Tesseract Vs Gocr A Comparative Study

Shivani Dhiman, A.J Singh

Abstract- Optical Character Recognition (OCR) is a technique used to convert scanned images into machine readable text formats. Different types of Optical Character Recognition (OCR) Tools are used in market from earlier times have their own strengths and weaknesses. They provided different results on the basis of different metrics or parameters. But in this paper we are going to compare two open source tools i.e. Tesseract and GOCR. This paper firstly provides the introduction of open source tools Tesseract and GOCR, architecture of Tesseract and description about their working. In this paper, Tools are compared on the basis of Precision as well as Accuracy by considering different parameters that are Image Type, Resolution, Brightness and Font Type.

Keywords- Optical Character Recognition (OCR), Open Source, Tesseract and GOCR.

I. INTRODUCTION

Optical Character Recognition (OCR) is a technique for converting handwritten or machine printed document to editable text. There is problem in recognition with hand written text as compare to machine printed text. This is due to some uncertainties such as variation in calligraphy over period of time, similarity in text, variation in styles of writing [1]. That so why, we preceded our study of comparing tools by considering machine printed text as base excluding handwritten text. OCR becomes useful when you want to use some text from the Image or any printed text. And you don't have time to write it down so you can easily extract that text with the help of different OCR Tools. OCR works with images that almost consist of text in it. The output of a tool is dependent on the type of input image [5]. Achieving 100% accuracy is not possible but it is better to have something rather than nothing**.**

II. TESSERACT

Tesseract is an open source Optical Character Recognition (OCR) Engine. It was developed by HP between 1984 and 1994[2]. It was one of the top 3 engines in the 1995 UNLV Accuracy test. Combined with the Leptonica Image Processing Library it can read a wide variety of image formats and convert them to text [3]. Its latest version is 3.02 which were released on 28th October, 2012 under Apache License. Tesseract can work over 60 Languages. It is under further development by Google [3].

A. ARCHITECTURE OF TESSERACT

The Architecture of Tesseract is explained with the help of below described Figure 1. The first step is to provide Input to the Engine that may be Gray scale or Colored Image [4]. That particular input image is then converted into Binary Image with the help of Adaptive Thresholding.

After that the Second step is Connected Component Analysis which is used to store the outlines of component. With the help of nesting these outlines are gathered together into Blobs [2]. Then these Blobs are organized into text lines and these text lines are further broken into words according to space between the characters. This process takes place with the help of Fuzzy spaces.The process of Recognition takes place with the help of two pass process. In the first pass, after recognizing each word the satisfactory word is passed to an adaptive classifier as training data. Whereas in second pass, the words are recognized that are not recognized during first pass [2]. After resolving Fuzzy spaces in final phase we get the output as extracted text from image [5].

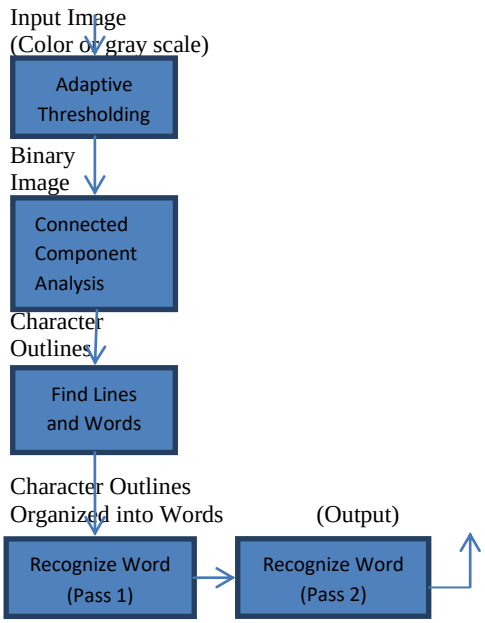

Fig1: Architecture of Tesseract

B. WORKING OF TESSERACT

Tesseract is a command based tool. Tesseract command takes two arguments: First argument is image file name that contains text and second argument is output text file in which, extracted text is stored [5]. Tesseract takes input as .tif image and gives output as .txt file. So first of all .jpg images from which we want to extract data are converted into .tif images with the help of Image magick software as shown.

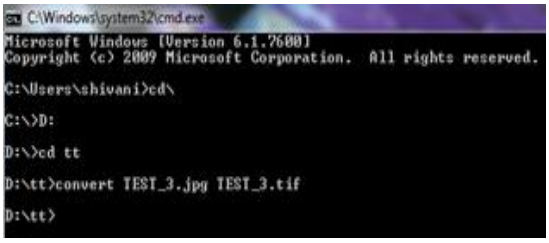

Fig2: Convert .jpg to .tif File

Manuscript received September 2013.

Ms. Shivani Dhiman, Dept. of Computer Science, Himachal Pradesh University, Shimla, India.

Dr. A.J Singh, Professor, Dept. ofComputer Science, Himachal Pradesh University, Shimla, India.

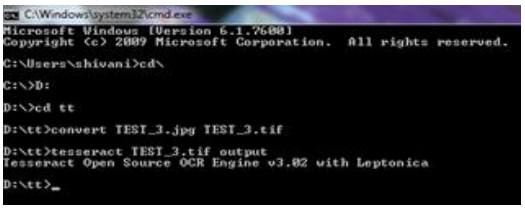

Fig3: Convert .tif to .txt File

The input File taken by Tesseract is .tif file shown in Fig5. So first of all .jpg file shown in fig4 should be converted into .tif file then using Tesseract it is converted into .txt file shown in Fig6 as output.

Fig4: .jpg File

Fig5: .tif File

Fig6: .txt File generated by Tesseract

III. GOCR

GOCR is Optical Character Recognition Tools. It is developed under the GNU public License [6]. It can read images in many formats (pnm, pgm, pbm, and ppm) and gives output as Text file [7]. It is a simple and fast engine which does not require any training data. Its recognition process takes two passes. In first pass, entire document is called. In second pass the unknown characters are called [8].

A. WORKING OF GOCR

GOCR is again a Command based tool same as Tesseract but differ from Tesseract in some aspects. The input taken by GOCR tool is generally in .pnm file and gives output as .txt file in Fig10. But before taking .pnm file shown in Fig9 as input .jpg file shown in Fig4 should be first converted into .pnm file and then with the help of GOCR .pnm file is converted into .txt file. Whole process is shown below:

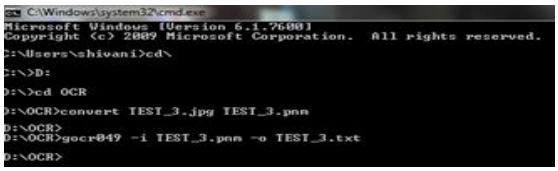

Fig7: Convert .jpg File to .pnm File

Fig8: Convert .pnm File to .txt File

Fig9: .pnm File

B. BRIGHTNESS

Third parameter considered for comparison is Brightness. Images are scanned with different Brightness that is 25, 50 and 100. Variable performance is shown by both tools due to change in brightness value. Image used is of color type for good results. Here In the case of color image with brightness value 100, GOCR provided 100% precision but for other values Tesseract again provided better results than GOCR. The results are shown in Table3.

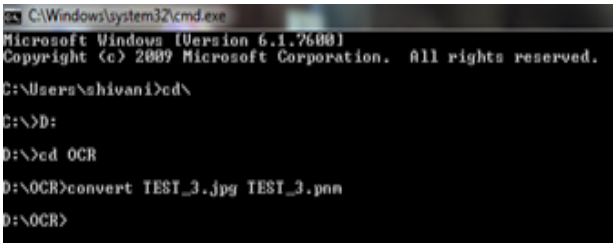

Fig10: .txt File generated by GOCR

IV. COMPARING TESSERACT AND GOCR

Different Optical Character recognition Tools are compared early but in this paper we are comparing two tools named Tesseract and GOCR on the basis of their Accuracy and Precision [9] by undertaking different parameters [10][11]. The images tested by these tools contain total number of 39 characters. The basic resolution used is 200dpi with white Background of an image and yellow background in case of colored image. The printed documents were digitalized using a Hewlett-Packard Deskjet 1050 Scanner with maximum resolution of 1200 dpi. Documents were digitized in True Color in position of Landscape. Different Parameters [10][11] used are as follows:

A. IMAGE TYPE

There are different types of Images that are used now days. We can observe that how an output of engine varies with different types of images as input. The types of Images used are as Color, Gray scale and Black-White images. It is observed from above Table1 that accuracy and precision for different image types are provided well with Tesseract as compare to GOCR. But GOCR provided good results for color image than other type of images.

B.RESOLUTION

Different types of images are scanned with different resolutions using HP Deskjet 1050 scanner having maximum resolution of 1200 dpi.The images are scanned with resolutions 75, 300,600 and 1200 dpi. The results observed are described below in Table 2.The accuracy and precision observed by Tesseract and GOCR are almost same in some cases but for some images due to variation of resolutions Tesseract provides better results than GOCR.

In some cases no character is determined or some error occurred.

D. FONT TYPE

And the last but not the least parameter used for comparison is Font type. Different Font types used are as Arial, Roman and Tahoma. Gray scale type of image is used. As shown in Table4.

V. ERROR ANALYSIS

In the study it is observed that there are number of mistakes arise. Some of them are:

1. 'T' is recognized as T.
2. 'q' is recognized as 0.
3. 'i' is recognized as ĩ.
4. 'o' is recognized as 0.
5. 'p' is recognized as 0.
6. 'd' is recognized as cl or 0.
7. 'g' is recognized as 0.
8. '!' is recognized as t.

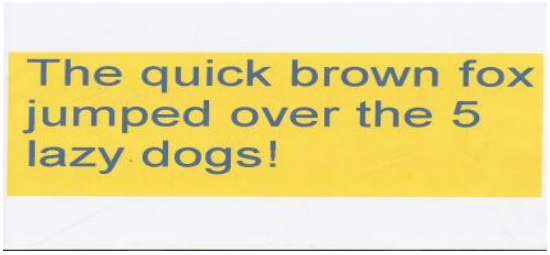

Fig11: Input Image to GOCR

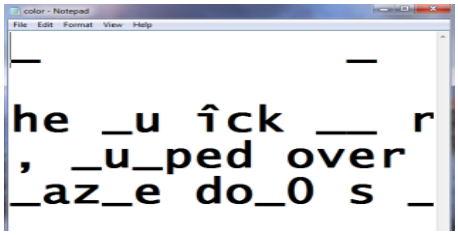

Fig12: Output generated by GOCR

VI. CONCLUSION

It is concluded from the study by considering different types of parameters that Tesseract has better accuracy and Precision than GOCR in most of the cases. But In rare cases GOCR also provided good results. We have used images of different types like colored, Gray scale and Black-White with different Resolutions, Brightness and Font type. Which provided different values of Precision and Accuracy for both the tools. The output of a tool depends on the type of input. Therefore results might be different for different types of input. It was difficult towork with handwritten documents as compare to machine printed documents. Therefore, we preceded our study byundertakingmachine printed documents.

ACKNOWLEDGEMENT

I would like to give sincere thanks to my Guide Dr. A.J. Singh, Professor, Himachal Pradesh University for all the help he has offered me during the development of this research.

Table1: Comparison between Tools using different Image types

IMAGE TYPE	TOTAL CHARACTERS EXTRACTED BY TESSERACT	CHARACTERS CORRECTLY EXTRACTED BY TESSERACT	TOTAL CHARACTERS EXTRACTED BY GOCR	CHARACTERS CORRECTLY EXTRACTED BY GOCR	TESSERACT ACCURACY (IN %)	GOCR ACCURACY (IN %)	TESSERACT PRECISION (IN %)	GOCR PRECISION (IN %)
Color	39	38	28	25	97.4	64.1	97.4	89.2
Gray scale	39	38	24	18	97.4	46.1	97.4	75.0
Black and White	39	38	27	19	97.4	48.7	97.4	70.3

Table2: Comparison between Tools using different Resolution values.

IMAGE TYPE	RESOLUTION	TOTAL CHARACTERS EXTRACTED BY TESSERACT	CHARACTERS CORRECTLY EXTRACTED BY TESSERACT	TOTAL CHARACTERS EXTRACTED BY GOCR	CHARACTERS CORRECTLY EXTRACTED BY GOCR	TESSERACT ACCURACY (IN %)	GOCR ACCURACY (IN %)	TESSERACT PRECISION (IN %)	GOCR PRECISION (IN %)
Color	75	39	38	39	38	97.4	97.4	97.4	97.4
Color	300	39	38	_	_	97.4	_	97.4	_
Color	1200	0	0	_	_	_	_	_	_
Gray scale	75	_	_	39	38	_	97.4	_	97.4
Gray scale	300	_	_	_	_	_	_	_	_
Gray scale	600	39	38	4	2	97.4	05.1	97.4	50.0
Black and White	75	39	38	39	38	97.4	97.4	97.4	97.4
Black and White	300	39	38	_	_	97.4	_	97.4	_
Black and White	1200	_	_	_	_	_	_	_	_

Table3: Comparison between Tools using different Brightness values.

BRIGHTNESS	TOTAL CHARACTERS EXTRACTED BY TESSERACT	CHARACTERS CORRECTLY EXTRACTED BY TESSERACT	TOTAL CHARACTERS EXTRACTED BY GOCR	CHARACTERS CORRECTLY EXTRACTED BY GOCR	TESSERACT ACCURACY (IN %)	GOCR ACCURACY (IN %)	TESSERACT PRECISION (IN %)	GOCR PRECISION (IN %)
25	39	37	28	23	94.8	58.9	94.8	82.1
50	39	38	27	26	97.4	66.6	97.4	96.2
100	39	37	1	1	94.8	02.5	94.8	100

Table4: Comparison between Tools using different Font Types.

FONT	TOTAL CHARACTERS EXTRACTED BY TESSERACT	CHARACTERS CORRECTLY EXTRACTED BY TESSERACT	TOTAL CHARACTERS EXTRACTED BY GOCR	CHARACTERS CORRECTLY EXTRACTED BY GOCR	TESSERACT ACCURACY (IN %)	GOCR ACCURACY (IN %)	TESSERACT PRECISION (IN %)	GOCR PRECISION (IN %)
Arial	39	37	7	0	94.8	0	94.8	0
Roman	39	38	6	2	97.4	05.1	97.4	33.3
Tahoma	39	37	3	0	94.8	0	94.8	0

REFERENCES

1. O.PSharma, M.K Ghose, K.B Shah and B.K Thakur, "Recent Trends and Tools for Feature Extraction in OCR Technology." International Journal of Soft Computing and Engineering (IJSCE), 2013, ISSN: 2231-2307, Volume-2, Issue-6.
2. R.Smith,"An Overview of the Tesseract OCR Engine." In proceedings of Document analysis and Recognition, ICDAR 2007, IEEE Ninth International Conference3.
3. The Tesseract open source OCR engine, http://code.google.com/p/tesseract-ocr.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

4. GOCR Reference, http://www.redhat.com/archives/blinux-list/2004-March/msg00201.html
5. A. Stromme and R. Carlson,"Minimally Supervised Methods to Correct Optical Character Recognition." Swarthmore College, Swarthmore, PA 19081.
6. T. Kanungo, G.A Marton and O. Bulbul, "Omni Page vs. Sakhr: Paired Model Evaluation of Two Arabic OCR Products."Centre for Automation Research University of Maryland College Park, MD 20742.
7. R.D Lins and N.F Alves, "A New Technique for Accessing the Performance of OCRs." IADIS International Conference on Applied Computing, 2005.
8. C.A.B Mello and R.D Lins,"A Comparative Study on OCR Tools." Vision Interface '99, Trois-Rivières, Canada, 19-21 May.
9. R. Mithe, S. Indalkar and N. Divekar,"Optical Character Recognition." International Journal of Recent Technology and Engineering (IJRTE) 2013, ISSN: 2277-3878, Volume-2, Issue-1.
10. C. Patel, A. Patel and D. 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.
11. GOCR open source OCR engine, http:// jocr.sourceforge.net/.