International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 05 | May - 2022 Impact Factor: 7.185 ISSN: 2582-3930

A STUDY ON OPTICAL CHARACTER RECOGNITION

Anirudh Dabral, Anika Bisht, Abhinav Sisodia, Ariba Khan and Devansh Gupta

Dept of Computer Science and Engineering
B. Tech, Babu Banarasi Das National Institute of Technology and Management & College

---

Abstract - The paperwork utilized in keeping diverse kinds of files in our daily life is time consuming and also it is far tough to preserve and recall the involved documents. This problem can be solved by the Optical Character Recognition (OCR) technique which is used for extracting texts and characters from an image which has a high accuracy rate for good quality images. In this project we will use machine learning techniques to train a model to recognize characters from images and then we compare it with results from Tesseract OCR Engine (which is an optical character recognition engine for various operating systems) to check accuracy of the model. The outcome of this project will be an application which implements machine learning techniques 36 model of OCR and a web version of OCR running on Tesseract API. It is observed that for the high resolution images the accuracy is above 84%.

***Key Words*:** Optical Character Recognition, Tesseract OCR Engine, text recognition.

1.INTRODUCTION *(Size 11, Times New roman)*

Supervised machine learning is a type of machine learning which is uses labelled datasets to train the algorithms so that it can predict the outcomes and classify the input data into the classes more accurately. As input data (which is labelled) is fed into the model, it adjusts its weights until the model has been fitted appropriately, which occurs as part of the cross-validation process. OCR stands for optical character recognition. It is used to distinguish printed text characters inside digital images of physical documents, such as a scanned paper document. It is the process taking the text as input from the document and examining it. Then this text is translated the characters of the text into the code for further processing. OCR is sometimes also referred as Optical Character Reader. For training the machine learning model, MNIST dataset is used for recognizing the text from images. It is a very big collection of handwritten digits. Various image processing systems are trained using this dataset. MNIST is Modified National Institute of Standards and Technology database. It was developed to train machine learning to recognize handwritten digits. First a physical form of document is processed into a digital form using a scanner. After this, the OCR converts the document into a black and white only form. The scanned-in image or bitmap is analysed for light and dark areas, where the dark areas are identified as characters that need to be recognized and light areas are identified as background. The dark areas are used to find alphabets or numeric digits. There are many techniques for recognition of characters using OCR, mostly it is selecting one character and run the algorithm on it.

Characters are generally identified using one of the two algorithms:

1. Pattern recognition: Pattern of dark and white areas of various formats are used by the OCR algorithms to compare the input problem with the dataset to recognize the characters and give a meaningful output.

2. Feature detection: Features like crossed lines, angled lines and curves of a character are used to differentiate characters from one another and recognize the characters. Like the digit '8' can be stored as two circles stacked on top of each other.

The most widely used open-source text recognition (OCR) engine is called Tesseract, it is made accessible through the Apache 2.0 license. It provides an API for programs; this API is used to get characters from the input images. It can be used for a vast range of languages such as Java, Python, JavaScript, etc.

The project aims to apply machine learning techniques in building a model which can recognize hand-written characters and to provide a quality tool accessible to all letting them discover the massive utility of OCR in our lives through an application.

2. Related Work

- (Karez Hamad et al., 2016) had proposed a way of organizing the various methods, algorithms and techniques of Optical Character Recognition. They have discussed challenges in OCR and about important phases involved in this process.
- (Chirag Patel et al., 2012) discussed about the Tesseract OCR engine and its applications in extracting vehicle registration numbers from a vehicle number plates and how accurate are the results and compare them with the results obtained by using Transym OCR engine.
- (Sebastian Raschka et al., 2020) has talked about various advances in the field of Machine Learning and the challenges and current state of solutions available for it. They have also discussed how the Quantum computing approaches in the future can help in solving the current challenges and make the available solutions more efficient then ever.

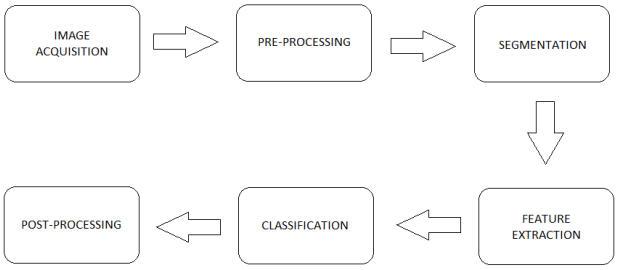
3. Proposed Work

- This task objectives at solving these issues confronted in the actual world with the aid of implementation of an OCR device to optically convert a digitally captured picture into device readable textual content shape to be able to assist in optimizing the precise workflows and enable toughness of user information through allowing digital storage and compression strategies for physical entities.

- Machine learning model for character recognition is built using MNIST dataset for recognizing digits between [0 to 9], whereas EMNIST dataset is being used for character recognition for letter [A to Z]. Matplotlib, pandas, numpy libraries of Python are being used.
- Google's Tesseract OCR Engine Library and API is being used to compare the recognized text for its accuracy. The IDEs used to build this project are VS Code, Kaggle Notebook, Google Colab.
- For the scope of this assignment, we'll be restricting it to a few specific use cases and mainstream languages like ENG (US) and ENG (INDIA) so we can match the timeline

Methodology

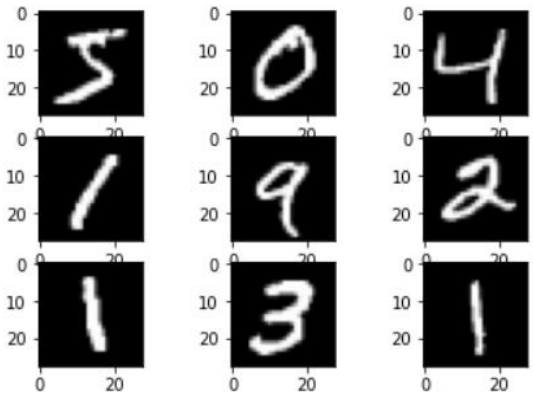
The operation of using images of typed text or letters and then transforming it into data that is understandable by a computer is known as OCR that stands for Optical Character Recognition. This process allows the person to extract text from an image that is taken by a person from any paper document. Therefore, it has become easier to transmute paper documents into computer files that are editable. Artificial intelligence and pattern recognition are some of the notable fields where OCR has proved its efficiency and use.



PROCESS OF OPTICAL CHARACTER RECOGNITION

Algorithm

1. Image Acquisition
   The initial step in the process of Optical Character Recognition is to capture pictures using optical scanners or cameras. By this, an image can be correctly taken and stored. A good Optical Character Reader or OCR scanner should be able to replace each pixel of these colored images with a black or white pixel i.e., to convert it to black and white image. This is done in further stages and the process is known as image segmentation.
2. Pre-processing
   Pre-processing is used to convert raw image data into usable form for computers. The images have noise in it and it should be optimized by removing the areas outside the text making the character clean and get better and more accurate results. Pre-processing is the essential step in recognition of handwritten documents which are more sensitive to noise.
3. Segmentation
   It is the process of grouping the characters obtained from images into meaningful blocks. These blocks of data are then scanned for patterns and matched by the predefined classes.
4. Feature Extraction
   This part of the process indicates dividing the data into subgroups, according to their features meaning is it very important to locate the important characteristics that will aide to recognize patterns. This makes the algorithm to classify each character into a particular class.
5. Training a Neural Network
   After the extraction of all the features they are then sent to the artificial neural network (ANN) so that it can train itself for character recognition. By feeding a large training dataset to the ANN, we can achieve more accurate results depending upon the given problem.



6. Post-processing
   This step involves refinement of the OCR model and correction of errors present in it. But it is difficult to reach 100% accuracy in character recognition. The context heavily influences the process of identification of characters. Human-in-the-loop approach is required for verification of the output which is not always possible.

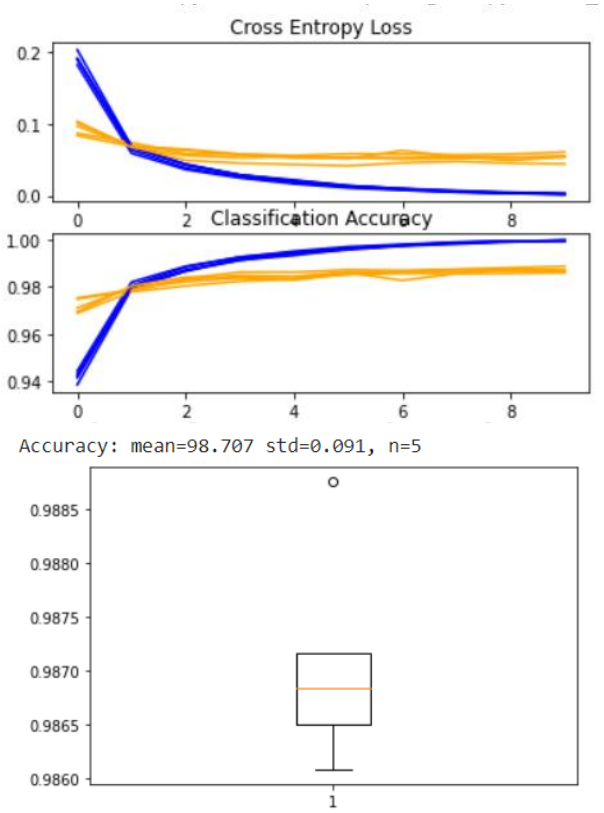
4. Results

We ran the program multiple times, and depending on the input image got different results each time. We ran the program several times and found out these results.

The OCR engine is able to differentiate between characters in most of the cases. It had difficulty in differentiating between the letter 'O' and digit '0' because both have similar circular shapes. It also faces problem in differentiating digit '1' with the digit '7' as the angle between the two lines is similar and has very slight difference.

The machine learning model trained with both the datasets (MNIST and EMNIST) found to be having mean accuracy of 98.707% when tested on the test input data. This percentage varied in various test runs from 98.3% to 98.8%.

We simultaneously compared our results with the results from the Tesseract OCR Engine and found out that both the model has similar accuracy in most of the cases. Sometimes Tesseract OCR Engine outperformed our model which is understandable considering that Tesseract has larger database.



5. CONCLUSIONS

The world is an ever-changing place and to keep up with these changing times, one needs to be open to the amazing world of technology that is making our lives easier and is also helping us to sustain our world with eco-friendly ways. The survival of our planet is depending on us to save our trees and to make less and less use of paper. This technology is a massive step towards our goal of making our earth greener and to be able to make each individual a part of this significant step towards change is what we aspire to achieve with our project. This project makes use of OCR technology to reduce the gigantic paper trail that is generated each year through almost every individual on the planet. The traits that set this project apart from others is the easy accessibility that it provides and the fact that it is free for use by anyone and has the hassle-free nature because the end user does not require to install any software on the device, all it needs is a web browser and anyone with it can get the instant results.

REFERENCES

[1] Karez Hamad & Mehmet Kaya (2016), A Detailed Analysis of Optical Character Recognition Technology, International Journal of Applied Mathematics Electronics and Computers, Issue Special Issue-1, 244 – 249.

[2] Chirag Patel, Atul Patel & Dharmendra Patel (2012), Optical Character Recognition by Open Source OCR Tool Tesseract: A Case Study, International Journal of Computer Applications, (0975-8887) Volume 55-No. 10.

[3] Sebastian Raschka, Joshua Patterson & Corey Nolet (2020), A Machine Learning in Python: Main Developments and Technology Trends in Data Science, Machine Learning, and Artificial Intelligence, Information 2020, 11, 193.

[4] Shamili Syed Rizvon & Karthikeyan Jayakumar (2021), Machine learning techniques for recycled aggregate concrete strength prediction and its characteristics between the hardened features of concrete, Arabian Journal of Geosciences 14(22).

[5] P. Rajesh Pandurangan (2021), Machine learning using Python, ResearchGate, 6/23/2021.

[6] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort & Vincent Michel (2012), Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, 12 (2011) 2825-2830.

[7] Mazdak Fatahi (2014), MNIST handwritten digits Description and using, ResearchGate, DOI: 10.13140/2.1.4601.1681.

[8] JET Akinsola (2017), Supervised Machine Learning Algorithms: Classification and Comparison, International Journal of Computer Trends and Technology, Volume 48 Number 3 June 2017.

AUTHORS' PROFILE

Anirudh Dabral. Currently pursuing his B.Tech. from BBDITM, Lucknow in Computer Science and Engineering.

Anika Bisht. Presently she is working as Assistant Professor at Computer Science department of BBDITM, Lucknow. In this project she worked as the Team Guide.

Ariba Khan. Currently pursuing her B.Tech. from BBDITM, Lucknow in Computer Science and Engineering.

Abhinav Sisodia. Currently pursuing his B.Tech. from BBDITM, Lucknow in Computer Science and Engineering.

Devansh Gupta. Currently pursuing his B.Tech. from BBDITM, Lucknow in Computer Science and Engineering.