Optical Character Recognition on Handheld Devices

Sravan Ch
Student
Amity School of Engineering and Technology
Sector 125, Noida

ShivankuMahna
Student
Amity School of Engineering and Technology
Sector 125, Noida

NirbhayKashyap
Assistant Professor
Amity School of Engineering and Technology
Sector 125, Noida

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 how OCR is more efficient and easier alternative to scanning a document using a scanner as the image captured using OCR is of exactly the same quality like its scanned copy, the only difference being that OCR is done with the help of a simple mobile phone camera whereas scanning is done using a bulky scanner. It then also explains the problems being faced by the developers in using OCR as a technology on a large scale and how that problem can be dealt with. The proposed OCR system provides many features that require no typing, editing raw data, quick translation, and memory utilization.In the end it also highlights the major emerging trends in the field of OCR and how OCR as a technology is evolving with every passing day.

Keywords:
Optical *Character Recognition System (OCR), Camera Captured Document Images, Handheld Device, Image Segmentation*

General Terms
Optical character recognition, Pattern Recognition, ImageSegmentation,Text Extraction, Tesseract.

1. INTRODUCTION
A person is able to see images because of the communication between our eyes and brain. Our eyes act as an optical mechanism and the images seen by our eyes are an input for our brain and the ability to understand visualise these images varies from person to person. Similarly we have the technology known as OCR, where OCR stands for Optical Character Recognition, which through its automated mechanism allows easier recognition of character and its processing.

Earlier scanners were the only working OCR application available in the market. The main disadvantage of scanners was that it was not portable and it takes a lot of time to capture an image.

But with today's devices having better processing speeds, larger internal memory and an excellent back camera, researchers have dared to think of running OCR applications on devices such as smart phones for having real time imaging results.Applications such as Cam Scanner and google translate are the prime examples of Optical character Recognition application. It also showcases the fact that this OCR 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.

2. HOW OCR WORKS
A. OCR
OCR allows for automatically recognizing characters through an optical mechanism. It is capable of recognizing both handwritten and printed text. Its performance can be judged based on the quality of the documents and the camera being used to capture the raw image. OCR system is so designed that it processes images with contain more text with very less number of graphic elements.

As mentioned before, most of the character recognition programs and algorithms will be working efficiently only on the images which are captured using a scanner or a digital camera and run on a computer software. But since the size and portability were the factors which were hampering further growth and usability of this technology, in order to overcome the above mentioned limitations, a character recognition system based on android devices is proposed. [1]

OCR works in Android mobile operating system by combining Google's open-source OCR engine, Tesseract and the text recognition OCR engine.[3] The text-to-speechsynthesizers in a mobile device allows users to take photographs of text using the camera in a mobile phone and have the text read aloud by the mobile phone.

OCR as a technology that enables us 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 is that the images captured by a digital camera differ from 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 done on the image to improve the accuracy of text recognition

How Our OCR WORKS

The proposed OCR system provides many features that require no typing, editing raw data, quick translation, and memory utilization.

Tesseract is chosen as the engine of the OCR because of its widespread approbation, extensibility and flexibility, its community of active developers, and the fact that it works out of the box. To perform the character recognition, our application has to go through two important steps which are as follows:-

1. Segmentation, i.e., given a binary input image, to identify the individual glyphs (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. [2]

B. TERSSERACT

Tesseract is an open source engine for optical character recognition. It is available on many operating systems. It is one of the most accurate OCR engine available. It can read and convert into over 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 is available at http://code.google.com/p/tesseract-ocr.[3]

a) Architecture of Tesseract

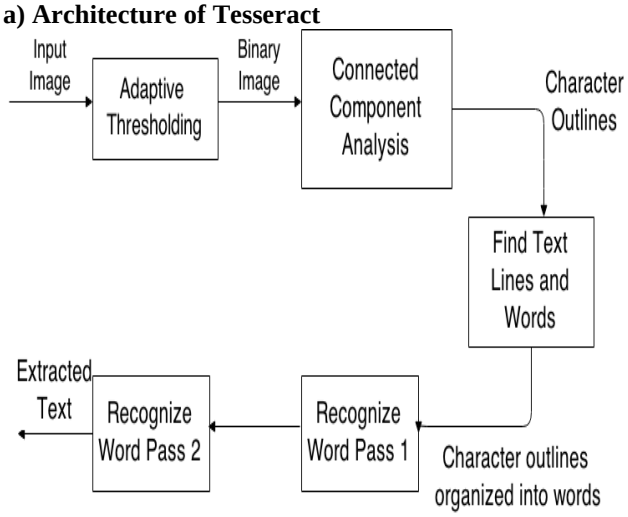

Figure 1 Architecture of Tesseract

Tesseract converts the input image into binary format using thresholding.Outlines of components are stored on connected Component Analysis. Nesting of outlines is done which gathers the outlines together to form a Blob.Text lines are analyzed for fixed pitch and proportional text. Then the lines are broken into words by analysis according to the character spacing. Fixed pitch ischopped in character cells and proportional text is broken into words by definite spaces and fuzzy spaces.

Tesseract recognises 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 recognised in the first pass are recognized again through a run over through the page. Finally Tesseract resolves fuzzy spaces. To locate small and capital text, Tesseract checks alternative hypothesis for x-height. [4]

3. CONSTRUCTION OF OCR

The Fig.2 given below is illustrates the overallfunctioning of Optical Character Recognition (OCR). The input image can be any document, live text, journals, magazines etc. The functioning of OCR contains the following steps: scanning, segmentation, pre-processing, feature extraction,recognition.[1]The input is first scanned 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 is extracted using feature extraction to finally recognise the text.

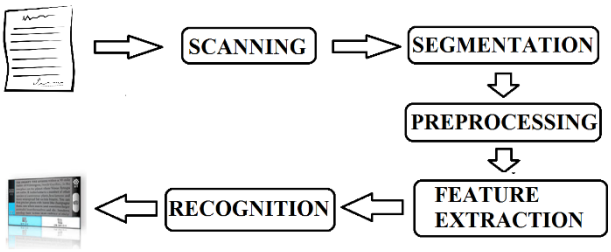

Figure 2 Construction of OCR

a) Scanning

Android mobile camera is used to capture the image of document. This process is called scanning.

This is nothingbut the process of scanning which converts the document into digital image. The digital image is then converted into a grayscale image using thresholding function.Thresholding is the process which converts multi levelimage 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 it easier to detect the text regions in an image. It also saves a lot of memory space and processing time.

b) Segmentation

Regions of text is detected using the process of segmentation. It differentiates the text from other graphical elements in the document. Splits and joints can cause confusion between text and graphic elements in the document resulting in incorrect segmentation of the text.[1] This generally occurs due to poor scanning which increases the noise in the digital document. Joints in characters occurs when the document is scanned at low threshold and splits occurs when the document is scanned at high threshold.

c) Pre-processing

During scanning stage, 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 done on the image using filling and thinning techniques. Normalization is responsibleto handle uniform size, slant and skew correction.

d) Feature Extraction

Feature extraction refers to the extraction of features 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 present in the character. Spaces, lines, intersections etc are some of the abstract features. Feature extraction is done using Tesseract algorithm. Tesseractalgorithm is used to implement feature extraction.

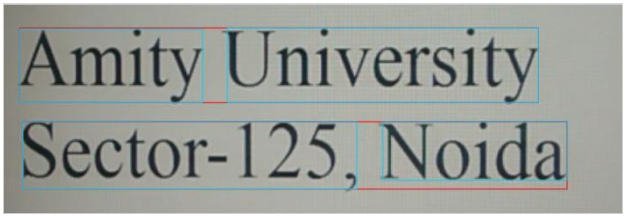

Figure 3 Text Regions extracted from the image

e) Recognition

OCR system uses Tesseract algorithm to identify characters from the image foreground pixels also called as blobs and recognizes the lines. These lines are then recognized into words or characters.In this phase the image is converted into character stream which represents letters. [5]

Flow Chart for OCR system

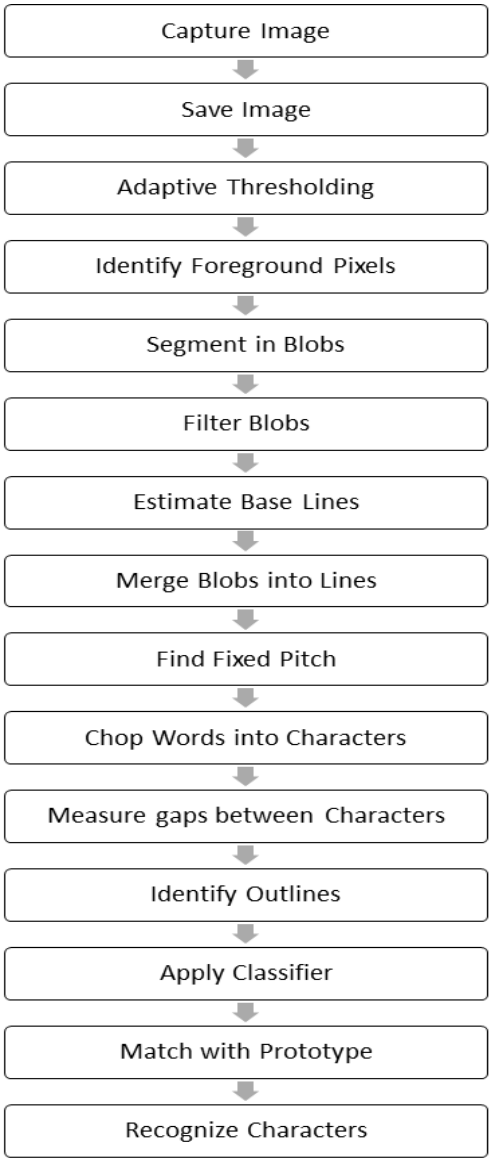

Figure 4 Flow Chart for OCR system

4. ISSUES ENCOUNTERED WHILE USING OCR ON HANDHELD DEVICES

However, computing images captured by handheld devices involves a number of challenges which are as follows :-

1. Images acquired by handheld device very often suffer from skew and perspective distortion, that is, the image captured is not as straight as the image captured by lets say a scanner.
2. Since the capturing process is manual, image can sometimes be subjected to uneven and insufficient illumination which in turn makes certain parts of the image abnormally brighter than the rest of the image.
3. Unavailability of technologies like auto focus, that helps in setting the focus at the correct point, in every smart phone yields poor quality images.
4. Though the processing speed and memory size of handheld devices have improved in recent times but it is still not sufficient enough to run desktop based OCR algorithms as they are computationally expensive and require very high amount of memory.
5. In addition to the above challenges, smart phones also do not have a Floating Point Unit (FPU) which is required for floating point arithmetic operations. However, floating point operations can be performed on such devices by using floating point emulators but that results in slower operation.

Therefore, the need of the hour is to design computationally efficient and light-weight OCR algorithms that can run smoothly on handheld mobile devices.

5. ACCURACY

Accuracy of a OCR system depends onthe quality of input document. Sometimes the output from OCR systems is often quite "noisy". Post processing is done on the text to correct the noise. The average time taken to recognize 20 words is 350ms and that of 100 words is 500ms.

The accuracy of the OCR system also depends on the camera used to capture the raw image of the document. Various factors affecting the quality are: Focus of the camera, resolution of the picture, amount of noise present etc. Tesseract engine achieved an average accuracy of 93%.

Toestimate the OCR accuracy, the OCR output can then be compared to the text of the original document,called the ground truth. [6]

6. CONCLUSION

This paper tells about Optical character recognition for handheld devices in recognizing characters in offline mode. The system has the ability to recognize characters with accuracy exceeding 90% mark. The advantage of the system is that it is easily portable and its scalability which can recognize various languages and also 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, theaccuracy can be increased.The future scope is to develop a software for automatic editing and searching.

7. ACKNOWLEDGMENTS

The authors would like to thank Department of Computer Engineering, Amity School of Engineering and Technology and indebted to our guide Prof.NirbhayKashyap for his guidance which helped us design this paper.

8. 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] "α-Soft: An English Language OCR", 2010 Second InternationalConference on Computer Engineering and Applications. Junaid Tariq,Umar Nauman Muhammad UmairNaru.

[6] A survey of modern optical character recognition techniques (DRAFT), February 2004