



TEXT RECOGNITION FROM IMAGES USING IMAGE PROCESSING TECHNIQUE

Bindu Bhargava Reddy C¹, C Tejasundar Reddy², B GopiAmarnath Reddy³, K Bharath Prakash Reddy⁴, Priyadarshini R⁵
^{1,2,3,4}UG Student, Dept. of CSE, REVA University,
⁵Assistant Professor, Dept. of CSE, REVA University, Bangalore, India.
¹bhargavchintam@gmail.com, ²tejareddychejarla465@gmail.com, ³gopiamarnathreddy@gmail.com,
⁴bharathprakashreddy4444@gmail.com, ⁵priyadarshinir@reva.edu.in,

Abstract— Image Processing is a next level character and object recognizer. At present image processing is a major technology that evolved in order to identify things as same as human eyes. It is developed by using a trained model in order to capture objects optically, it has greater error rate in order to capture the character and objects. Nowadays it is popularly used everywhere in various businesses and software tools. A paper document is best source of raw data but filtering required information from it is difficult. In order to gather in such a way, we need information into textual document which easily underpass into filters and produces required output from it. Image processing algorithms are not accurate enough to give efficient output of the text document. It can be improved by using trained model which frequently updates the algorithm efficiency and in order to decrease the word error rate. In this way text can be recognized efficiently from the image. The major distractors in this are image quality, image contrast, pixel level impurities and image backgrounds acts a lot in text recognition. Various improvements are brought in order to get the best out of the image content.

Keywords- OpenCV, Tesseract, Complication Nonaligned Network (CNN), Image Recognition, Image Filters, ASCII filter, Artificial Intelligence (AI).

I. INTRODUCTION

Nowadays there is increased demand for softwares that recognizes valuable data from the paper manuscripts and historic books which are in image format. Recognition of text from these images is bit difficult as there are various disturbances implanted in them. In order to remove them various algorithms are used to extract information from images. These documents are occupying the large amount of physical space in order to store these copies.

These hardcopies cannot be handled for longer durations as days pass the tarnishing of documents takes place. In order to overcome such issues, it needed be to converted into text copy i.e., softcopy. These softcopies can be reused and replenished for future use. In order to convert image text to softcopy we need special algorithms to handle it. These algorithms detect the images in both line by line and word by word.

Due to various font structures and sizes used in different paper documents it becomes harder to identify characters/words. Therefore, algorithms must be efficient enough in order to capture such text from images. Removal polluted text and scraps from image is to done in order retain best quality from image text.

Hence, Image processing Algorithms have done great work in order to recognize the text from image. In this fashion text recognition is performed on various other languages in order to bring back text from images of their respective languages. To skin out various character from images where the infectious backgrounds are removed by using various python tools and modules.

Many researchers proposed various approaches to prevent word error rate and character error rate in order to improve the efficiency. Thereafter, these proposed algorithms are implemented and developed in various ways to decrease error rate drastically.

II. LITERATURE SURVEY

Text recognition from images is a vigour research in order to guess the text from images. Many researchers involved in developing and extracting smart resources in various ways. The common entry point for texted images is pixel clarity which is required and essential as well. At present circumstances the research on text recognition is updated by various legendary persons. Some of them are as follows:

Yang et al. [1] he told about binarization method that is being used filtering the image content in order to get best information out of it. His methods are more accurate and faster in processing and implemented in handsets also. He implemented these contexts in major conferences to put forward the information to society.

Sankaran et al. [2] he described that text recognition process will decrease the word error rate for Indian languages in various manner. He designed various algorithms in order to get text out of images.

Gur et al. [3] he told that there are problems in text recognition and development for that issues in text recognition process. He used special tools which are highly sophisticated in nature and they are implemented

for various languages. He even implemented it for holy books also. His results are highly accurate and more reliable for even handy writings also. He even recovered errors while detecting character in wrong manner and updated them in correct fashion.

Rhead et al. [4] he condemned that number plates used various algorithms in this specifically to apply in general scenario. He started developing various models and methods and even discussed about problems faced while in text identification.

Badawy, W. et al. [5] he conflicted in various issues related to character mapping in text and image observations are taken into consideration. Various applications are implemented by him in various levels of general life style. Miscommunication and mistaken data is taken into observation and he tried to reduce the error rate and negligence rate.

Jawahar et al. [6] he had explained various cases of Indian languages and their matching with the character and object that are present in reports. These image reports helped him in order to decrease the character error ratio by 20% and word ratio by 35%. He specialized method in identifying the text best out of an image.

Ntirogiannis et al [7] his studies tell that document image has various polluted information which can be further used processing it. Pixilation of image done and then process this image in order to get accurate results from handy imaged text to textual form. The performance of his methods is high and can handle larger datasets even in short span of time. This process is continuous and efficient in nature to handle corrupted images also.

Malakar et al. [8] he had described the information has been constantly processes from images to retain correct stuff from it. It had been a challenging task to perform check for this system. Written information is highly difficult to identify each and every character from images. To solve this issue in 2019 he started his work in designing modern style model to get text from images more easily than before. Hence, it replaced may model in that year and listed on behalf of his name universally.

III. METHODOLOGY

The main operation of image processing is to extract information from the images. This extraction of image requires the pixel level clarity to implement the algorithms more efficiently than before. De-pixelating the pixel breakage to formally retrieve the pixel from broken image. Recognition of font and size also plays an important role identifying the text from images. This scenario of implementation requires high-level organization of information by use of simplifying tools and machine learning tools.

This modelling of text in image takes greater effort of python tools and modules. These python AI tools will help us to move forward and bound text with their services. Hence, these specialized methods help us to implement various image processing techniques.

A. Theory Behind Image Processing

Image processing uses various image processing algorithms to convert images to text. This text images are in pixel format in order to convert ASCII codes are implemented by extensive learning models or tools. Various mathematical and character related formulas are implemented in order to recognize text from images. Multi-level layer images are processed before they got implemented in major processing techniques. This is the key role underneath this image processing.

B. Area of Working

- Traffic police are also using image processing in order to convert vehicle registration number to text format which in return helps in tracking details of vehicle.
- Various online resources use conversion of image text to doc or pdf formats.
- Even the medical industry use image processing in order to recognize x-rays and scan reports.
- Banking sector implements conversion passbook image data to text.
- Law department uses this image processing to digitalize their bulky legal libraries.
- Recognition Text for Blind and skim Out (No need of braille text).

IV. DESIGN, IMPLEMENTATION AND ALGORITHMS

The font characteristics of text in a computing system is unlike the font characteristics of text in scripts like pictures and pages in a book, etc., This backlog results in the computer system unable to recognize the text characters while scanning the script. Image processing is a phenomenon in which the contents of script document are stored in the memory of computing system rather than searching and reading the content.

The idea of image processing is not only used to detect English language but its scope can be extended to process various languages around the globe. This motivated us to move forward with the research to check few issues.

A. Objectives

- Pre-processing Module (Processing Image into OCR form)
- System Training Module (Machine Learning)
- Text Recognition Module (Various Text /Styles /Fonts /Languages Conversion)
- Post-processing Module (Conversion to respective text format (Digital)).

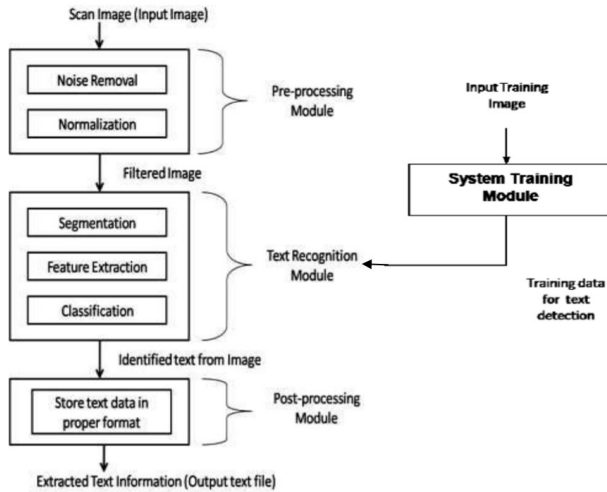
Other major objectives: -

- Line detection from image
- Character detection from the line

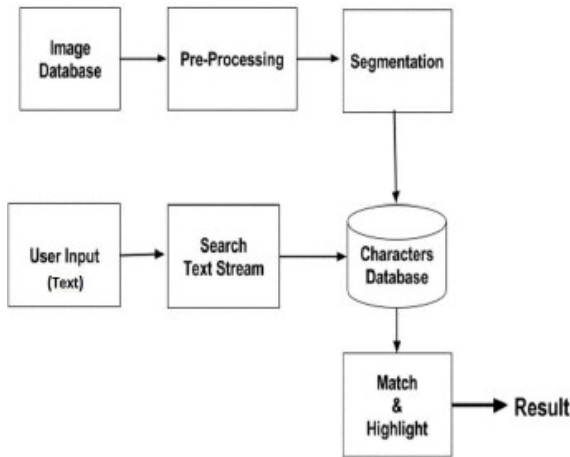
B. System Requirements

Software Requirements	Hardware Requirements
Python 3.7 Tesseract OCR(Google) Pillow OpenCV Numpy Wheel Database (MySQL)	Camera

C. Implementation

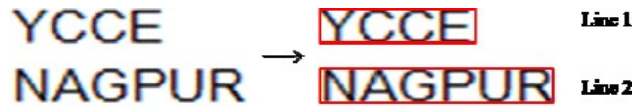


D. Block Diagram



V. APPLICATION OVERVIEW

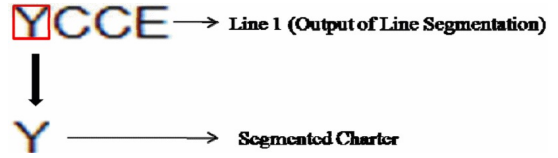
Output of Line Segmentation



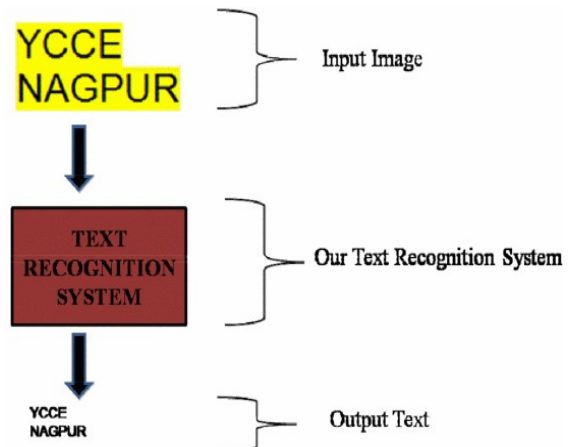
Output of preprocessing

Line Segmentation

Output of Character Segmentation



Output Text Recognition System



A. Expected New Applications

To scan a paper document, primarily the script will be scanned by an optical scanner and the duplicate of the script is stored in the form of images. The amalgamation of picture elements well known as pixels forms the images. The Important information is to be retrieved from these images, since the information is in the form of images, they must be further analyzed. So, we use our method to recognize text which we already discussed in this paper and the output results are shown in the form of ASCII characters.

VI. CONCLUSION

This Paper proposes and discusses the Text Recognition Method (Recognizing text from images). A wide range of researchers use Optical Character Recognition (OCR) for pattern recognition and loads of research work has been done based on the Optical Character Recognition (OCR) for various languages around the globe along with English. This Image

Processing method attracted the interest of many more researchers. Every stage of OCR has its own implication and will be updated for improved results: -

- Traffic police are also using image processing in order to convert vehicle registration number to text format which in return helps in tracking details of vehicle.
- Various online resources use conversion of image text to doc or pdf formats.
- Even the medical industry use image processing in order to recognize x-rays and scan reports.
- Banking sector implements conversion passbook image data to text.
- Law department uses this image processing to digitalize their bulky legal libraries.
- Recognition Text for Blind and skim Out (No need of braille text).

VII. ACKNOWLEDGMENT

We take it a considered entitlement to precise our honest because of all concerned. We acquiesce our thankfulness and truthfulness because of our University and Department of Computer Science Engineering (CSE), for his or her continuous inspiration and care and has helped us form the matter and providing awareness towards the response. We are obliged to the REVA University, Bangalore, India for providing the required amenities for finishing this work.

VIII. REFERENCES

- [1] Yang, Jufeng, Kai Wang, Jiaofeng Li, Jiao Jiao, and Jing Xu, "A debauched adaptive binarization method for complex section images," 19th IEEE International Conference on Image Processing (ICIP), 2019.
- [2] License plate detection; Ching-Tang Hsieh, Yu-Shan Juan, Kuo-Ming Hung, *emph*{Proceedings of the 19th International Conference on Innovative Information Networking and Applications (IINA'05) 2 017 IEEE}
- [3] Optical Character Recognition by Open Source OCR Tool Tesseract: A Case Study; Chirag Patel, Atul Patel, Dharmendra Patel, *emph* {International Journal of Computer Applications (0975 – 8887) Volume 55– No.10, October 2018}.
- [4] Rhead, Mke, "Accuracy of Automatic Number Plate Recognition (ANPR) and real-world UK number plate problems." IEEE International Camahan Conference on Security Technology (ICCST), 2018.
- [5] Badawy, W. "Automatic License Plate Recognition (ALPR): A National of the Art Assessment." IEEE International Conference on Document Analysis and Recognition, 2018.
- [6] Naveen Sankaran and C.V Jawahar, "Recognition of Printed Devanagari Text Using BLSTM Neural Network," IEEE, 2018.
- [7] Ntirogiannis, Konstantinos, Basilis Gatos, and Ioannis Pratikakis. "A Performance Evaluation Methodology for Historical Document Image Binarization." IEEE International Conference on Document Analysis and Recognition, 2019.
- [8] Malakar, Samir, et al. "Text-line extraction from handwritten document pages using spiral run-length smearing algorithm," IEEE International Conference on Communications, Devices and Intelligent Systems (CODIS), 2019.
- [9] Itseez leads the development of the renowned computer vision library OpenCV. [http:// itseez.com](http://itseez.com) .
- [10] Bradski, Gary; Kaehler, Adrian (2018). Computer vision with the OpenCV library. O'Reilly Media, Inc. p. 6.
- [11] [http:// search.cpan.org/~ yuta/Cv-0.29/](http://search.cpan.org/~yuta/Cv-0.29/)
- [12] "The Red-Monk Programming Language Rankings: June 2015 – ecosystems". Redmonk.com. 1 July 2016. Retrieved 10 September 2019.
- [13] Shrey Dutta, Naveen Sankaran, Pramod Sankar K., C.V. Jawahar, "Robust Recognition of Degraded Documents Using Character N - Grams," IEEE, 2019.
- [14] Gur, Eran, and Zeev Zelavsky, "Retrieval of Rashi Semi-Cursive Handwriting" in the IEEE International Conference on Frontiers in Handwriting Recognition (ICFHR), 2018.
- [15] Rhead, Mke, "Accuracy of Automatic Number Plate Recognition (ANPR) and real-world UK number plate problems." IEEE International Camahan Conference on Security Technology (ICCST), 2019.
- [16] Badawy, W. "Automatic License Plate Recognition (ALPR)" in IEEE International Conference on Document Analysis and Recognition, 2019.
- [17] Application of OCR, from [http:// www.cvisiontech.com](http://www.cvisiontech.com).
- [18] Kay, Anthony (July 2019). "Tesseract: an Open-Source Optical Character Recognition Engine". Linux Journal. Retrieved 28 September 2019.
- [19] Vincent, Luc (August 2006). "Announcing Tesseract OCR". Archived from the innovation on October 16, 2017. Retrieved 2019-06-2.