Text Recognition Using Tesseract OCR Facilitating Multilingualism: A Review

Lakshmi Aravind[1], Shabin P[2]

[1]*M. tech Student, Dept of ECE, College of Engineering Thalassery, Kerala, India*
[2]*Assistant Professor, Dept of ECE, College of Engineering Thalassery, Kerala, India*

Abstract – *Text detection and recognition are the major areas of experimentation under image processing domain. It is a process by which the system locates any kind of text present and extract them from an image. OCR (Optical Character Recognition) is the electronic or mechanical conversion of images of typed, handwritten or printed text into machine encoded text whether from a document or image. Then later this extracted or recognized text is converted to user desired language. Tesseract OCR is a free software released under Apache License for various operating systems (Linux, Windows and macOS).*

Key Words: **OCR, Tesseract OCR, Apache Licence, Linux, Windows, macOS.**

1.INTRODUCTION

Character recognition is an art of detecting, segmenting and recognizing characters from an image which belongs to the area of pattern recognition and artificial intelligence.

Text detection and its recognition has vast number of applications:

- Text recognized from a recorded image converted into an audio output, helps visually impaired people.
- Assisting the tourists by providing standardized instructions of notice boards and sign boards.
- Conversion of handwritten characters is important for making several important documents related to our history, such as manuscripts converted to machine editable form.
- Another important application of OCR is in banking, where it is used to process cheques without human involvement.
- Automatic number plate recognition is used as a mass surveillance technique making use of optical character recognition on images to identify vehicle registration plates.

2. COMPONENTS OF OCR SYSTEM

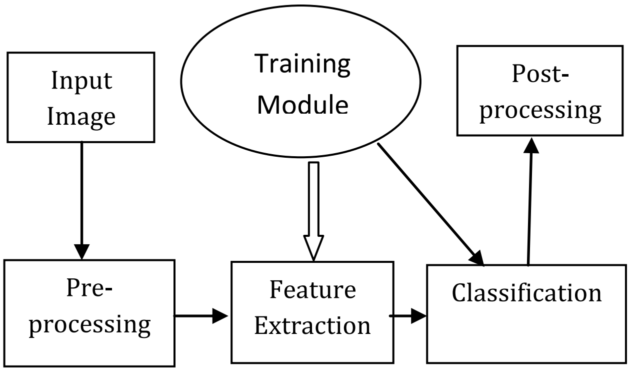

Fig-1: OCR Architecture

Images used to implement the system can be captured live using a camera or also by accessing an existing file from a device. Systems can be trained, tested and validated using the dataset such as ICDAR, MNIST, OSTD, MSRA etc.

Segmentation is a process that determines the constituents of an image, it is very important to segment out the textual and non-textual regions from image. Edge detection is also a vital part in segmentation. In case of text recognition bounding box is implemented for segmentation.

Pre-processing stage includes grey scaling, thresh olding, binarization, removal of noise from the image. Morphological techniques such as dilation and erosion for noise removal. Otherwise filters such as median filter, gaussian filter, bilateral filter etc can be applied for noise removal. Thresholding is implemented using Otsu thresholding method.

The feature extraction is the stage to extract the important information, capture the essential characteristics of the symbols, and it is the most important, difficult step in pattern recognition. In case of text recognition feature extraction is facilitated mainly using geometrical properties (Aspect ratio, Euler's number, Eccentricity, etc.)

The feature vector obtained from previous step is assigned with a class label and recognized using unsupervised and supervised method. The data set is divided training set, validation set and test set. Character classifier can be Bayes classifier, nearest neighbour classifier, Radial basis function, Support vector machine, Linear discriminate functions and Neural networks with or without back propagation.

Post-processing step involves grouping of symbols. This step also ensures that the detected characters correctly recognized as in the image. The process of performing the association of symbols into strings is referred to as grouping.

3. TRANSLATOR

There are different translator API's available in the market such as Google API, IBM CLOUD AZURE API, Linguatools, SYSTRAN io, Yandex etc. But from of all these available API's only few of them such as Google API, IBM CLOUD AZURE API has found to have got a significant accuracy in translation of text from one language to another. IBM API for language identification is available for around 68 languages but only language translation is available for the 12 languages of which only Hindi is found to be the native Indian language. Google translate API is preferred because it supports the greatest number of Indian languages such as (Malayalam, Telegu, Tamil, Kannada etc), which best suits the proposed method

4. EXISTING WORKS ON TEXT RECOGNITION AND TRANSLATION

Venkata Rao, N., Sastry, A. S. C. S., Chakravarthy, A. S. N., & Kalyanchakravarthi proposed a modified back propagation method. Elements of neural network: Inputs, connection weights, desired output, error signal. Claims 100% accuracy [1]. P Sathiapriya R, Manoj J proposed a tesseract OCR based character recognition and its translation from English to Chinese using Bing translator [3]. Rosemol E, Jilu E proposed a method template matching method for character recognition of Malayalam characters where the image is pre-processed using median filter, Sobel operator and bottom profile method (skew-correction) [7]. Deepak C B, Alok A proposed a method of character recognition using HOG (Histogram of Orientation Gradients) and SVM (Linear Support Vector Machine). Translation is implemented using Google Translator API [8]. Nikhil C, Sai Rohit B proposed a method implemented using Dart programming language in Flutter application adopting tesseract OCR and Google translator API [10]. Sukhvinder Singh, Surender Kumar Grewal Rectangular proposed a method using structuring element of size 2*3 for morphological operations. Bounding box and connected component analysis for text classification. OCR for text recognition [11].

5. PROPOSED METHOD

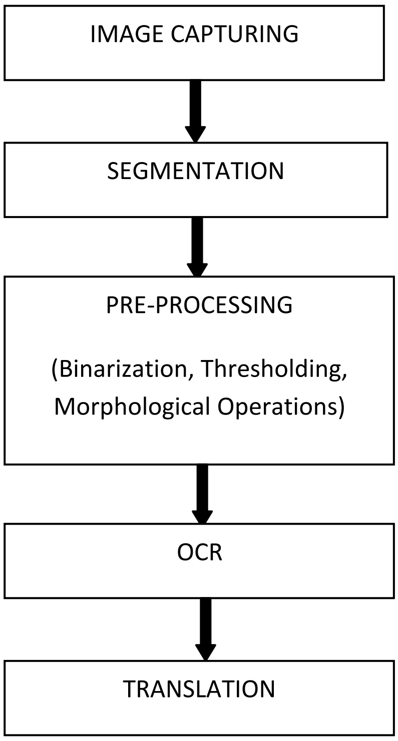

Fig-2: Block Diagram of Proposed Method

The research work is implemented in the platform python. Importing all the packages and tools: Tesseract OCR for character recognition. NumPy package for classification and pattern analysis. Tkinter for GUI. The dataset/images required for the implementation of research work is captured using the camera/accessed from an existing file. Then the input image is pre-processed which includes grey scaling and binarization. Then morphological operations such as dilation and erosion are applied to remove noise. These pre-processing steps are carried out using the libraries in Open CV. Then the processed image is passed to the tesseract OCR for character recognition. Then using Tkinter a message box pops up to select the required language for translation.

The paper concentrates on the major Indian languages such as Hindi, Kannada, Marathi, Malayalam, Tamil, Telugu, Urdu. On choosing the preferred language, Google translator API is used to translate to the desired language. The whole proposed system is implemented as a android application for user compatibility.

6. CONCLUSION

This survey is expected to carry out different text recognition techniques in all type of formats when the input image is provided by the user for recognition. Moreover, the packages and engines used for detection and recognition makes the user to easily carry out text recognition. OpenCV and Tesseract are used widely, and so it is tried to incorporate these technologies in this research for character recognition.

REFERENCES

[1] Venkata Rao, N., Sastry, A. S. C. S., Chakravarthy, A. S. N., & Kalyanchakravarthi, P. (2016). Optical character recognition technique algorithms. Journal of Theoretical and Applied Information Technology,83(2), 275-282.

[2] P Ranjitha; Shamjiith, ; K, Rajashekar (2020). [IEEE 2020 International Conference for Emerging Technology (INCET) - Belgaum, India (2020.6.5-2020.6.7)] 2020 International Conference for Emerging Technology (INCET) - Multi-Oriented Text Recognition and Classification in Natural Images using MSER. , (), 1–5.

[3] Ramiah, Sathiapriya; Liong, Tan Yu; Jayabalan, Manoj (2015). [IEEE 2015 IEEE Student Conference on Research and Development (SCOReD) - Kuala Lumpur, Malaysia (2015.12.13-2015.12.14)] 2015 IEEE Student Conference on Research and Development (SCOReD) - Detecting text-based image with optical character recognition for English translation and speech using Android.

[4] Manage, P., Ambe, V., Gokhale, P., Patil, V., Kulkarni, R. M., & Kalburgimath, P. R. (2020). An Intelligent Text Reader based on Python. 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS).

[5] Adapting the Tesseract Open-Source OCR Engine for Multilingual OCR Ray Smith, Daria Antonova, Dar Shyang Lee, International Workshop on Multilingual OCR 2009, Barcelona, Spain.

[6] Andrew S. Agbemenu, Jepthah Yankey, Ernest O. Addo, “An Automatic Number Plate Recognition System using OpenCV and Tesseract OCR Engine” International Journal of Computer Applications, Volume 180 -No. 43, May 2018. pp. 1-5.

[7] Rosemol Emmanuel, Jilu George, Automatic detection and recognition of Malayalam text from natural scene images, January 2013, IOSR Journal of VLSI and Signal processing 3(2):55-61,DOI:10.9790/4200-0325561.

[8] Bijalwan, Deepak Chandra; Aggarwal, Alok (2014). [IEEE 2014 International Conference on Parallel, Distributed and Grid Computing (PDGC) - Solan, India (2014.12.11-2014.12.13)] 2014 International Conference on Parallel, Distributed and Grid Computing - Automatic text recognition in natural scene and its translation into user defined language.

[9] T. Damak, O. Kriaa, A. Baccar, M. A. Ben Ayed, N. Masmoudi, Automatic Number Plate Recognition System Based on Deep Learning International Journal of Computer and Information Engineering.2020

[10] Chigali, Nikhil; Bobba, Sai Rohith; Suvarna Vani, K; Rajeswari, S (2020). [IEEE 2020 7th International Conference on Smart Structures and Systems (ICSSS) - Chennai, India (2020.7.23-2020.7.24)] 2020 7th International Conference on Smart Structures and Systems (ICSSS) - OCR Assisted Translator, 1–4.

[11] Sukhvinder Singh, Surender Kumar Grewal, Text Extraction and Character Recognition Form Image using Mathematical Morphology and OCR Technique June 2014, International Journal of Science and Research (IJSR) 3(6):952-955.

[12] S.Thiyagarajan , Dr.G.Saravana Kumar, E.Praveen Kumar3 , G.Sakana, Implementation of Optical Character Recognition Using Raspberry Pi for Visually Challenged Person, September 2018, International Journal of Engineering & Technology 7(3):65-67.

[13] Kumar, Vedant; Kaware, Pratyush; Singh, Pradhuman; Sonkusare, Reena; Kumar, Siddhant (2020). [IEEE 2020 International Conference on Smart Electronics and Communication (ICOSEC) - Trichy, India (2020.9.10-2020.9.12)] 2020 International Conference on Smart Electronics and Communication (ICOSEC) - Extraction of information from bill receipts using optical character recognition, 72–77.

[14] Rushikesh Laxmikant Kulkarni, “Handwritten Character Recognition Using HOG, COM by OpenCV & Python”, International Journal of Advance Research in Computer Science and Management Studies, Volume 5, Issues 4, April 2017.

[15] Naidu, C. Dhanunjaya. “A Literature Survey on Character Recognition of Indian Scripts for New Researchers.” (2016).