ISSN (ONLINE): 2250-0758, ISSN (PRINT): 2394-6962

Volume-7, Issue-2, March-April 2017

International Journal of Engineering and Management Research

Page Number: 159-161

Optical Character Recognition (OCR)

Atul Pandey[1], Vivek Sharma[2], Shruti Paanchbhai[3], Neha Hedaoo[4], Prof. S.D. Zade[5]
[1,2,3,4]B.E Final Year Student, Department of Computer Science and Engineering, Priyadarshini Institute of Engineering and Technology, Nagpur, INDIA
[5]Assistant Professor, Department of Computer Science and Engineering, Priyadarshini Institute of Engineering and Technology, Nagpur, INDIA

ABSTRACT

This paper presets a review on the optical character recognition techniques. The optical character recognition is a mobile application. It uses smart mobiles phones of android platform. The objective is to develop user friendly application performs images to editable text. The OCR takes images as the input, and gives the plain text (editable text) as the output. Our proposed algorithm was based on Tesseract library. This system can be useful in various applications like banking, legal industry, other industries, print media and home and office automation. In this paper, OCR technology and android phone with higher quality camera is being used to recognize the characters.

Keywords— OCR, Tesseract, Character Recognition, camera, document, editable

I. INTRODUCTION

Optical character recognition (OCR) is that the mechanical or electronic conversion of pictures of written or written text into machine-encoded text. it's wide used as a kind of knowledge entry from written paper knowledge records, whether or not passport documents, invoices, bank statements, computerised receipts, business cards, mail, printouts of static-data, or any appropriate documentation. it's a standard methodology of digitizing written texts in order that it is electronically emended, searched, keep additional succinctly, displayed on-line, and utilized in machine processes like artificial intelligence, text-to-speech, key knowledge and text mining. OCR could be a field of analysis in pattern recognition, computer science and laptop vision.

Early versions required to be trained with pictures of every character, and worked on one font at a time. Advanced systems that have a high degree of recognition accuracy for many fonts square measure currently common. Some systems square measure capable of reproducing formatted output that closely approximates the first page as well as pictures, columns, and alternative non-textual elements. OCR have very wide applications editing the large print material, postal address reading, direct processing of documents, foreign language recognition and well known ANPR (Automatic Number Plate Recognition) etc.

OCR consists of many steps like pre-processing, segmentation, classification and recognition. The task of the pre-processing relates to the removal of noise from the sentence. OCR is being used in several areas like mail sorting, bank processing, and document reading

II. PHASES OF OCR

Digitization:

Digitization is the technique of converting a paper-based handwritten document into electronic format. Here, each document consists of only one character. The electronic conversion is accomplished by using a method whereby a document is scanned and an electronic illustration of the original document as an image file format is produced. We used various scanners for digitization, and the digital image was go for next step that is pre-processing phase.

Pre-processing:

In The pre-processing step, there is a series of operations performed on the scanned input image. It enhances the image rendering it suitable for segmentation the gray-level character image is normalized into a window size. After noise attrition, we produced a bitmap image. Then, the bitmap image was transformed into a thinned image.

Segmentation:

The Segmentation phase is the most critical process. Segmentation is done by separation from the individual characters of an image. Segmentation of handwritten characters into different zones (upper, middle

and lower zone) and characters is more crucial than that of printed documents that are in standard form. This is mainly because of variability in paragraph, words of line and characters of a word, skew, slant, size and curved. At times, components of two adjacent characters may be touched or overlapped and this situation creates difficulties in the segmentation task. Touching or overlapping issue occurs frequently because of modified characters in upper-zone and lower-zone. Segmentation is an important stage.

Extraction:

In this phase, features of individual characters are extracted. The performance of an each character recognition system that depends on the features that are extracted. The extracted features from input character should allow classification of a character in a different way. We used diagonal features, intersection and open endpoints features, transition features, zoning features, directional features, parabola curve fitting–based features, including power curve fitting–based features in order to find the feature set for a given character.

III. LITERATURE REVIEW

Malakar et al.(2012)[1] has depicted that extraction of text lines from document images is one of the important steps in the process of an Optical Character Recognition (OCR) system. In event of handwritten document images, presence of skewed, touching or overlapping text line(s) makes this process a real challenge to the researcher. The present technique extracts 87.09% and 89.35% text lines successfully from the said databases respectively.

Automatic number plate recognition [2] is used for mass surveillance technique making use of optical character recognition on images to identify vehicle registration plates. ANPR has also been made to save the images captured by the cameras including the numbers captured from license plate. ANPR technology own to plate variation from place to place as it is a region specific technology. They are used by various police forces and as a way of electronic toll collection on pay-per-use roads and cataloging the movements of traffic or individuals.

MAJIDA ALI ABED HAMID ALI ABED ALASADI [2005][3] This manuscript considers a new approach to Simplifying Handwritten Characters Recognition based on simulation of the behaviour of schools of fish and flocks of birds, called the Particle Swarm Optimization Approach (PSOA).

Ivan Dervisevic [2006][4] Success of optical character recognition depends on a number of factors, two of which are feature extraction and classification algorithms. In this paper we look at the results of the application of a set of classifiers to datasets achieved through various basic feature extraction methods.

Optical Music Recognition [5] without any human correction or human effort, Automatic number plate recognition [6] and Handwritten Recognition [7] .

After the recognition stage, if there are some unrecognised characters found, those characters are given their meaning in the post-processing stage. Extra templates can be added to the system for providing a wide range of compatibility checking in the systems database [8].

IV. PROPOSED APPLICATION

The aim is to provide high accuracy for OCR on a mobile based application without involving the complications of the internet.

The application provides a highly accurate OCR system. The main feature of the application is that, it performs without the internet.

V. WORKING

Step 1: Input is provided to the application ie. Image.
Step 2: Whole line or Sentence is scanned.
Step 3: Segmented line is then broken into words.
Step 4: Words are classified into characters.
Step 5: Then the whole process is reversed to formed the sentence

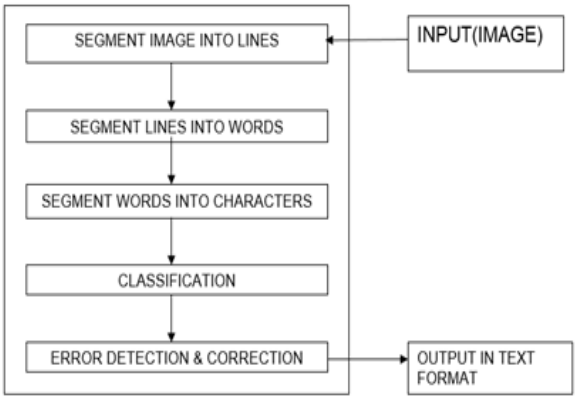

Figure 1 : Block Diagram of OCR

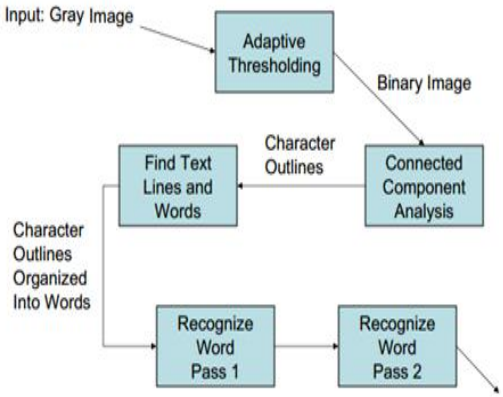

Figure 2: Architecture of OCR

VI. APPLICATIONS

1. Data entry for business documents ie. cheque , bank statement and receipt.
2. Automatic number plate recognition.
3. Documentation of historical books and documents.
4. Mail sorting etc.

VII. CONCLUSION AND FUTURE SCOPE

This is very easy, efficient and low cost method for documentation. This application allows us to get rid from multiple cards and documents carrying in our pocket. In the future this application will be modified and to have more accuracy. More features will be added to the application to make it more advance.

REFERENCES

[1] Malakar, Samir, et al. "Text line extraction from handwritten document pages using spiral run length smearing algorithm." Communications, Devices and Intelligent Systems (CODIS), 2012 International Conference on. IEEE, 2012.
[2] Badawy, W. "Automatic License Plate Recognition (ALPR): A State of the Art Review." (2012): 1-1.
[3] 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 ISSN 2286-4822
[4] Ivan Dervisevic Machine Learning Methods for Optical Character Recognition December 18, 2006
[5] A. Singh, K. Bacchuwar, A. Choubey, S. Karanam, D. Kumar, “An OMR Based Automatic Music Player”, in 3rd International Conference on Computer Research and Development (ICCRD 2011) in, (IEEE Xplore), 2011, Vol. 1, pp. 174-178.
[6] S.L. Chang, T. Taiwan , L.S. Chen, Y.C. Chung, S.W. Chen, “ Automatic license plate recognition” in IEEE Transactions on Intelligent Transportation Systems, 2004, Vol: 5 , Issue: 1, p.p. 42 - 53
[7] R Plamondon, S. N. Srihari, "On-line and off-line handwriting recognition: a comprehensive survey" IEEE transaction on pattern Analysis and machine Intelligence, 2000, 22(1), 63-84
[8] “Combination of Document Image Binarization Techniques”, 2011 International Conference on Document Analysis and Recognition