International Journal Of Multidisciplinary Research In Science, Engineering and Technology (IJMRSET)

| ISSN: 2582-7219 | www.ijmrset.com |

| Volume 2, Issue 5, May 2019 |

Portable Handheld Optical Character Recognition Scanner

Shivani Magdum, Milind Sutar

Department of Electronics Engineering, Vishwakarma Institute of Technology, Pune, India

ABSTRACT: The paper explains how an Optical Character Recognition system (OCR) works and how this system enables us in capturing an image of a text document. It also explains the fact how OCR is more efficient and easier tool which can be an alternative to scanning a document as the image captured using OCR is of exactly the same quality which is obtained like in a scanned copy, the only difference which is there is that OCR is done with the help of a simple mobile phone camera whereas scanning requires a bulky scanner. It also explains the problems which are being faced by the developers in using Optical character recognition as a technology on a huge scale and what are the probable solutions to this. The proposed OCR system provides many features which requires no typing, no editing of raw data, quick translation, and memory utilization. In the end it also highlights the important emerging trends in the field of OCR and how OCR can be an evolving technology.

KEYWORDS: Optical Character Recognition System (OCR), Camera Captured Document Images, Handheld Device, Image Segmentation, Pattern Recognition, Image Segmentation, Text Extraction, Tesseract.

I.INTRODUCTION

A Person is able to capture any image because of the communication between our eyes and brain. Our eyes act as an optical lens and the images seen by our eyes are an input for our brain and the ability to understand visualize these images varies from person to person. Similarly we have the technology called as OCR, where OCR stands for Optical Character Recognition, which has an automated mechanism which allows easier recognition of character and its processing. Earlier there were scanners which were the only working OCR applications available in the market. The main disadvantage of scanners was that it was heavy, bulky, and not portable and it takes a lot of time to capture an image. But with today's device which have better processing speeds, larger internal memory and an excellent rear camera, researchers have to think of running OCR applications on devices such as smart gadgets like mobile phones for having real time imaging monitoring. Applications such as Cam Scanner and Google translate powered by Google are the major examples of Optical character Recognition applications. It also showcases the fact that this technology can be put to use in a wide array of streams and hence is a very important concept which requires more attention towards research and development.

II.LITERATURE SURVEY

OCR is capable of automatically recognizing characters through an optical mechanism. It is capable of recognizing both handwritten and printed text. Its performance depends upon the quality of the documents and the camera being usedto capture the raw image. OCR system is designed in such a way that it processes images which contain maximum text with very less number of graphic elements. As mentioned earlier, most of the character recognition programs and algorithms will be working efficiently only on those images which are captured using a scanner or a digital camera and which can run on a computer software. But since in this case the size and portability are the most important factors which will further hamper the growth and usability of this technology, in order to overcome the above mentioned drawbacks, a character recognition system based on android devices is proposed[1]. OCR enables working on Android mobile operating system by combining open-source OCR engine powered by Google called, Tesseract and the text recognition OCR engine[3]. The text-to- speech synthesizers in a mobile device allows the users to take photographs of text using the camera which is present inside the mobile phones and have the capability to read the text read aloud by the mobile phone. OCR as a technology provides us different ways to convert various types of documents such as scanned papers, PDF files or images captured by a digital camera into editable and searchable data. A point worth noting here is that the

| ISSN: 2582-7219 | www.ijmrset.com |

| Volume 2, Issue 5, May 2019 |

images obtained by a digital camera differ from those scanned documents or images as they often have distortions in their captured images. These distortions and noise makes it difficult to recognize the text accurately. Pre-processing is carried out on the image to improve the accuracy of text recognition.

A. How OCR works

The proposed OCR model provides many features which do not require typing, editing of raw data, quick translation, and memory consumption.Tesseract is voted as the engine for the OCR because of its widespread approbation, extensibility and flexibility properties, its community of developers which are always active, and the fact that it works out of the box. To perform the character recognition process, our application has to go through two major steps which are as follows:- 1. Segmentation, i.e., given a binary input image, to identify the individual glyphs present (basic units representing one or more characters, usually contiguous). 2. Feature extraction, i.e., to compute from each glyph a vector of numbers that will serve as input features for an ANN.

B. What is Tesseract?

Tesseract is an open source engine powered by Google for optical character recognition. It is available on many operating systems. It is considered as one of the most accurate OCR

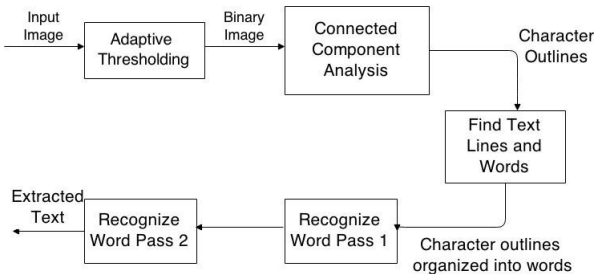

Fig. 1. Architecture of Tessaract

engine available. It can read images and convert them into as many as 60 languages. It was developed at HP between years 1984 to 1994 but its first working copy was released only in 2005 as open source by HP. Tesseract converts the input image into binary image format using the process called thresholding. Outlines of components are stored on connected Component Analysis. Nesting of outlines is done which gathers the out- lines together to form a Blob. Text lines present are analyzed for fixed pitch and proportional text. Then the lines obtained are broken into words by analysis according to the process of character spacing. Fixed pitch is chopped into character cells and proportional text present is broken into words by definite spaces and fuzzy spaces. Tesseract recognizes a word in two passes, that is, it tries to recognize the words in the first pass. If the match is found, then the found word is passed on to the Adaptive Classifier, which recognizes the text more accurately. During the second pass, the words which were not at all recognized or were not well recognized in the first pass are recognized again through a run over through the page. Finally Tesseract resolves fuzzy spaces. In order to locate small and capital text, Tesseract always checks alternative hypothesis for x-height[4].

III. CONSTRUCTION OF OCR

Using Raspberry Pie, we are developing OCR scanner which is faster and extremely portable than the desktop scanners. It can capture the exact segment of the text we need; it instantly inserts scanned data into the desired field within our application. Nevertheless, of, curved, laminated or patterned all surfaces can be scanned using this scanner. To achieve this, we are using image processing on python platform which will detect the characters of the scanned surface and using database management we will access the stored information(offline or online) to compare, edit or enter a new data. To make it completely portable, we are creating an android app studio which we display and scan the number plates, credit card, water- gas meters and many more. The input image can be of any type. It can be a document, live text,

| ISSN: 2582-7219 | www.ijmrset.com |

| Volume 2, Issue 5, May 2019 |

journals, magazines etc. Here we are using an vehicle number plagte as an example for from which text is to be extracted. The functioning of OCR consists of the following steps: scanning, segmentation, pre- processing, feature extraction and recognition.[1] The input image is firstly a scanned copy using an Android mobile camera. This is done to digitize the document. Segmentation extracts any symbols in the text region. Noise is removed by pre-processing each symbol, and the characteristics of each symbol are extracted using feature extraction to finally recognize the text.

A. Scanning

Android mobile camera is used to capture the image of document. This process is called scanning. This is nothing but the process of scanning which converts the document into digital image. The digital image is then converted into a gray scale image using thresholding function hresholding is the process which converts multi-level image into bi-level image
i.e. black and white image. Black is represented if the gray level is below the threshold level, and it is represented by white if the gray level is above the threshold level. This makes very easy to detect the text regions in an image. It also saves a lot of memory space and processing time.

B. Segmentation

Regions of text are detected using the process of segmenta- tion. It differentiates the text from other graphical elements in the document. Splits and joints present in an image can cause confusion between text and graphic elements present in the document which results in incorrect segmentation of the text.
[1] This mainly occurs due to poor scanning which increases the noise levels in the digital document. Joints in characters occur when the document is scanned at low threshold and splits occur when the document is scanned at high threshold.

C. Pre-processing

During scanning process, some noise is produced in the scanned image. This results in poor recognition of characters. This noise can be reduced by pre-processing. Pre-processing is done using smoothing and normalization. Smoothing is performed on the image with the help of filling and thinning techniques. Normalization is responsible to handle uniform size, slant and skew correction

D. Feature Extraction

Feature extraction refers to the process of extraction of fea- tures of symbols from the image. In this step, only important attributes are taken into account and any unnecessary attributes are ignored. This technique takes into account the abstract features which are present in the character. Spaces, lines, intersections etc are some of the abstract features. Feature ex- traction is done using Tesseract algorithm. Tesseractalgorithm is used to implement feature extraction.

E. Recognition

OCR system uses algorithm defined by Tessaract to identify characters from the image foreground pixels which are called as blobs and recognizes the lines present. These lines are then recognized further into letters or words or characters. In this process the image is converted into character stream which represents letters or group of letters called words.[5].

| ISSN: 2582-7219 | www.ijmrset.com |

| Volume 2, Issue 5, May 2019 |

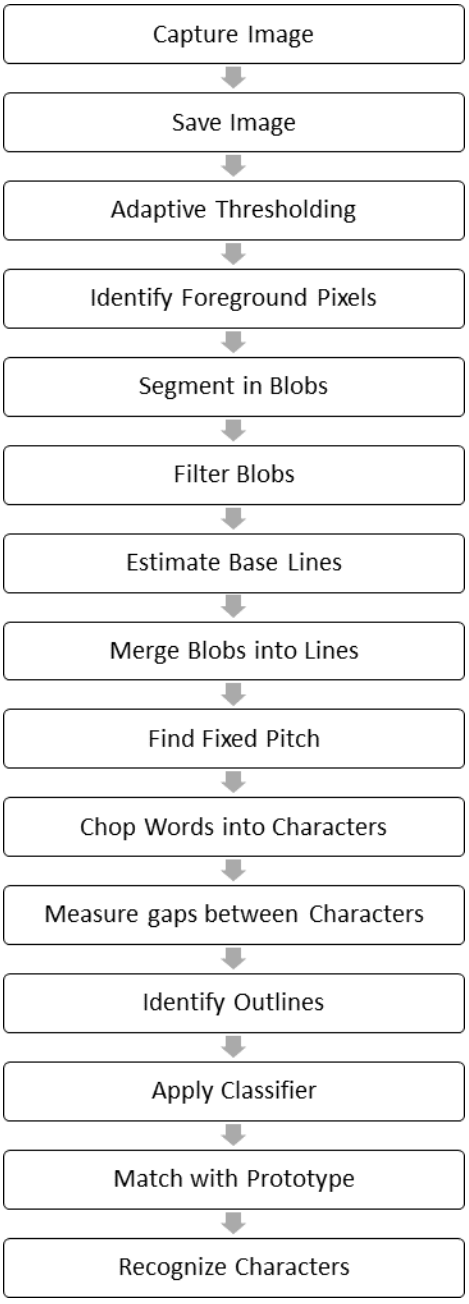

Fig. 2. Flowchart for the OCR system

IV.RESULTS

As we can see, first we acquire the image of a number plate. Then the image goes through pre-processing and segmentation. After successful feature extraction and recognition of charac- ters using OCR technique, the result is obtained. That can be seen in the picture shown below. Also this result get stored in CSV file which remains in a database and this database can be accessed or updated as per the requirement.

| ISSN: 2582-7219 | www.ijmrset.com |

| Volume 2, Issue 5, May 2019 |

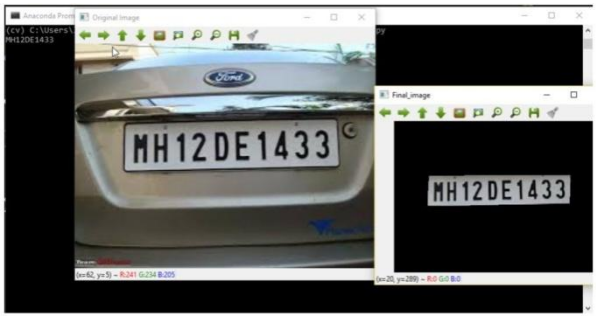

Fig. 3. Number plate to be recognised

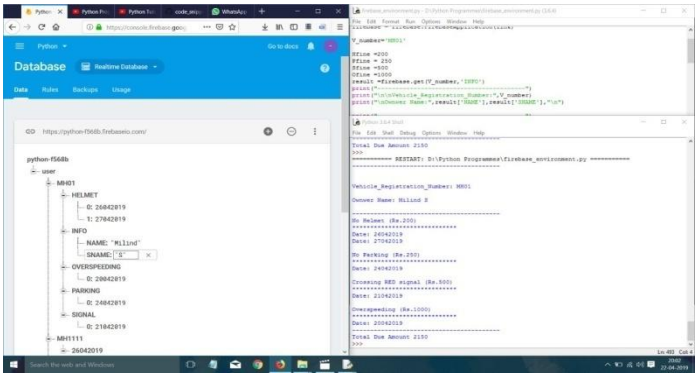

Fig. 4. Formation of Database and Result Obtained

V.CONCLUSION

This paper explains about Optical character recognition for the devices as well as handheld devices in recognizing characters in an offline mode. The system has the ability to recognize characters with accuracy exceeding 90 percent mark. The advantage of the system is that it is easily portable and its property of scalability which can recognize various languages and providing help in translating the text in various languages. Recognition is often followed by a post-processing stage. If post-processing is done on the output image, the accuracy can be increased. The future scope is to develop software for automatic editing and searching.

REFERENCES

[1]Heuristic-Based OCR Post-Correction for Smart Phone Applications, the University of North Carolina at Chapel Hill department of computer science honors thesis Author: Wing-Soon Wilson Lian 2009.
[2]R.W. Smith, The Extraction and Recognition of Text from Multimedia Document Images, PhD Thesis, University of Bristol, November 1987
[3]The Tesseract open source OCR engine, http://code.google.com/p/tesseract-ocr.
[4]R. Smith. "An overview of the Tesseract OCR Engine." Proc 9th Int.Conf. on Document Analysis and Recognition, IEEE, Curitiba, Brazil,Sep 2007
[5]An English Language OCR, 2010 Second International Conference on Computer Engineering and Applications. Junaid Tariq, Umar Neumann Muhammad Marinara
[6]A survey of modern optical character recognition techniques (DRAFT), February 2004
[7]An overview of character recognition methodologies - by J.Mantas, volume 19, issue 6, feb-2004
[8]Character recognition — A review- by V.K. Govindan, A.P. Shivaprasad, volume23, issue 7, 1990.
[9]Text detection and recognition in images and video frames- by Datong Chen, volume 37, issue 3, march 2004
[10]OPTICAL CHARACTER RECOGNITION — A SURVEY, Interna- tional Journal of Pattern Recognition and Artificial Intelligence, volume- 05-1991
[11]Document Image Retrieval through Word Shape Coding- by Shijian Lu, Linlin Li, Chew Lim Tan, IEEE Transactions, volume30, issue 1, April- 2008
[12]Optical Character Recognition Implementation Using Pattern Matching- by Faisal Mohammad, Jyoti Anarase, Milan Shingote, Pratik Ghanwat , International Journal of Computer Science and Information Technolo- gies, Vol. 5 (2) , 2014,