ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Volume 4, Issue 4, April 2017

Extraction of Text from an Image and its Language Translation Using OCR

[1]G.R.Hemalakshmi, [2]M.Sakthimanimala, [3]J.Salai Ani Muthu
[1]Assistant Professor, [2] [3]U.G. Student
[1] [2] [3]Department of Computer Science and Engineering, National Engineering College, TamilNadu

Abstract :- — **In our day to day life the people are facing many problems in understand the languages. For example, if the people move from one state to the other they don't understand their language at that time this Mobile Application will help them. Existing system, having a separate application for each and every process like camera, Google translator and Optical Character Recognition(OCR) text scanner. But, people expect the application consists of all the three facilities together. So this proposed application provides a new idea to the people to translate the other language text into their known language. This application contains three steps. 1.Take a photo image of the unknown language text which you want to translate(either handwritten or printed material), 2.Tessaract is an open source Optical Character Recognition (OCR) technology, which is used to extract the text from the image then Google API and Bing API is used for translation of language. 3.The translated text is generated in PDF format.**

Keywords: **- Text Extraction, Android, OCR, Tesseract.**

I. INTRODUCTION

Text extraction from image is one of the complicated areas in digital image processing. It is a complex process to detect and recognize the text from image. It's possible of computer software can provide extracted text from image using most complicated algorithm. So it can't be use anywhere in this existing environment. Here different types of language translators are available such as voice based translator, keyboard based translator etc. But those translators are not easy to use. The purpose of this work is to demonstrate that a tight dynamical connection may be made between text and interactive visualization imagery. The Android device camera can prove this type of extraction and also the algorithm will easily implemented using java language. Millions of mobile users in this world and they always have mobile in their hand, so simply they can capture the image to extract the text.

The purpose of this project is to implement text extraction from the image and translating the text. Captured text information from camera in natural scene images can serve as indicative marks in many image based applications such as assistive navigation, auxiliary reading, image retrieval, scene understanding, etc. Extracting text from natural scene images is a more challenging problem as compared to scanned document because of complex backgrounds and also large variations of text patterns such as font, color, scale, intensity and orientation.

Therefore, to extract text from camera captured images, text detection & extraction is an important and essential step which computes the sub-regions of the images containing text characters or strings. Once the image is captured from camera, the image went through various processes whose task is to detect the text within the image and extract those texts then translates that text.

II. LITERATURE REVIEW

This material serves as a guide and update for readers working in the Character Recognition area. Ayatullah Faruk Mollah, Nabamita Majumder, Subhadip Basu, and Mita Nasipuri(2011) [1] presents a complete Optical Character Recognition(OCR) system for camera captured image textual documents for handheld devices. Firstly, text regions are extracted and skew corrected. Then, these text regions are binarized and segmented into lines and characters. Characters are passed into the recognition module. Pranob K Charles, V.Harish, M.Swathi(2012)[2] describes the techniques for converting textual content from a paper document into machine readable form. The computer actually recognizes the characters in the document through a revolutionizing technique called Optical Character Recognition. Chirag Patel ,Atul Patel, Dharmendra Patel(2012) [3] recognize the characters in a given scanned documents and study the changes in the Models of Artificial Neural Network. It describes the behaviors of different Models of Neural Network used in Optical Character Recognition.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Volume 4, Issue 4, April 2017

Neural network mostly uses the OCR. Dileep Kumar Patel, Tanmoy Som, Sushil Kumar Yadav ,Manoj Kumar Singh [2012][4] gives the solution to the problem of handwritten character recognition. It has been tackled with multi resolution technique using Discrete wavelet transform (DWT) and Euclidean distance metric (EDM). The technique has been tested and found to be more accurate and faster. Characters is classified into 26 pattern classes based on appropriate properties. Chi et al. (2012) [5] has proposed an effective algorithm to deal with bleed-through effects existing in the images of financial documents.Double-sided images scanned simultaneously are used as inputs, and the bleed-through effect is detected then removed after the registration of the side images. Satyajitsaha, Dnyaneshwar, Hagawane, Pravin C.Kulkarni, Swapni R.Dhamane (2013)[6] proposes the objective to recognize and extract the text from images captured by camera based mobile device, and once the text is recognized then information about the text can be obtain via Dictionary or via Web. Majida Ali Abed et al.(2013)[7] presents a new approach to simplify Handwritten Characters Recognition based on simulation of the behavior of schools of fish and flocks of birds that is called the Particle Swarm Optimization Approach (PSOA).PSOA is convergent and more accurate in solutions that minimize the error recognition rate. Vijay Laxmi Sahu et al(2013)[8] explains that characteristics of the classification methods that have been successfully applied to character recognition and remaining problems that can be potentially solved by learning methods. Argha Roy, Diptam Dutta K Austav, Choudhury (2013)[9] explains the IRIS plant classification using Neural Network.It provides the adaptation of network weights using Particle Swarm Optimization (PSO) was proposed as a mechanism to improve the performance of Artificial Neural Network (ANN) in classification of IRIS dataset. Classification method is a machine learning technique used to predict group membership for data instances. Amir Bahador Bayat(2013)[10] proposes an efficient system that includes two main modules, the feature extraction module and the classifier module. In the first module, seven sets of discriminative features are extracted and used in the recognition system. In the second module,the adaptive neuro-fuzzy inference system is investigated. N.K.Gundu, S.M.Jadhav, T.S.Kulkarni, A.S.Kumbhar(2014)[11] explains the best ideas from the text extraction with the help of character description and stroke configuration, web context search and web mining with the help of semantic web and synaptic web at low entropy. Faisal Mohammad, Jyoti Anarase, Milan Shingote, Pratik Ghanwat(2014)[12] presents an algorithm for implementation of Optical Character Recognition (OCR) to translate images of typewritten or handwritten characters into electronically editable format by preserving font properties.OCR can easily do this by applying pattern matching algorithm. The recognized text characters are stored in editable format. Shalin A. Chopra, Amit A. Ghadge, Onkar A. Padwal, KaranS. Punjabi, and Prof. Gandhali S. Gurjar(2014) [13] presents a simple, efficient and minimum cost approach to construct OCR for reading any document that has fix font size and style or handwritten style.In this the systems have the ability to yield excellent results. It is mostly used with existing OCR methods, especially for English text. Sravan, ShivankuMahna, NirbhayKashyap (2015)[14] explains that problems being faced by the developers in using OCR as a technology on a large scale and give the solution to that problem. This system provides many features that require no typing, editing raw data, quick translation, and memory utilization.Surabhi Dusane, Monica Ahuja, Rucha Ghodke & Prathamesh Kothawade (2016)[15]The objective in this paper is to develop user friendly system which will extract text from images and convert the extracted text into user friendly language then it will convert it into audio which describes the text more efficiently.

III. PROPOSED SYSTEM

Optical Character Recognition (OCR), is a technology that enables you to convert different types of documents, such as scanned paper documents, PDF files or images captured by a digital camera into editable and searchable data. It is the mechanical or electronic conversion of images of typewritten or printed text into machine encoded text. Images captured by a digital camera differ from scanned documents or image-only PDFs. They often have defects such as distortion at the edges and dimmed light, making it difficult for most OCR applications, to correctly recognize the text. The latest version of ABBYY Fine Reader supports adaptive recognition technology specifically designed for processing camera images. It offers a range of features to improve the quality of such images, providing you with the ability to fully use the capabilities of your digital devices.

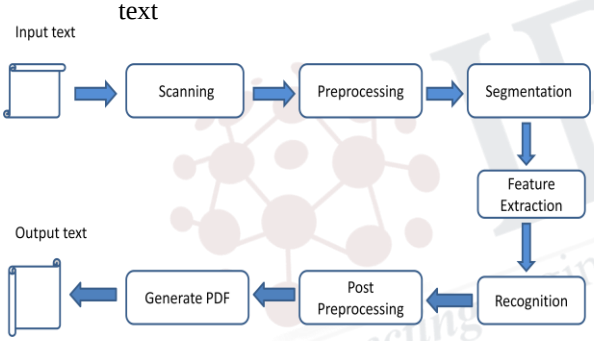
A common problem faced by travelers is that of understanding unfamiliar language. Failing to understand unknown languages, when travelling can lead to minor problems. These systems are usually composed of two subsytems that perform text extraction and text translation respectively. The extraction and translation parts are relatively well developed and there exist a large variety of software packages or web services that perform these tasks. The challenge is with extracting the exact text from the

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Volume 4, Issue 4, April 2017

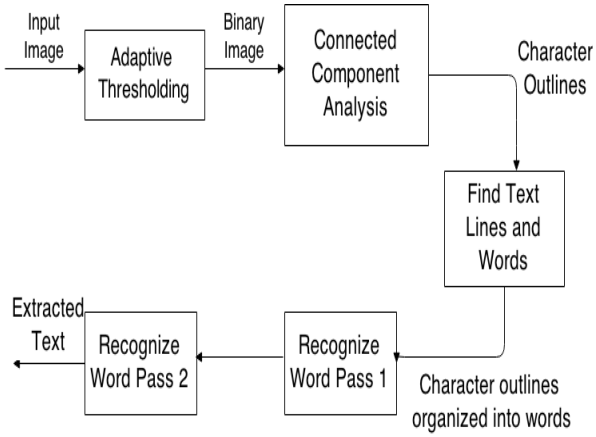
images and translating it to known language. In a typical scenario, a user takes a picture of a text area with the cell phone camera, the text is extracted from the image. In image processing and computer vision, edge detection treats the localization of significant variations of a gray level image and the identification of the physical and geometrical properties of objects of the scene. The variations in the gray level image commonly include discontinuities (step edges), local extreme (line edges) and junctions. Most recent edge detectors are autonomous and multiscale then include three main processing steps smoothing, differentiation and labeling. The edge detectors vary according to these processing steps, to their goals, and to their mathematical and computational complexity. The extracted text is then translated using translation engine which contains the database of languages. Then the translated text is given as output.

The Purpose of this project is to implement text extraction from the image and translating the given text. Many different methods are used for extracting the text from the images. Properties like color, intensity, edges etc are related in extracting the text



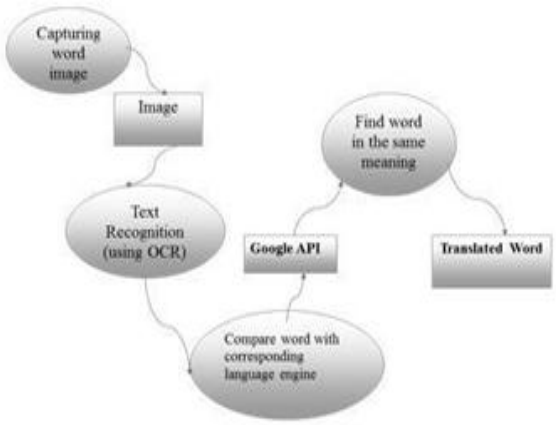
To carry out this task we have four modules those are Image Capture, Text Identification, Language conversion, PDF generation.

A mobile camera is used to capture the image. It is important to learn how to use a mobile camera properly so that you can convert an text to image effectively and get the most accurate results. The text is given as input and image is get as output The input image is first pre-processed to remove the noise present in the image. The image is converted into a grayscale image which can then be converted into binary image. Tesseract is an Optical Character Recognition engine for various operating systems. It is free software, released under the Apache License, Version 2.0, and development has been sponsored by Google since 2006.Tesseract is considered one of the most accurate open source OCR engines currently available. The total count of support languages to over 60.It is the tool used to extract the text from an image.



After extracting the words from image by using the Optical Character Recognition (OCR) Engine, those words are translated into known language to do this the Bing Translator API service is used. This is a free service. It provides many libraries for translation. The first thing to remember is that translation is the transfer of meaning from one language to another.

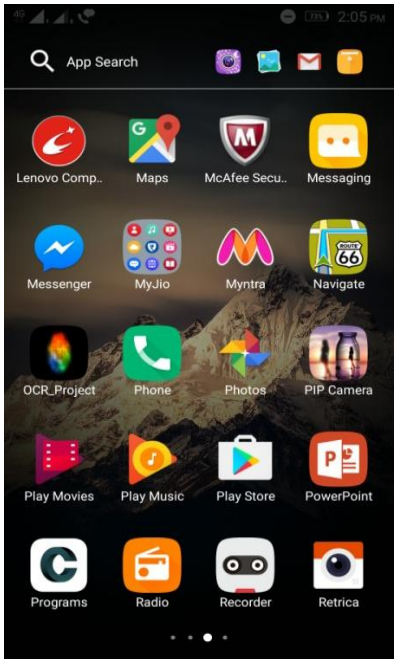
The efforts have been dedicated to extracting the text content from an image taken by a cell phone camera, and getting it as close as possible to the original. The final step is translating the text using the language translator. The language is translated with the help of being translator. After that the input text and output text is generated in PDF format. Generate and edit high volumes of PDFs programmatically with iText, you can assemble, expand, extract, split and interact with any PDF file. iText allows you to spend your time more productively by automating routine documentation, invoicing and archiving tasks. Here the input text is generaterd as pdf format in left side and translated text is generated as pdf format in right side and this pdf is stored in internal memory.

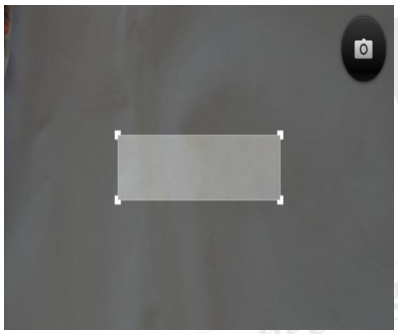


ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Volume 4, Issue 4, April 2017

IV. RESULTS





V. CONCLUSION

This is the discussion about optical character recognition techniques to translate the text from unknown language text into known language.The system has the capability to recognize characters with accuracy exceeding 90% mark. The advantage of this system is that it is easily portable and its scalability which can recognize various languages and also help in translating the text in different languages. The accurate recognition is directly depending on the nature of the material to be read and by its quality.

REFERENCES

[1] Ayatullah Faruk Mollah, Nabamita Majumder, Subhadip Basu and Mita NasipuriDesign of an Optical Character Recognition System for Camera-based Handheld Devices,IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 4, No 1, July 2011.

[2] Pranob K Charles, V.Harish,M.Swathi, CH. Deepthi,"A Review on the Various Techniques used for Optical Character Recognition",International Journal of Engineering Research and Applications (IJERA) ISSN:2248-9622, Vol. 2, Issue 1, Jan-Feb 2012.

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

[4] Dileep Kumar Patel, Tanmoy Som1, Sushil Kumar Yadav, Manoj Kumar Singh," Handwritten Character Recognition Using Multiresolution Technique and Euclidean Distance Metric" JSIP 2012, 208-214 .

[5] Chi, Bingyu, and Youbin Chen. "Reduction of Bleed-through Effect in Images of Chinese BankItems." Frontiers in Handwriting Recognition(ICFHR), 2012 International Conference on IEEE.

[6] Satyajit S.Saha, Dnyaneshwar S. Hagawane , Pravin C.Kulkarni ,Swapnil R.Dhamane ,Prof. S.A. Agrawal, Mobile Based Text Detection and Extraction from an Image, International Journal of Emerging Technology and Advanced Engineering (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 3, Issue 11, November 2013).

[7] Majida Ali Abed, Hamid Ali Abed Alasadi "Simplifying Handwritten Characters Recognition Using a Particle Swarm Optimization Approach" European Academic Research, Vol. I, Issue 5/ August 2013.

[8] Vijay Laxmi Sahu, Babita Kubde "Offline Handwritten Character Recognition Techniques using Neural Network: A Review"International Journal of Science and Research (IJSR), India Online ISSN: 2319-7064 Volume 2 Issue 1, January 2013.
[9] Argha Roy, Diptam Dutta, Kaustav Choudhury," Training Artificial Neural Network using Particle Swarm Optimization Algorithm" IJARCS SE Volume 3, Issue 3, March 2013.
[10] Amir Bahador Bayat, Recognition of Handwritten Digits Using Optimized Adaptive Neuro-Fuzzy Inference Systems and Effective Features, Journal of Pattern Recognition and Intelligent Systems Aug. 2013, Vol. 1 Iss. 2, PP. 25-37.
[11] N.K. Gundu1, S.M. Jadhav2, T.S. Kulkarni3, A.S. Kumbhar, Text Extraction from Image and Displaying its Related Information, International Journal of Scientific and Research Publications, Volume 4, Issue 12, December 2014 1 ISSN 2250-3153.
[12] Faisal Mohammad, Jyoti Anarase, Milan Shingote, Pratik Ghanwat, Optical Character Recognition Implementation Using Pattern Matching, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (2), 2014, 2088-2090
[13] Shalin A. Chopra, Amit A. Ghadge, Onkar A. Padwal, KaranS. Punjabi, and Prof. Gandhali S. Gurjar , " Optical Character Recognition" International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 1, January 2014.
[14] Sravan,ShivankuMahna,NirbhayKashyap, Optical Character Recognition on Handheld Devices, International Journal of Computer Applications (0975 – 8887) Volume 115 – No. 22, April 2015.
[15] Surabhi Dusane, Monica Ahuja, Rucha Ghodke & Prathamesh Kothawade,Text To Speech Synthesizer, Imperial Journal of Interdisciplinary Research (IJIR) Vol-2, Issue-5, 2016 ISSN: 2454-1362.

● ● ●