgres SQL is Object-Oriented which is fully ACID compliant. It is highly extensible, enabling the community to add new features and capabilities as workload demands increase. PostGRE SQL provides a variety of built-in data types including JSON, XML, HSTORE (key-value pairs), Geo-Spatial (PostGIS), IPv6, flexible indexing, featuring composite indexes, GiST, SPGiST, GIN, full-text search, online index reorganization, background workers such as managed processes known as Mongress, which accepts MongoDB queries to interface with PostGRE’s data, a contrib module interface pgcryptp (for data encryption), pg\_trgm (find similar data), HSTORE (schema-less data), and extensive SQL Support. PostGRE SQL runs on all major Operating Systems such as Windows, Linux, macOS, etc. It offers the following programming languages: PL/pgSQL, PL/SQL, Java, Python, Ruby, C/C++, PHP, Perl, Tcl, Scheme. PostGRE also offers the following Library interfaces: OCI, libpq, JDBC, ODBC, .NET, Perl, Python, Ruby, C/C++, PHP, Lisp, Scheme, and Qt. Its databases provide enterprise-class database solutions and are used by various enterprises across many industries, including financial services, information technology, government, and media, and communications.

#### D. PYTESSERACT

PyTesseract, the abbreviation for Python Tesseract, is an Optical Character Recognition (OCR) tool for Python. It can identify and read the text

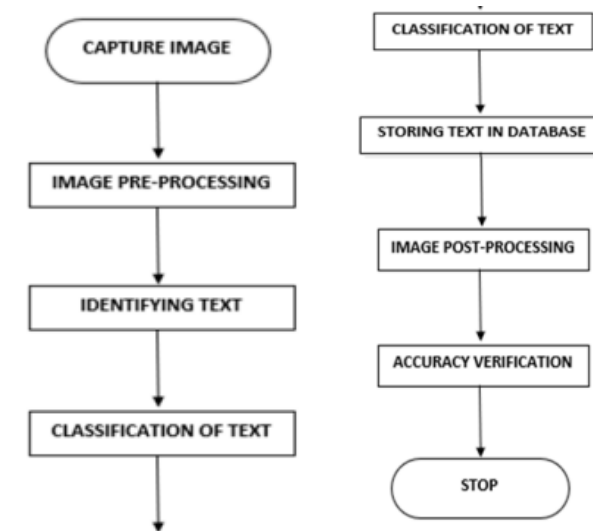
embedded in an image. It is a wrapper for Google's Tesseract-OCR Engine. It is useful as a standalone invocation script to Tesseract. It can read all types of images supported by the Pillow and Leptonica imaging libraries, which include jpeg, png, gif, BMP, tiff, etc. It can print the text instead of writing it in a file if we use PyTesseract as a script.

### E. BASIC FLOW OF THE APPLICATION

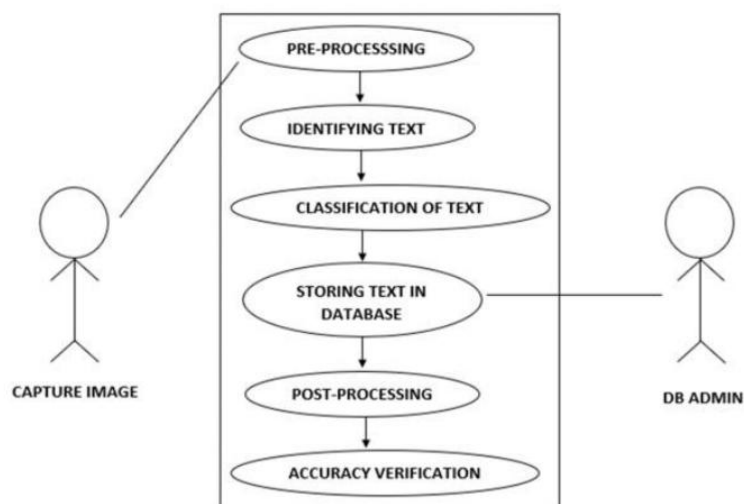
#### Algorithm:

- Step 1: Start
- Step 2: Capture Image and Send to API
- Step 3: Pre-Process the Image
- Step 4: Identify the text
- Step 5: Classify the text in its respective category
- Step 6: Store the text in the database
- Step 7: Post-Process the Image Step
- Step 8: Verify the Accuracy and Improve over time
- Step 9: Stop

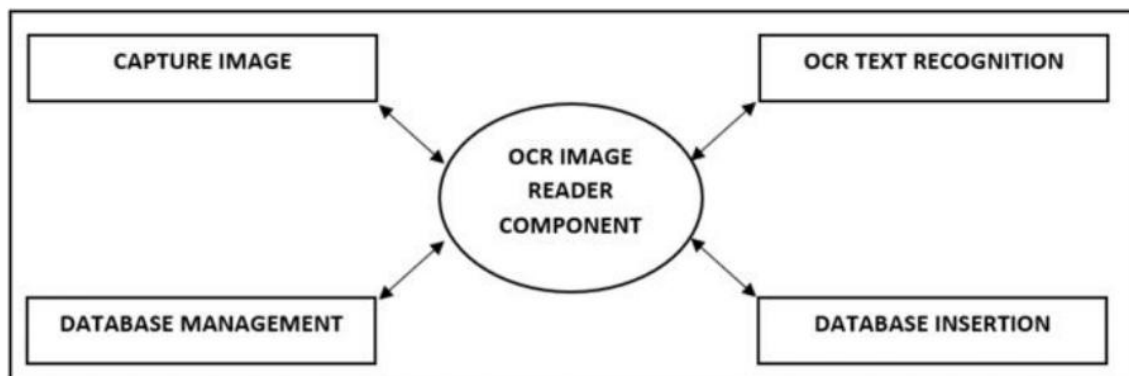
#### Flowchart:



#### Use Case Diagram:

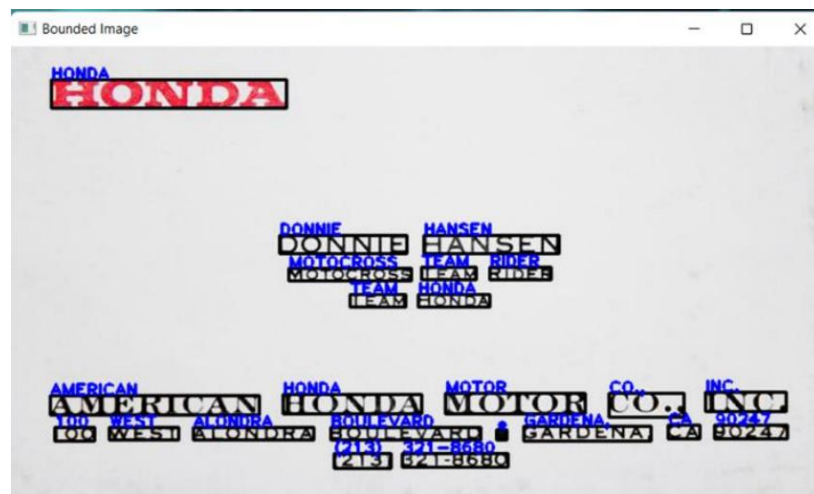


Data Flow Diagram:



## ACHIEVEMENTS TO DATE

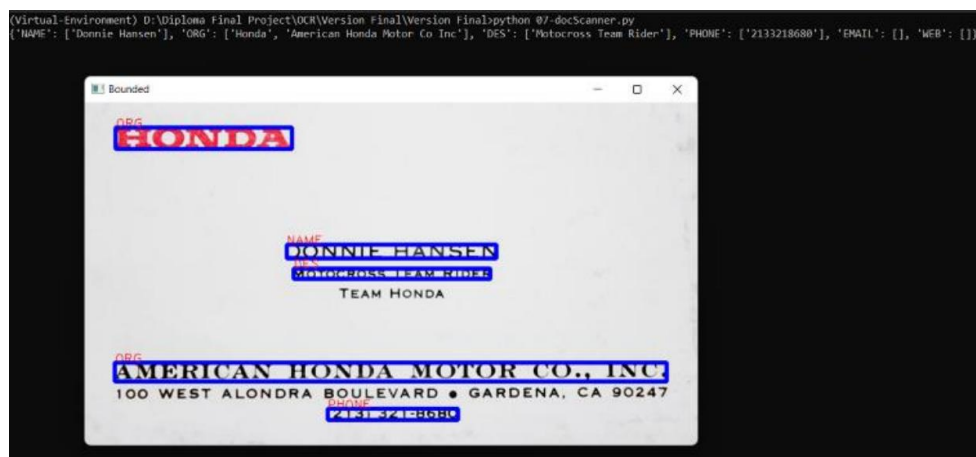
Plotting of Text on the Image:



Classification of Text and Verifying Accuracy:

| (Virtual-Environment) | D:\Diploma | Final     | Project\OCR\Version | Final\Version | Final\python | 01-practice-with-ocr.py |     |       |        |      |       |           |
|-----------------------|------------|-----------|---------------------|---------------|--------------|-------------------------|-----|-------|--------|------|-------|-----------|
| level                 | page_num   | block_num | par_num             | line_num      | word_num     | left                    | top | width | height | conf | text  |           |
| 0                     | 1          | 1         | 0                   | 0             | 0            | 0                       | 0   | 720   | 401    | -1   |       |           |
| 1                     | 2          | 1         | 1                   | 0             | 0            | 36                      | 29  | 207   | 25     | -1   |       |           |
| 2                     | 3          | 1         | 1                   | 1             | 0            | 36                      | 29  | 207   | 25     | -1   |       |           |
| 3                     | 4          | 1         | 1                   | 1             | 1            | 36                      | 29  | 207   | 25     | -1   |       |           |
| 4                     | 5          | 1         | 1                   | 1             | 1            | 36                      | 29  | 207   | 25     | 96   | HONDA |           |
| 5                     | 2          | 1         | 2                   | 0             | 0            | 0                       | 237 | 166   | 246    | 63   | -1    |           |
| 6                     | 3          | 1         | 2                   | 1             | 0            | 0                       | 237 | 166   | 246    | 63   | -1    |           |
| 7                     | 4          | 1         | 2                   | 1             | 1            | 0                       | 237 | 166   | 246    | 16   | -1    |           |
| 8                     | 5          | 1         | 2                   | 1             | 1            | 1                       | 237 | 166   | 112    | 16   | 96    | DONNIE    |
| 9                     | 5          | 1         | 2                   | 1             | 1            | 2                       | 364 | 166   | 119    | 16   | 95    | HANSEN    |
| 10                    | 4          | 1         | 2                   | 1             | 2            | 0                       | 245 | 194   | 231    | 12   | -1    |           |
| 11                    | 5          | 1         | 2                   | 1             | 2            | 1                       | 245 | 195   | 108    | 11   | 96    | MOTOCROSS |
| 12                    | 5          | 1         | 2                   | 1             | 2            | 2                       | 362 | 194   | 48     | 11   | 96    | TEAM      |
| 13                    | 5          | 1         | 2                   | 1             | 2            | 3                       | 422 | 194   | 54     | 11   | 96    | RIDER     |
| 14                    | 4          | 1         | 2                   | 1             | 3            | 0                       | 299 | 218   | 123    | 11   | -1    |           |
| 15                    | 5          | 1         | 2                   | 1             | 3            | 1                       | 299 | 218   | 49     | 11   | 94    | TEAM      |
| 16                    | 5          | 1         | 2                   | 1             | 3            | 2                       | 359 | 218   | 63     | 11   | 94    | HONDA     |
| 17                    | 2          | 1         | 3                   | 0             | 0            | 0                       | 35  | 304   | 649    | 67   | -1    |           |
| 18                    | 3          | 1         | 3                   | 1             | 0            | 0                       | 35  | 304   | 649    | 67   | -1    |           |
| 19                    | 4          | 1         | 3                   | 1             | 1            | 0                       | 35  | 304   | 648    | 22   | -1    |           |
| 20                    | 5          | 1         | 3                   | 1             | 1            | 1                       | 35  | 307   | 184    | 17   | 96    | AMERICAN  |
| 21                    | 5          | 1         | 3                   | 1             | 1            | 2                       | 240 | 306   | 122    | 17   | 96    | HONDA     |
| 22                    | 5          | 1         | 3                   | 1             | 1            | 3                       | 383 | 305   | 123    | 18   | 93    | MOTOR     |
| 23                    | 5          | 1         | 3                   | 1             | 1            | 4                       | 527 | 305   | 66     | 21   | 75    | CO.,      |
| 24                    | 5          | 1         | 3                   | 1             | 1            | 5                       | 612 | 304   | 71     | 17   | 94    | INC.      |
| 25                    | 4          | 1         | 3                   | 1             | 2            | 0                       | 39  | 333   | 645    | 14   | -1    |           |
| 26                    | 5          | 1         | 3                   | 1             | 2            | 1                       | 39  | 335   | 35     | 12   | 96    | 100       |
| 27                    | 5          | 1         | 3                   | 1             | 2            | 2                       | 87  | 335   | 61     | 11   | 93    | WEST      |
| 28                    | 5          | 1         | 3                   | 1             | 2            | 3                       | 161 | 335   | 106    | 11   | 90    | ALONDRA   |
| 29                    | 5          | 1         | 3                   | 1             | 2            | 4                       | 282 | 335   | 132    | 11   | 93    | BOULEVARD |
| 30                    | 5          | 1         | 3                   | 1             | 2            | 5                       | 428 | 337   | 9      | 9    | 92    | e         |
| 31                    | 5          | 1         | 3                   | 1             | 2            | 6                       | 452 | 334   | 112    | 13   | 92    | GARDENA,  |
| 32                    | 5          | 1         | 3                   | 1             | 2            | 7                       | 579 | 333   | 28     | 12   | 96    | CA        |
| 33                    | 5          | 1         | 3                   | 1             | 2            | 8                       | 620 | 333   | 64     | 12   | 96    | 90247     |
| 34                    | 4          | 1         | 3                   | 1             | 3            | 0                       | 285 | 358   | 153    | 13   | -1    |           |
| 35                    | 5          | 1         | 3                   | 1             | 3            | 1                       | 285 | 358   | 50     | 13   | 96    | (213)     |
| 36                    | 5          | 1         | 3                   | 1             | 3            | 2                       | 345 | 358   | 93     | 12   | 92    | 321-8680  |

### Plotting of Classification on the Image



### FUTURE RESEARCH

Optical Character Recognition has now started moving away from just seeing and matching. The new phase of OCR, driven by Deep Learning, first recognizes the text and then makes some meaning of it. The software that provides the most powerful character extraction and highest quality insights will be given the competitive edge. Since every business category has its document type, structure and consideration, there can be multiple companies that succeed based on vertical-specific competencies. Tradition OCR users must re-evaluate their current licenses and payment terms. Free services like Amazon's Textract and Google's Tesseract can also be tried to see the latest advancement in OCR and determine if those advancements align with their business goals. It will also be important to scope independent providers in the Robotic Process Automation (RPA) and Artificial Intelligence (AI) space that are making strides for the industry overall. In approximately 5 years, we expect that the past 30 to 50 years of static advancement will be unrecognizable.

### CONCLUSION

In this paper, our approach to Optical Character Recognition has been overviewed and presented. OCR is not a simple and single-phase process, it comprises various phases such as acquisition, pre-processing, feature extraction, classification, and post-processing. Each technology, programming language, external libraries used in our approach has been explained in detail in this paper. The practical applications of OCR can be: number plate recognition, text translation, etc. Despite the significant amount of research in OCR, recognition of characters for languages such as Arabic, Sindhi, and Urdu remains an open challenge. OCR for these languages has been planned as future work. Other important areas of research are multi-lingual character recognition systems. The employment of Optical Character Recognition Systems in practical applications remains an active area of research.

### REFERENCES

- [1]<https://www.forbes.com/sites/forbestechcouncil/2019/09/10/the-future-of-ocr-is-deeplearning/?sh=3b4743c96a04>
- [2]<https://www.mygreatlearning.com/blog/pythonnumpytutorial/#:~:text=NumPy%2C%20which%20stands%20for%20Numerical,stands%20for%20Numerical%20Python'>
- [3][https://pandas.pydata.org/docs/getting\\_started/overview.html#:~:text=pandas%20is%20a%20Python%20package,world%20data%20analysis%20in%20Python](https://pandas.pydata.org/docs/getting_started/overview.html#:~:text=pandas%20is%20a%20Python%20package,world%20data%20analysis%20in%20Python)
- [4]<https://becominghuman.ai/face-detection-with-opencv-and-deep-learning-from-video-part-2-592e2dee648#:~:text=Imutils%20are%20a%20series%20of,Python%202.7%20and%20Python%203>
- [5]<https://www.tensorflow.org/learn>
- [6]<https://spacy.io/>
- [7]<https://matplotlib.org/>