International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-2 Issue-1, March 2013

Optical Character Recognition

Ravina Mithe, Supriya Indalkar, Nilam Divekar

Abstract: ***The Optical Character Recognition is a mobile application. It uses smart mobile phones of android platform. This paper combines the functionality of Optical Character Recognition and speech synthesizer. The objective is to develop user friendly application which performs image to speech conversion system using android phones. The OCR takes image as the input, gets text from that image and then converts it into speech. This system can be useful in various applications like banking, legal industry, other industries, and home and office automation. It mainly designed for people who are unable to read any type of text documents. In this paper, the character recognition method is presented by using OCR technology and android phone with higher quality camera.***

Index Terms: Binerization, Optical Character Recognition, Pattern Matching, Segmentation, Tesseract, Text Extraction,.

I. INTRODUCTION

A. OCR

OCR is the acronym for Optical Character Recognition. This technology allows to automatically recognizing characters through an optical mechanism. In case of human beings, our eyes are optical mechanism. The image seen by eyes is input for brain. The ability to understand these inputs varies in each person according to many factors [2]. OCR is a technology that functions like human ability of reading. Although OCR is not able to compete with human reading capabilities.

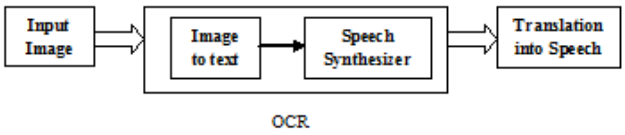

Fig.1.1 Block Diagram of OCR

OCR can recognize both handwritten and printed text. But the performance of OCR is directly dependant on quality of input documents.OCR is designed to process images that consist almost entirely of text, with very little non-text clutter obtain from picture captured by mobile camera. This application is for the Android mobile operating system that combines Google's open-source OCR engine, Tesseract, text recognition OCR engine [5]. Google's language translation service, and the Android operating system's text-to-speech synthesizer to allow users to take photographs of text using a camera phone and have the text read aloud.

Most of the character recognition program will be recognized through the input image with a scanner or a digital camera and computer software. There is a problem in the spatial size of the computer and scanner. If you do not have a scanner and a digital camera, a hardware problem occurs. In order to overcome the limitations of computer occupying a large space, character recognition system based on android phone is proposed [4].

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. Images captured by a digital camera differ from scanned documents or image. 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. We have chosen Tesseract because of widespread approbation, its extensibility and flexibility, its community of active developers, and the fact that it "just works" out of the box. To perform the character recognition, our application has to go through three important steps. The first is Segmentation, i.e., given a binary input image, to identify the individual glyphs (basic units representing one or more characters, usually contiguous). The second step is feature extraction, i.e., to compute from each glyph a vector of numbers that will serve as input features for an ANN [3]. This step is the most difficult in the sense that there is no obvious way to obtain these features. The final task is classification.

B. TERSSERACT

Tesseract is Open source OCR engine. It was firstly developed between 1984 to 1994 at HP. In 1995, it was sent to UNLV for Annual Test of OCR Accuracy after the joint project between HP Labs Bristol and HP's Scanner Division in Colorado. Finally in 2005, Tesseract was released as open source by HP and available at http://code.google.com/p/tesseract-ocr [2].

a) Architecture of Tesseract

Tesseract works with independently developed Page Layout Analysis Technology. Hence Tesseract accepts input image as a binary image. Tesseract can handle both, the traditional-Black on White text and also inverse-White on Black text.

Outlines of component are stored on connected Component Analysis. Nesting of outlines is done which gathers the outlines together to form a Blob. Such Blobs are organized into text lines. 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 is chopped in character cells and proportional text is broken into words by definite spaces and fuzzy spaces.

Revised Manuscript Received on 30 March 2013.
* Correspondence Author
Ms. Ravina Mithe, Computer Science Department, JSPM's BSIOTR (W), Pune, India.
Ms. Supriya Indalkar, Computer Science Department, JSPM's BSIOTR (W), Pune, India.
Ms. Nilam Divekar, Computer Science Department, JSPM's BSIOTR (W), Pune, India.

Tesseract performs activity to recognize words. This recognition activity is mainly consists of two passes. The first pass tries to recognize the words. Then satisfactory word is passed to Adaptive Classifier as training data, which recognizes the text more accurately. During second pass, the words which were not recognized well in first pass are recognized again through run over the page. Finally Tesseract resolves fuzzy spaces. To locate small and capital text Tesseract checks alternative hypothesis for x-height [1].

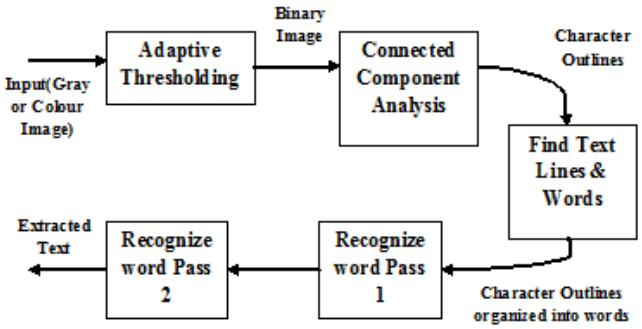

Fig.1.2.1 Architecture of Tesseract

C. TEXT TO SPEECH

A text to speech (TTS) synthesizer is a system that can read text aloud automatically, which is extracted from Optical Character Recognition (OCR). A speech synthesizer can be implemented by both hardware and software. Speech synthesis is the artificial production of human speech [9]. A computer system used for this purpose is called a speech synthesizer. A text-to-speech (TTS) system converts normal language text into speech. A synthesizer can incorporate a model of the vocal tract and other human voice characteristics to create a completely "synthetic" voice output.

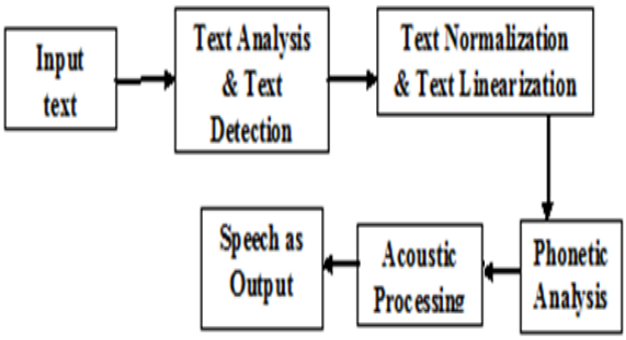

Fig.1.3 Text to Speech Conversion

a) Text Analysis & Detection:

The Text Analysis is part of preprocessing. It analyzes the input text and organizes into manageable list of words. Then it transforms them into full text. Text detection localizes the text areas from printed documents.

b) Text Normalization & Linearization:

Text Normalization is the transformation of text to pronounceable form. Text normalization is often performed before text is processed in some way, such as generating synthesized speech or automated language translation. The main objective of this process is to identify punctuation marks and pauses between words. Usually the text normalization process is done for converting all letters of lowercase or upper case, to remove punctuations, accent marks , stopwords or "too common words "and other diacritics from letters.

c) Phonetic Analysis

It provides phonetic alphabets. The grapheme to phoneme conversion is done. It is actually a conversion of orthographical symbols into phonological symbols.

d) Acoustic Processing.

It performs formant synthesis. It works intelligently and thus does not require any kind of database of speech samples. For speak out the text, it uses voice characteristics of a person.

II. PREVIOUS SYSTEM

OCR engines converts images of machine-printed characters into machine-readable characters. Images of machine-printed characters are extracted from a bitmap. Forms can be scanned through an imaging scanner, faxed, or computer generated to produce the bitmap. There are two separate methods in previous OCR system: First text extraction and speech translation. In a typical OCR systems input characters are digitized by an optical scanner. Each character is then located and segmented, and the resulting character image is fed into a preprocessor for noise reduction and normalization. Certain characteristics are the extracted from the character for classification. The feature extraction is critical and many different techniques exist, each having its strengths and weaknesses. After classification the identified characters are grouped to reconstruct the original symbol strings, and context may then be applied to detect and correct errors.

III. PROPOSED SYSTEM

OCR technology allows the conversion of scanned images of printed text into text or information that can be understood or edited using android mobile phones. Optical character recognition, usually abbreviated to **OCR**, is the mechanical or electronic translation of scanned images of handwritten, typewritten, or printed text into machine-encoded text. Open source OCR software called Tesseract as a basis for Optical Recognition project, which is considered as the most accurate free OCR engine in existence.

OCR technology uses three steps- Scanning acquisition of printed documents as optical images. Recognition- involves converting these images to character streams representing letters of recognized words and the final element involves accessing or storing the converted text. Converted text is referred as extracted text. When, the user begins by capturing an image containing text of interest using the Mobile camera. The specified area of the image is processed on the device in order to optimize it for transfer and input to the OCR. Speech synthesizer is used to convert extracted text into the voice. Firstly it analyses text, transforms text into pronounceable form. Speech synthesizer performs conversion of grapheme to phoneme form and uses voice characteristics of a person.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

Most of the character recognition systems will be recognized through the input image with a scanner and computer software. There is a problem in the size of the computer and scanner, as computer and scanner requires large amount of space.

In order to overcome this problem of computer and scanner occupying a large space, optical character recognition (OCR) system based on android phone is proposed. Because the performances of smart phone and computer are different, the speed of character recognition is slow. In this paper, the character recognition method is presented by using OCR technology and android phone with higher quality camera.

IV. APPLICATIONS

A. Banking

One widely known application is in banking, where OCR is used to process checks without human involvement. A image of check can be captured by mobile camera, the writing on it is scanned instantly, and the correct amount of money is transferred. This technology has nearly been perfected for printed checks, and is fairly accurate for handwritten checks as well, though it occasionally requires manual confirmation. Overall, this reduces wait times in many banks.

B. Legal Industry

In the legal industry, there has also been a significant movement to digitize paper documents. In order to save space and eliminate the need to sift through boxes of paper files, documents are being scanned. OCR further simplifies the process by making documents text-searchable, so that they are easier to locate and work with once in the database. Legal professionals now have fast, easy access to a huge library of documents in electronic format, which they can find simply by typing in a few keywords.

C. Other Industries

OCR is widely used in many other fields, including education, finance, and government agencies. OCR has made countless texts available online, saving money for students and allowing knowledge to be shared.

D. Vocal Monitoring

In some cases, oral information is more efficient than written messages. The appeal is stronger, while the attention may still focus on other visual sources of information. Hence the idea of incorporating speech from documents.

V. METHODOLOGY

A. Components of OCR

The Fig.4.1 given below is illustrates the overall functioning of Optical Character Recognition (OCR). It contains some steps to recognize text. These steps are: scanning, segmentation, preprocessing, feature extraction, recognition. Here the input image to OCR is any hand written or printed text like books, gerenals, magazines, news papers, etc. Such input is given to OCR. Firstly it is scanned using android mobile camera. It means it digitizes the analog document. Text regions within image are located, extracts symbols through segmentation, preprocess each symbol, extracts feature, and recognise them.

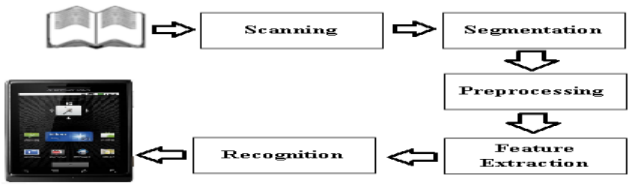

Fig. 5.1 Optical Character Recognition (OCR)

a) Scanning

This application i.e. OCR system uses Android mobile camera. Camera captures image of document. This is nothing but the process of scanning. In short we can say that scanning makes original document as digital image. Generally, original douments are made up of the black coloured text print on the white coloured background. Scanning comes with thresholding which makes the digital image as gray scale image. Thresholding is the process which converts multi level image into bilevel image i.e. black and white image.[4] Fixed threshold level is defined in thresholding. If the gray levels are below the threshold level, identified as black.Whereas if gray level is above the threshold level, identified as white. This results insaving memory space and computational efforts.[4]

b) Segmentation

The process of locating regions of printed or handwritten text is segmentation. Segmentation differs text from figures and graphics. When segmentation is applied to text, it isolates characters or words. The mostly occurred problem in segmentation is: it cuases confusion between text and graphics in case of joined and split characters. Usually, splits and joints in the characters causes due to scanning. If document is dark photocopy or if it scanned at low threshold, joints in characters will occur. And splits in charcters will occur if document is light photocopy or scanned at high threshold.[4] OCR system also gets confused during segmentation when characters are connected to graphics.

c) Preprocessing

As we seen above, some noise may occurred during scanning process. This results in poor recognition of characters. This usually occurred problem is overcomed by preprocessing. It consists of smoothing and normalization. In smoothing, certain rules are applied to the contents of image with the help of filling and thining techniques. Normalization is responsible to handle uniform size, slant and rotation of characters.

d) Feature Extraction

It extracts the features of symbols. Features are the characteristics. In this, symbols are characterized and unimportant attributes are left out. The feature extraction technique does not match concrete character patterns, but rather makes note of abstract features present in a character such as intersections, open spaces, lines, etc. [7] Tesseract algorithm is used to implement feature extraction. Feature extraction is concerned with the representation of the symbols. The character image is mapped to a higher level by extracting special characteristics of the image in the feature extraction phase.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

e) Recognition

OCR system works with Tesseract algorithm which recognizes characters. Tesseract identifies characters in foreground pixels, called as blobs, and then it finds lines. Word by word recognition of characters is done throughout the lines. Recognition involves converting these images to character streams representing letters of recognized words [8]. In short, recognition extracts text from images of documents.

B. FLOW CHART

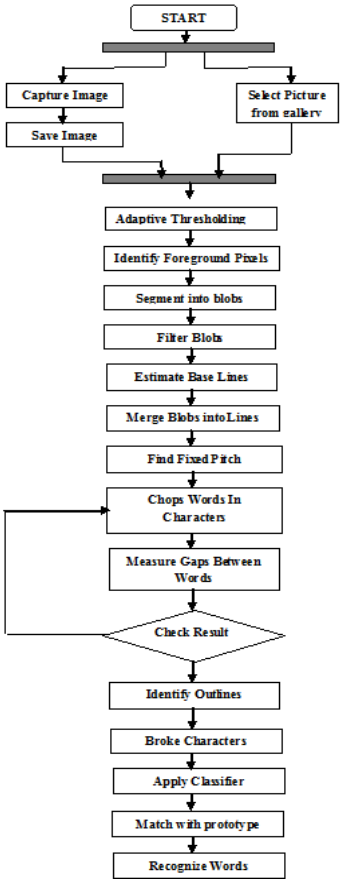

Fig.5.2 Flowchart of image to text

VI. CONCLUSION

The Optical Character Recognition deals with recognition of optically processed characters. Reliably interpreting text from real-world photos is a challenging problem due to variations in environmental factors even it becomes easier using the best open source OCR engine.

A. ADVANTAGES

1. This approach uses Android Operating System.
2. Android is free and open software. Thus OCR mobile application is significantly of lower cost.
3. Early OCR requires expensive scanners and special purpose electronic or optical hardware, but this approach is a mobile application having inbuilt camera.
4. Converts raster image to text and text to voice using OCR techniques
5. As Android based on Linux kernel it has feature of safety from virus infection.

B. LIMITATIONS

1. Accuracy of an OCR system is directly dependent on the quality of input document.
2. The output from OCR systems is often quite “noisy” and garbled. In order to correct this, the application will perform some post processing on the text after it has received a response from the OCR package.

C. FUTURE SCOPE

Our next works with OCR Mobile Application will include the improvement of the results by the use of table boundaries detection techniques and the use of text post-processing techniques to detect the noise and to correct bad-recognized words. OCR application will also display the signatures and the other symbols as it is in the document. It will also update its features including the translation of one language to another. So that it will helpful for people from other countries who can’t understand the local language.

APPENDIX

1. OCR- Optical Character Recognition
2. ANN-Artificial Neural Network
3. UNLV-University of Nevada Las Vegas.
4. TTS-Text To Speech

ACKNOWLEDGMENT

It is indeed a matter of great pleasure & privilege to publish this paper on “OPTICAL CHARACTER RECOGNITION” under the valuable guidance of Prof. Nirmal Chouhan and Prof. Anita Gunjal, We would like to express my deep sense of gratitude to my guides for this valuable guidance, advice, & constant aspiration to our work. F. A. Author thanks to Prof. Jayant Jadhav project coordinator & principal of Bhivarabai Sawant Institute of Technology & Research (For Women) for providing us constant support & facilities. F. A. Author thanks to our guides, who has helped us a lot or the completion for our project work.

REFERENCES

1. R. Smith. “An overview of the Tesseract OCR Engine.” Proc 9th Int. Conf. on Document Analysis and Recognition, IEEE, Curitiba, Brazil, Sep 2007, pp629-633.
2. The Tesseract open source OCR engine, http://code.google.com/p/tesseract-ocr.
3. R.W. Smith, The Extraction and Recognition of Text from
4. Multimedia Document Images, PhD Thesis, University of Bristol, November 1987.
5. 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.
6. Implementing Optical Character Recognition on the Android Operating System for Business Cards By Sonia Bhaskar, Nicholas Lavassar, Scott Green EE 368 Digital Image Processing.
7. Hybrid Page Layout Analysis via Tab-Stop Detection Ray Smith Google Inc. 1600 Amphitheatre Parkway, Mountain View, CA 94043, USA. theraysmith@gmail.com, 2009.
8. Optical Character Recognition Line Eikvil December 1993.
9. NLP Applications of Sinhala: TTS & OCR Ruvan Weerasinghe, Asanka Wasala, Dulip Herath and Viraj Welgama Language Technology Research Laboratory, University of Colombo School of Computing, 35, Reid Avenue, Colombo 00700, Sri Lanka.
10. Text To Speech: A Simple Tutorial D.Sasirekha, E.Chandra, March 2012.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)