ScanIT- Extracting text of the Document from its Image and Creating its exact Soft copy on Laptop using Android Platform

O.S.KAMATE, H.B.MALUNJKAR , S.N.RAUT, V.M.DESAI, M.A.PULSE,

Final Year,Bachelor of Engineering,Department of Computer Engineering, ZealCollege of Engineering and Research, Maharashtra, India

ABSTRACT

ScanIt is an android app basically designed for implementing the Optical Character Recognition (OCR) technology through android platform . The main objective of the app will be to recognize the alpha-numeric characters from a hard copy of a document as well as from the screen and project it on the screen. The basic concept is to read the alpha-numeric characters and display them onto the screen. This will be done by performing complex image processing. Then these recognized characters will be transferred to the laptop. The connection between a laptop and a mobile device would be done using the 802.11 (Wi-Fi) technology. This would save a lot of time that a user wastes on typing that document. Thus, a generalized and simple solution for implementing the Optical Character Recognition(OCR) in an Android device.

Keyword : - *Image processing, Wi-Fi, Alpha-numeric, Image, Mobile, Android, Laptop.*

1.INTRODUCTION

ScanIT is an android app basically designed for implementing the OCR technology through android platform .Till date a lot of work has been done in the field of Image Processing thus, obviously in its sub-branch i.e. OCR. But till today emphasis has been given only to extraction of text from the image. Many different algorithms and libraries have been designed for the purpose of performing OCR on images. Many of them are also available on online OCR engines. But there are only a few of the OCR engines or Libraries that are accurate .And even those which are the most accurate give only 60% of accuracy. And even with that accuracy only the extracted text is kept as it is and is not processed further.

In this paper, an Android application known as ScanIT will be designed to utilize the OCR technology at its maximum potential. This will be done by using a Optical Character Recognition(OCR) library "Tesseract " provided by the Google itself. But only this is not at all enough for improving the accuracy. In order for the OCR app ScanIT to be more accurate the taken image should be processed before taking it to the actual OCR library for extracting text. For the Tesseract Library to extract text more accurately from the image the image has to be processed first. This includes the image enhancement, adjusting contrast, adjusting monochrome so as to make the image text more readable and more exact. This would be the improvement in the extraction of text. Now after the text has been extracted, normally it is kept as it is. Its not used further anymore, but the ScanIT app will not only extract the text but also manipulate the text so that we can perform operations like searching a specific word. Now other part of the app includes sending this extracted text on to the laptop using Wi-Fi and thus creating a document of the picture taken .Thus, ScanIt will be saving much of the time that we spend on typing the document i.e making a soft copy from the hard copy.

2.LITERATURE SURVEY

A lot of work has been done in the field of Optical Character Recognition (OCR) technology. In Recent days the applications of the OCR technology has been increased to a very vast scope. From previous research, we take three paper to make it as base of this research. The first research is Book Search by Capturing Text from Digital Images Using Optical Character Recognition. This paper shows how to use MATLAB and its image processing toolbox functions in order to recognize characters in an image. They implement this using MATLAB for segmentation using edge detection, identification of characters, and storing the vector of characters.

Optical Character Recognition (OCR) service enables application to retrieve the text that appears in a photograph. They have to first pre-process the image and image extraction as followed to find characters in a photograph. The resulting vector can be used in many applications. They wish to develop this process for developing an application for searching a book, based on the characters recognized in the input image (usually the cover page of the book) . The last process of this application is letting the customers pick the information from search engine by them self. Besides research on book search application, we also look for a research in OCR used on Android platform device. This research let us know the use of Tesseract-OCR on Android platform and what we get from it. This research make an experiment for four different languages and make a conclusion that Tesseract-OCR can recognize the text and have a high accuracy . There is more research on OCR library for Android platform. research make a library for OCR on Android that called OCRdroid which can handle blur, shading, tilted and any other picture condition. We use this research for a
Comparison to Tesseract-OCR.

The second research is “Scene Text Extraction using Stroke Width Transform for Tourist Translator on Android Platform”. In this paper, considering few of challenges, real time application named as TravelMate is designed and developed. Extraction is performed using stroke width transform and connected component based approach. Proposed application assists the tourists, while they are roaming in foreign countries. The performance of system is tested based on extraction rate. With proposed application almost all the text in horizontal orientation extracted correctly, whereas performance of real time images varies with lighting condition and camera resolution. Proposed android application can be further extended to deal with any target and source language for translation. It can be further modified to deal with text having vertical or arbitrary orientation.

The third research is ,”anyOCR: A Sequence Learning Based OCR System for Unlabeled Historical Documents” .In this research the input that was to be given by the user is reduced which is very helpful when there is a very little training data or difficulties arise with the images or characters of old handwritten or historic scripts.

3. OPTICAL CHARACTER RECOGNITION(OCR)

Optical character recognition (also optical character reader, OCR) is the mechanical or electronic conversion of images of typed, handwritten or printed text into machine-encoded text, whether from a scanned document, a photo of a document, a scene-photo (for example the text on signs and billboards in a landscape photo) or from subtitle text superimposed on an image (for example from a television broadcast). It is widely used as a form of information entry from printed paper data records, whether passport documents, invoices, bank statements, computerised receipts, business cards, mail, printouts of static-data, or any suitable documentation. It is a common method of digitising printed texts so that they can be electronically edited, searched, stored more compactly, displayed on-line, and used in machine processes such as cognitive computing, machine translation, (extracted) text-to-speech, key data and text mining. OCR is a field of research in pattern recognition, artificial intelligence and computer vision.

Early versions needed to be trained with images of each character, and worked on one font at a time. Advanced systems capable of producing a high degree of recognition accuracy for most fonts are now common, and with support for a variety of digital image file format inputs. Some systems are capable of reproducing formatted output that closely approximates the original page including images, columns, and other non-textual components. Special fonts like OCR-A, OCR-B, or MICR fonts, with precisely specified sizing, spacing, and distinctive character shapes, allow a higher accuracy rate during transcription. These were often used in early matrix-matching systems.

In recent years, the major OCR technology providers began to tweak OCR systems to better deal with specific types of input. Beyond an application-specific lexicon, better performance can be had by taking into account business rules, standard expression, or rich information contained in colour images. This strategy is called "Application-Oriented OCR" or "Customised OCR", and has been applied to OCR of license plates , invoices, screenshots, ID cards, driver licenses, and automobile manufacturing.

"Comb fields" are pre-printed boxes that encourage humans to write more legibly – one glyph per box. These are often printed in a "dropout color" which can be easily removed by the OCR system. Palm OS used a special set of glyphs, known as "Graffiti" which are similar to printed English characters but simplified or modified for easier recognition on the platform's computationally limited hardware. Users would need to learn how to write these special glyphs. Zone-based OCR restricts the image to a specific part of a document. This is often referred to as "Template OCR".

4. CURRENT OCR METHODOLOGIES USING DIFFERENT APIS

	Google Cloud Vision API	Microsoft Vision API	FreeOCR API
Total Images Processed	3001	3001	3001
Extracted data	2762	1328	1531
Correctly Extracted	1839	463	542
Incorrectly Extraction	923	865	989
Not Extracted anything	239	1673	1470
Recall %	92%	44%	51%
Precision %	66%	34%	35%

The above table illustrates the present popular Optical Character Recognition(OCR) APIs . These currently present APIs have produced different results in different light conditions when tested. Here among the three APIs compared i.e. Google Cloud Vision API, Microsoft Vision API, FreeOCR API. The Google Cloud Vision API has proved to be producing the best result in the tests with 92% recall and 66% precision . Even being the best among the present APIs the Google Cloud Vision API proved ineffective and useless when there was no Internet Connection. Thus, it proved to be the drawback of this best API . Though being the best, it comes with a price . That is why this paid version has a very limited use and is affordable only by big companies. This is where the best failed ,so comes into picture the open sourced “Tesseract Library” . This is a offline library provided under apache license and over all the benefits its free.

5.TESSERACT LIBRARY

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. In 2006 Tesseract was considered one of the most accurate open-source OCR engines then available.

The Tesseract engine was originally developed as proprietary software at Hewlett Packard labs in Bristol, England and Greeley, Colorado between 1985 and 1994, with some more changes made in 1996 to port to Windows, and some migration from C to C++ in 1998. A lot of the code was written in C, and then some more was written in C++. Since then all the code has been converted to at least compile with a C++ compiler.[3] Very little work was done in the following decade. It was then released as open source in 2005 by Hewlett Packard and the University of Nevada, Las Vegas (UNLV). Tesseract development has been sponsored by Google since 2006.

Tesseract was in the top three OCR engines in terms of character accuracy in 1995. It is available for Linux, Windows and Mac OS X. However, due to limited resources it is only rigorously tested by developers under Windows and Ubuntu.

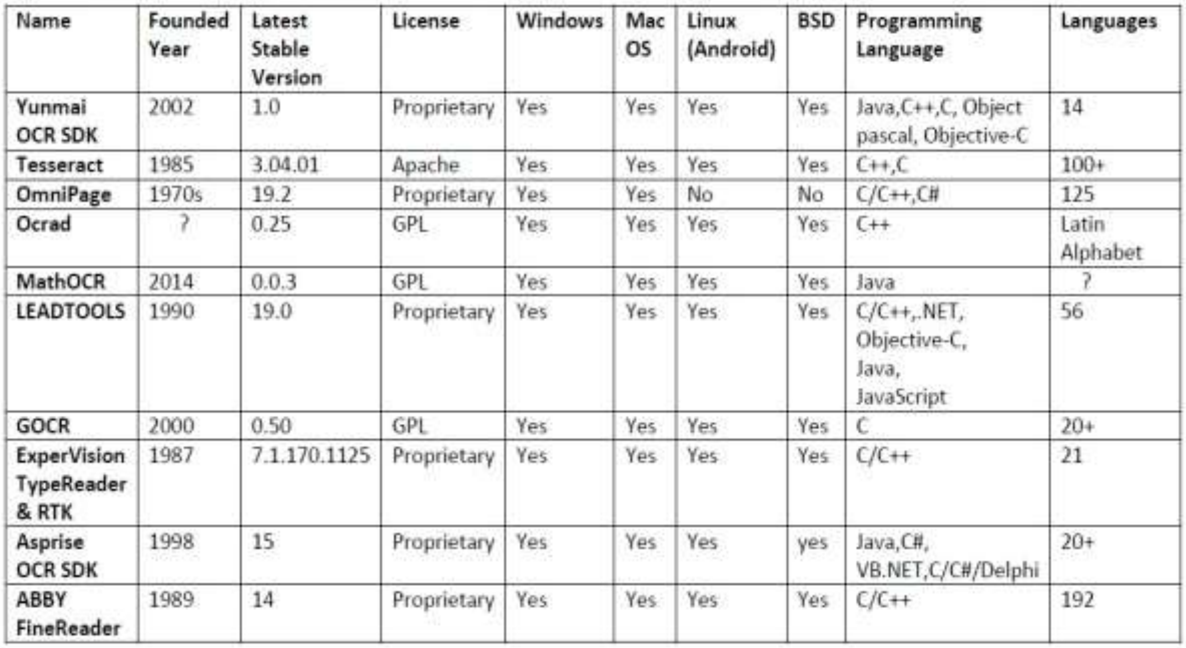
Tesseract up to and including version 2 could only accept TIFF images of simple one-column text as inputs. These early versions did not include layout analysis, and so inputting multi-columned text, images, or equations produced garbled output. Since version 3.00 Tesseract has supported output text formatting, hOCR positional information and page-layout analysis. Support for a number of new image formats was added using the word.Leptonica library. Tesseract can detect whether text is monospaced or proportionally spaced.

The Tesseract library functions in two main stages in the system. These are as follows:

1) Recognizing a Text:-
On the onPhotoTaken activity, firtsly, Tesseract library will stabilize the final picture orientation. After that, it takes the width and the height of the final picture then converts it to ARGB_8888 for being processed by Tesseract function. The detection have done by Tesseract by calling baseApi.getUTF8Text() function. Tesseract recognize all of the alphabet and number that contain the word except special character. Then Tesseract will replace the detection result with a alphabet and number that depend on the language in trained data.

2) Separating a Text:-
Separating a text have a purpose if there is/are any word that can't recognize clearly, the title still can be compared with data in database using another word contain on the editable text title. This process is done by separating the word if system finds a space between them.

6. COMPARISON OF TESSERACT WITH OTHER LIBRARIES



Name	Founded Year	Latest Stable Version	License	Windows	Mac OS	Linux (Android)	BSD	Programming Language	Languages
Yunmai OCR SDK	2002	1.0	Proprietary	Yes	Yes	Yes	Yes	Java,C++,C, Object pascal, Objective-C	14
Tesseract	1985	3.04.01	Apache	Yes	Yes	Yes	Yes	C++,C	100+
OmniPage	1970s	19.2	Proprietary	Yes	Yes	No	No	C/C++,C#	125
Ocrad	?	0.25	GPL	Yes	Yes	Yes	Yes	C++	Latin Alphabet
MathOCR	2014	0.0.3	GPL	Yes	Yes	Yes	Yes	Java	?
LEADTOOLS	1990	19.0	Proprietary	Yes	Yes	Yes	Yes	C/C++,.NET, Objective-C, Java, JavaScript	56
GOCR	2000	0.50	GPL	Yes	Yes	Yes	Yes	C	20+
ExperVision TypeReader & RTK	1987	7.1.170.1125	Proprietary	Yes	Yes	Yes	Yes	C/C++	21
Asprise OCR SDK	1998	15	Proprietary	Yes	Yes	Yes	yes	Java,C#, VB.NET,C/C#/Delphi	20+
ABBY FineReader	1989	14	Proprietary	Yes	Yes	Yes	Yes	C/C++	192

7. PROPOSED SYSTEM

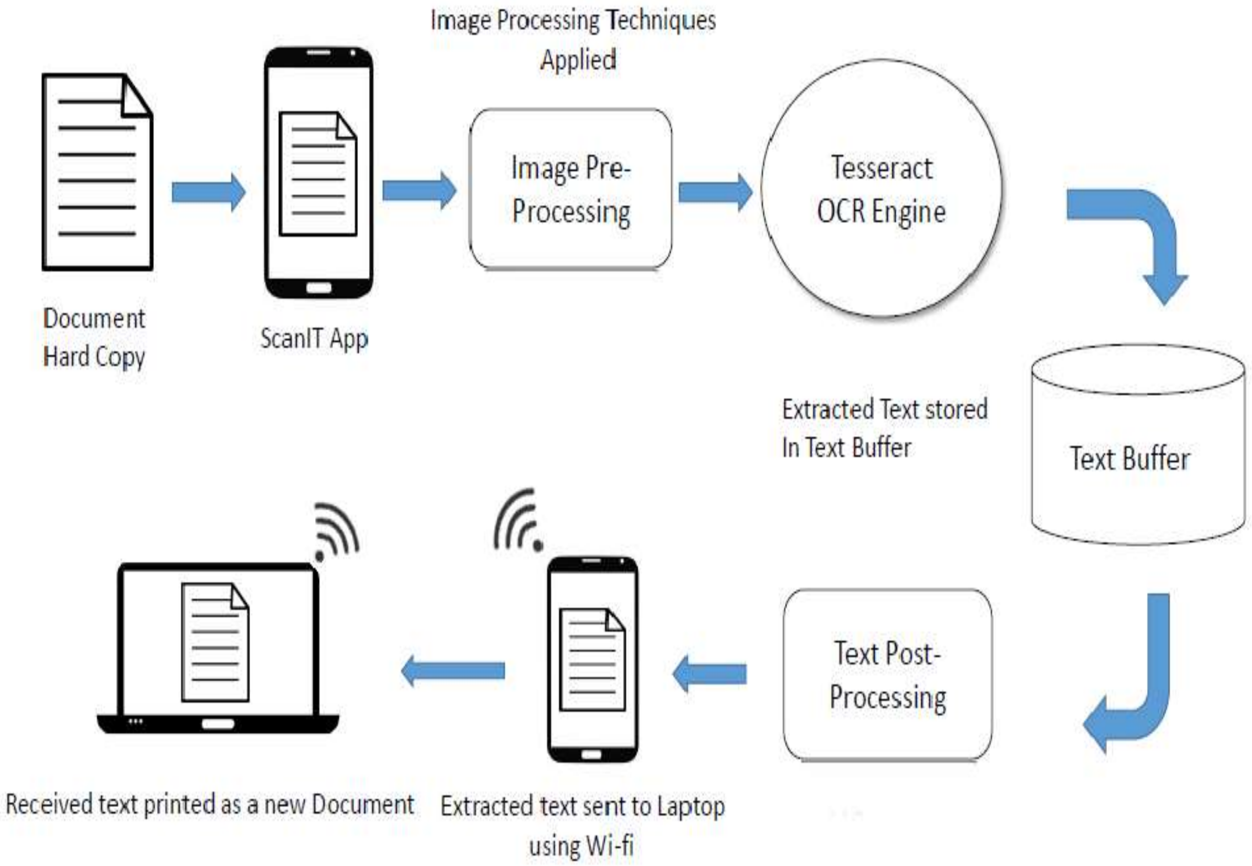


Image Processing Techniques Applied
Document Hard Copy
ScanIT App
Image Pre-Processing
Tesseract OCR Engine
Extracted Text stored In Text Buffer
Text Buffer
Text Post-Processing
Received text printed as a new Document
Extracted text sent to Laptop using Wi-fi

The above Diagram illustrates the proposed System regarding the overall functionality of the ScanIT App Respectively. In this system the targeted document is scanned through the app. Then the scanned document is firstly cropped as per the user then the resultant image is processed using Image processing techniques, After the completion of this stage the image is then applied as a input0 of the ScanIT OCR Engine which extracts the text for the document efficiently with maximum accuracy. Now the Extracted Text is stored in a temporary storage named Text Buffer Storage respectively. After this stage, the text processing is done on the resultant text to increase the accuracy of the overall result. When the required text is obtained the last result is transferred to the User's laptop through Wi-Fi technology. At last when the text is received from the mobile onto the laptop the text is printed on a newly created Document and hence the softcopy of the text in the captured image is created.

8. CONCLUSION

So ,the conclusion that can be drawn from this paper is that the Optical Character Recognition (OCR) technology can be implemented in an Android device efficiently using the Tesseract Library. Various tests of the tesseract library have been conducted and it has produced very much optimal results. The succes rate of testing is 100% for light condition, 80% for position condition, and 85% for noise condition. And further using the app to create a soft copy of the document on the laptop would prove to be very time saving and cost effective solution to the problem of typing a specific document and wasting many hours in it. Thus, a generalized solution for the implementation of Optical Character Recognition (OCR) in a free and efficient manner is provided in this paper.

9. REFERENCES

1) Dr S.Vasavi, Srikanth Varma.Ch, Anil kumar.Ch, Santosh.D.M, SaiRam.S (2014). Book Search by Capturing Text from Digital Images Using Optical Character Recognition, International Journal of Computer Science and Information Technologies, Vol. 5 .

2) Nana Ramadijanti, Achmad Basuki, Agrippina G.J.W, ”Designing Mobile Application for Retrieving Book Information using Optical Character Recognition”, 2016 Knowledge Creation and Intelligent Computing (KCIC)

3) Dr. Archana Ghotkar , Miss. Pooja Chavre “Scene Text Extraction using Stroke Width Transform for Tourist Translator on Android Platform” 2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT) International Institute of Information Technology (I^{2}IT), Pune.
4) Martin Jenckel, Syed Saqib Bukhari,”anyOCR: A Sequence Learning Based OCR System for Unlabeled Historical Documents” 2016 23rd
5) "The History of OCR". Data processing magazine. **12**: 46. 1970.