ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICITCSA 2017

Pioneer College of Arts and Science, Coimbatore

Vol. 6, Special Issue 1, January 2017

Online Optical Character Recognition (OCR) Tools - Performance Analysis

Dr. S.Vijayarani[1], **Ms. A.Sakila**[2]

[1]Assistant Professor, Department of Computer Science, Bharathiar University, Coimbatore

[2]Ph.D Research Scholar, Department of Computer Science, Bharathiar University, Coimbatore

Abstract- Optical Character Recognition (OCR) is a technique, which is used to convert the document images into editable text format. Many different types of OCR tools are freely and commercially available today. The primary objective of this work is to compare the performance of the open source OCR tools for extracting the text information from the image (Table format). The main functions of these tools are to convert the images into text format. Eight different types OCR tools are considered for this analysis. From this analysis is observed that the performances of OCR Convert and My Free OCRtools are better than other OCR tools.

Keywords-Optical Character Recognition, Online OCR, Free Online OCR, OCR Convert, My free OCR, Free OCR, i2OCR, To-text.net, Google Docs.

I. INTRODUCTION

Optical Character Recognition (OCR) is a technique, which is used to identify the text from the images, and then convert into their text format. OCR technique is Recognition Based Information Retrieval; it retrieves the text from the images. Retrieved text should be stored in various word files like Note pad, Rich text documents, MS Office Word, PDF etc. [8]. It supports all types of image formats such as JPG, PNG, BMP, GIF, TIFF and multi-page PDF files [19]. An OCR technique analyzes the captured or scanned document images and then translates character images into character codes (e.g. ASCII codes), therefore it should be easy to edited and searched [3]. There are different types of OCR tools are commercially and freely available today. Commercial OCR tools are Abbyy Fine Reader PRO, OmniPage Standar, Readiris Pro, Captricity, Top OCR, etc. It works with images that almost consist of text in it [1]. To improve the accuracy of the text most of the OCR tools use dictionaries to recognize individual characters then it try to recognize entire words that exist in the selected dictionary. Sometimes these tools are very difficult to extract text from the image because of different font size, style, symbols and dark background [21]. If we are using high resolution documents the OCR tools will produce best results.

The remaining portion of this paper is discussed as follows. Section II describes various types of OCR tools and they results. Section III discusses the performance analysis and conclusion is given in Section IV.

II. OCR TOOLS COMPARISON

This work compares eight different types of Online OCR tools. They are Online OCR, Free Online OCR, OCR Convert, My free OCR, Free OCR, i2OCR, To-text.net, Google Docs. Figure 1 shows the sample input image considered for performance analysis, this input image download from goolgle image.

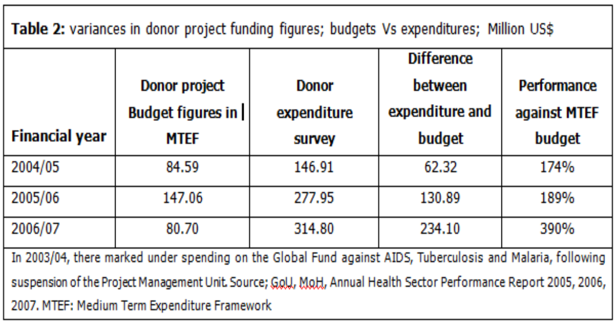

Table 2: variances in donor project funding figures; budgets Vs expenditures; Million US$

Financial year	Donor project Budget figures in | MTEF	Donor expenditure survey	Difference between expenditure and budget	Performance against MTEF budget
2004/05	84.59	146.91	62.32	174%
2005/06	147.06	277.95	130.89	189%
2006/07	80.70	314.80	234.10	390%

In 2003/04, there marked under spending on the Global Fund against AIDS, Tuberculosis and Malaria, following suspension of the Project Management Unit. Source; GoU, MoH, Annual Health Sector Performance Report 2005, 2006, 2007. MTEF: Medium Term Expenditure Framework

Fig. 1 Sample Input Image

A. *Online OCR*

Online OCR tool is open source OCR software that permits to reform (convert) scanned PDF documents, faxes, photographs or digital camera captured images into editable and searchable documents [4]. The result which is displayed in this tool has different formats and supports various languages [4].Its maximum input file size is 100 MB [4] [19]. The conversion output of the sample input image is given in below coding.

Tablet: variances in donor project funding figures; budgets Vs expenditures; Million US$ Financial year DOOM pr.. Budget figures in i MTEE Donor expenditure survey Difference between expenditure and budget Perfomunce against MTEE budget 2004/05 84.59 146.91 62.32 17496 2005/06 147.06 277.95 130.89 18996 2006/07 60.70 314.60 239.10 390.

B. *Free Online OCR*

NewOCR.com is free online OCR software that can analyze and converts the text from images. Input files supported by this tool are JPEG, JFIF, PNG, GIF, BMP, PBM, PGM, PPM, PCX and multipage [4]. After conversion the result is displayed in different formats like Plain text (TXT), Microsoft Word (DOC) and Adobe Acrobat (PDF). It supports different languages and also supports several font types [21]. The advantage of this software, it has taken unlimited uploads [19]. The resultant output [5] is illustrated below.

Difference
Donor project Donor between Performance
Budget figura in I expenditure expenditure and against
MTEF
suspension ofthe Project Management Unit. Source; 9.9.”:
Mott, Annual Health Sector Performance Report 2005,
2006,
2007. MTEF: Medium Term Expenditure Framework

C. *OCR Convert*

OCR Convert is free online OCR software, which provides the facility to convert the scanned image into text [19]. It supports different image formats such as JPG, PNG, BMP, GIF, TIFF and multi-page PDF files and also support low resolution images[6] [19]. The result may be in text format and this tool supports simultaneous uploads and able to perform conversion process of files up to 5MB (aggregated). The output text result is shown below.

Table 2: variances in donor project funding figures; budget» Vs expenditures; Million US$
Difference
Donor project Donor between Performance
Budget figurs in I expenditure expenditure and against
MTEF
Fimmlal Y9-3" MTEF survey budget budget
2004/05 84.59 146.91 62.32 174%
2005/06 147.06 277.95 130.89 189%
2006/07 80.70 314.80 234.10 390%
In 2003/04, there marked under spending on the Global Fund against AIDS, Tuberculosis and Malaria, following
suspension ofthe Project Management Unit Source; 5,911,
|)(Lo,i;1, Annual Health Sector Performance Report 2005,
2006,
2007. MTEF: Medium Term Expenditure Framework

D. *My free OCR*

My Free OCR is a tool which is used to recognize the characters from an image and the documents which are uploaded in this tool are automatically deleted after conversion. The output result as follows.

Table 2: variances in donor project funding figures; budgets Vs expenditures; Million USS Financial year Donor project Budget figures in I MIFF Donor expenditure survey Difference between expenditure and budget Performance against MIFF budget 2004/05 84.59 146.91 62.32 174% 2005/06 147.06 277.95 130.89 189% 2006/07 80.70 314.80 234.10 3909’o In 2003/04, there marked under spending on the Global Fund against AIDS, Tuberculosis and Malaria, following suspension of the Project Management Unit Source; CPU, tI Annual Health Sector Performance Report 2005, 2006, 2007. MTEF: Medium Term Expenditure Framework

E. *Free OCR*

Free-OCR.com is a free online OCR tool, which is used to extract text from any images and convert these images into an editable text document.

It takes a JPG, GIF, TIFF BMP or PDF (only first page) file formats and supports30 languages. This tool only supports less than 2MB [8]. Result is illustrated below.

Table 2: variances in donor project funding figures; budgets Vs expenditures; Million US$
Difference
Donor project Donor between Performance
Budget figure in I expenditure Bqlenditure and against
MTEF
Fi"a"¢ia| Veal’ MTEF survey budget budget
2005/00 147.06 277.95 130.09 2006/07 80.70 314.80 234.10 390%

In 2003/04, there marked under spending on the Global Fund against AIDS, Tuberculosis and Malaria, following
suspension ofthe Project Management Unit. Source; 591,1,
Mod, Annual Health Sector Performance Report 2005,
2006,
2007. MTEF: Medium Term Expenditure Framework

F. *i2OCR*

Converting text from images usingi2OCR,it’s a free online Optical Character Recognition software.After converting text can be edited, formatted, indexed, searched, or translated [19]. **Input image file types are**TIF, JPEG, PNG, BMP, GIF, PBM, PGM and PPM [21]. It takes unlimited uploads and supports more than 60Languages. The output result of the i2OCR [9] is given below.

Table 2: variances in donor project funding figures; budgets Vs expenditures; Million US$
Difference
Donor project Donor between Performance
Budget figura in I expenditure expenditure and against
MTEF
Fimficlal Y9-3" MTEF budget budget
survey

In 2003/04, there marked under spending on the Global Fund against AIDS, Tuberculosis and Malaria, following
suspension of the Project Management Unit. Source; gag,
uLo,1;1, Annual Health Sector Performance Report 2005,
2006, 2007. MTEF: Medium Term Expenditure Framework

DOI 10.17148/IJARCCE

G. *To-Text.net*

To-Text.net is freeonline OCR software, it supports PDF, JPEG and scanned images into editable documents. This software supports processing documents in 40 recognition languages[20]. The output result follows.

Fable 2: vznznces m donor project funding figures; budgets vs expenanmes; Mllllon uss
fiffuulm
Dam! nnipct Dnunr hem... Psfuilulm
nu-met figuvs in I Etlluldilme zxnuldilme and aginsl mgr
fir-av-2'2! var ms: mlvzv budget budget
zacu/as M59 146.91 52.32 174%
zoos/as 147.06 277.95 1341.39 IE9“/as
2aa6/a7 M.7a sum 234.10 390%
m zuu:/no. um malkcd undu spumg an I11: cum: Fund against mas, Tubumlnsns and Malaria, fnllnwmfl
susmsm nrmsnn;u2nsnagam-munnsuurce; aw. um. Annual Haalltw S1-nnr Pufnlmanua mm zuus, zuua.
2uu7. ME: ma.-Am Tam Bvuwdlmrc Framewmk

H. *Google Docs*

Google Docs converts images and scanned pdf into text format.

It performs OCR on images and PDFs as large as 2 MB [17], in the output format of Google docs are ODT, PDF, TXT, RTF, DOC and HTML. It supports 30languages [11], the output text resultis represented as,

Table 2: variances in donor project funding figures; budgets Vs expenditures; Million US$
Difference
Donor project Donor between Performance
Budget figura in I expenditure expenditure and against MTEF
Fimficlal Y9-3" MTEF budget budget
survey

In 2003/04, there marked under spending on the Global Fund against AIDS, Tuberculosis and Malaria, following suspension of the Project Management Unit. Source; gag, uLo,1;1, Annual Health Sector Performance Report 2005, 2006, 2007. MTEF: Medium Term Expenditure Framework

III. PERFORMANCE ANALYSIS BETWEEN OCR TOOLS

In order to perform the comparative analysis of the OCR tools, this work has applied two performance measures; they are conversion accuracy and error rate. Conversion accuracy is used to identify whether the alphabets are converted accurately or not. Error rate helps to identify number of alphabets not converted properly. Table 1 shows the Accuracy and Error rate of different OCR tools.

TABLE 1 Performance Analysis of OCR Tools

S. No	OCR Tools	Character Accuracy %	Character Error Rate %
1	Online OCR	47.06	52.94
2	Free Online OCR	62.35	37.65
3	OCR Convert	95.29	4.71
4	My free OCR	95.29	4.71
5	Free OCR	92.94	7.06
6	I2ocr	94.12	5.88
7	To-text	28.24	71.76
8	Google Docs	94.12	5.88

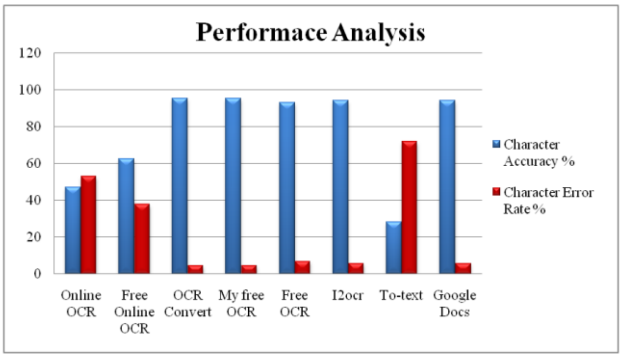

Fig. 2 Character accuracy and Error rate between the OCR tools

IV. CONCLUSION

This paper analyzes the eight different types of OCR tools. From this analysis, OCR Convert and My Free OCR tools are better than other OCR tools, it gives higher accuracy than other OCR tools. But all converted text can't be stored in proper layout, because all contents to be merged, hence very difficult to understand the converted text. In Future, both table with text images will store proper layout in the word documents and these issues are to be handled by developing new techniques and algorithms.

REFERENCES

[1] ShivaniDhiman, A.J Singh, “TesseractVsGocr A Comparative Study”,International Journal of Recent Technology and Engineering, ISSN: 2277-3878, Volume-2, Issue-4.
[2] Chirag Patel, Atul Patel, Dharmendra Patel, “Optical Character Recognition by Open Source OCR Tool Tesseract: A Case Study”,International Journal of Computer Applications (0975 – 8887), Volume 55– No.10
[3] http://en.wikipedia.org/wiki/Optical_character_recognition
[4] http://www.onlineocr.net/
[5] http://www.newocr.com/
[6] http://www.ocrconvert.com/
[7] http://www.convertimagetotext.net/
[8] http://www.free-ocr.com/
[9] http://www.i2ocr.com/
[10] http://www.ocrtoword.com/
[11] https://docs.google.com/
[12] Yasser Alginahi, “Preprocessing Techniques in Character Recognition”

[13] Oivind due trier, Anil K.Jain, TorfinnTaxt, “Future extraction methods for character recognition A survey”.
[14] Pritpal Singh, SumitBudhiraja, “Feature Extraction and Classification Techniques in O.C.R. Systems for Handwritten Gurmukhi Script – A Survey”, International Journal of Engineering Research and Applications (IJERA), Vol. 1, Issue 4, pp. 1736-1739.
[15] Youssef Bassil, Mohammad Alwani, “OCR Post-Processing Error Correction Algorithm Using Google's Online Spelling Suggestion”, Journal of Emerging Trends in Computing and Information Sciences, Vol.3, No. 1
[16] Archana A. Shinde, D.G.Chougule, “Text Pre-processing and Text Segmentation for OCR”, IJCSET, Vol2.
[17] http://www.labnol.org/software/convert-images-to-text-with-ocr/17418/
[18] Sandeep Dangi, Ashish Oberoi, Nishi Goel “Performance Comparison between Different Feature Extraction Techniques with SVM Using Gurumukhi Script”, International journal of Engineering Research and Applications, ISSN : 2248-9622, Vol. 4, Issue 7, July 2014, pp.123-128
[19] Dr. S.Vijayarani and Ms. A.Sakila “PERFORMANCE COMPARISON OF OCR TOOLS”, International Journal of UbiComp (IJU), ISSN : 0975 –8992 (Online) ; (Online) ; 0976 – 2213 (Print), Vol.6, No.3, July 2015. Pp.19-30 .
[20] http://www.to-text.net/
[21] http://www.slideshare.net/ijujournal/performance-comparison-of-ocr-tools

BIOGRAPHIES

Dr. S. Vijayarani has completed MCA, M.Phil, Ph.D in Computer Science. She is working as Assistant Professor in the School of Computer Science and Engineering, Bharathiar University, Coimbatore. Her fields of research interest are data mining, privacy and security issues in data mining and data streams. She has published papers in the international journals and presented research papers in international and national conferences.

Ms. A. Sakila has completed M.Sc, M.Phil in Computer Science. She is currently pursuing her Ph.D in Computer Science in the School of Computer Science and Engineering, Bharathiar University, Coimbatore. Her fields of research interest are Image mining, Data Mining and Multimedia Mining. She has published papers in the international journals and presented research papers in international and national conferences.