Comparative Study of Character Recognition Tools

Purna Vithlani
Research Scholar,
Department of Computer Science,
Saurashtra University, Rajkot.

C.K.Kumbharana, Ph.D
Head,
Department of Computer Science,
Saurashtra University, Rajkot.

ABSTRACT

Optical Character Recognition tools are used to convert handwritten or typewritten characters to machine editable form. OCR tools are either a stand-alone command-line application or GUI application. In this paper, Tesseract tool, GOCR tool and other Desktop and Web OCR tools are presented. Among them, Researchers have selected six open source and free tools for handwritten character recognition and presented comparative study of it. Researchers have collected 2520 handwritten character and digit samples from the various persons and compare it with six different tools and found out highest recognition rate of each character and digit.

General Terms

Character Recognition

Keywords

Tesseract tool, GOCR tool, Web OCR, Desktop OCR.

1. INTRODUCTION

Optical character recognition tools convert a scanned image or PDF file to searchable and editable text file. OCR tools are used as either frontend or backend for character recognition application. There are two types of OCR tools available: Web OCR (Online OCR), Desktop OCR.

Web OCR is also known as Online OCR. Web OCR will require that you upload your files on the internet to their servers, so there may be privacy concerns as well as time/bandwidth concerns if your document is big. Most have limits to file size and count of pages to process daily/weekly that they will process for free; for bigger jobs they require to buy extra processing power. On the flip side, many of these services are really good at the OCR itself [1].

Desktop OCR is also known as Offline OCR. With Desktop OCR you don’t need to worry about uploading sensitive information to foreign servers, or whether your file will take too long to upload. Some desktop OCR programs generally give better text review options, and some have scanning functionality integrated [1].

2. CHARACTER RECOGNITION TOOLS

In this paper, OCR tools which are used to identify handwritten and typewritten text are presented. Some of them are open source and free.

2.1 Tesseract Tool

Tesseract is free software, released under the Apache License and development has been sponsored by Google since 2006. Tesseract is one of the most accurate open source OCR engines currently available. It is available for Linux, Windows and Mac OS X [2]. Tesseract does not come with a GUI and is instead run from the command-line interface [3].

2.1.1 GUI Desktop Applications using Tesseract Tool

Free OCR, PDF OCR X, YAGF, gImageReader, VietOCR, OCRFeeder, Lector, Lime OCR, QTesseract and SunnyPage are GUI desktop application for Tesseract tool.

Free OCR is a useful tool for scanning and extracting text from images and PDFs. It supports most image files and multi-page TIFF files. It is also compatible with TWAIN devices like scanners. Free OCR is a complete scan OCR program [4]. But it cannot recognize formatting.

PDF OCR X is a simple drag-and-drop utility for Mac OS X and Windows that converts PDFs and images into text documents or searchable PDF files. It uses advanced OCR technology to extract the text of the PDF even if that text is contained in an image. It supports over 60 languages. It also supports batch processing.

YAGF is a graphical interface for cuneiform and tesseract text recognition tools on the Linux platform. With YAGF you can scan images via XSane, import pages from PDF documents, perform image preprocessing and recognize texts using cuneiform from a single command centre [5].

GImageReader reads text from images and PDFs. User can manually define and adjust recognition regions. It supports multipage PDF documents. It supports Spellchecking for output text (if corresponding dictionary installed). It removes line breaks in output text [6].

VietOCR, available in Java and .NET executable, is a GUI frontend for Tesseract OCR engine. VietOCR runs on multi-platform (Java version only) i.e. Windows, Solaris, Linux/Unix, Mac OS X and Others. It extracts text from PDF, TIFF, JPEG, GIF, PNG, BMP image formats. It supports batch processing and scanning facility.

OCRFeeder is a document layout analysis and optical character recognition system. It automatically outlines its contents, distinguish between what's graphics and text and perform OCR over the latter [7].

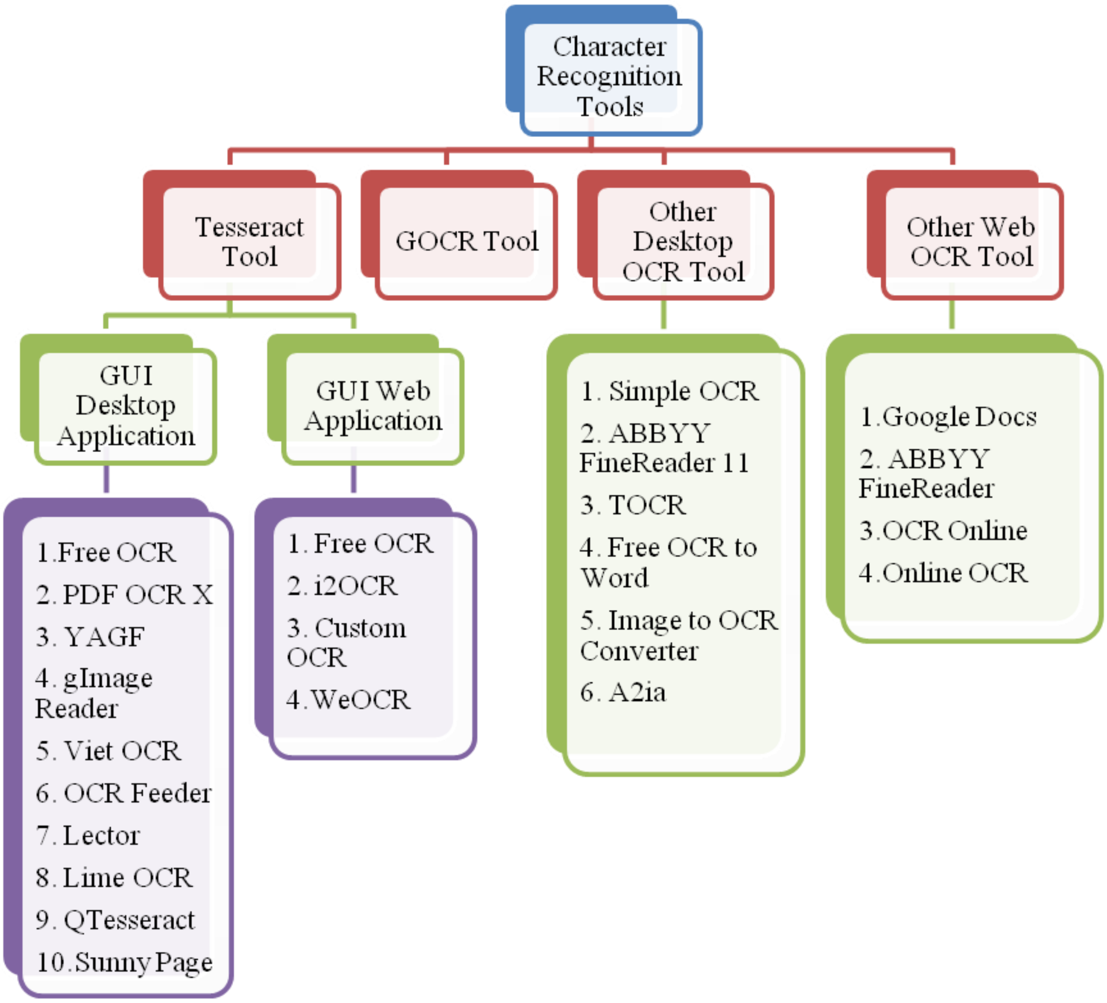

Figure 1: Character Recognition Tools

Lector is a graphical OCR solution for GNU/Linux based on Python, Qt4 and tesseract OCR. The resulting text can be proofread, formatted and edited directly in Lector [8]. It supports rotation of images, spellchecking, text editing.

Lime OCR is build with tessearact-ocr. Lime OCR was initially developed for internal use of Lime Consultants, and now published under GNU General Public License v3. Lime OCR is free, simple to use and currently supports 29 languages, and supports all tesseract-ocr trained data files [9].

QTesseract is a Graphical User Interface for the Tesseract OCR.

SunnyPage OCR is a GUI frontend for Tesseract OCR engine with automatic adjustment of image brightness; image processing and PDF support [10].

2.1.2 GUI Web Applications using Tesseract Tool

Free OCR, i2OCR, CustomOCR, and WeOCR are web applications using Tesseract tool.

Free-OCR is a free online OCR tool. It takes a JPG, GIF, TIFF, BMP or PDF (only first page) file. It can handle images with multi-column text and supports many languages. The images must not be larger than 2 MB, no wider or higher than 5000 pixels and there is a limit of 10 images uploads per hour [11].

I2OCR is an online service of OCR. I2OCR enables you to upload an image file from a URL (web, cloud etc). It supports 64 recognition languages. It analyzes the layout of the document and can extract text from multiple columns [12]. It gives facility to edit extracted text online using Google Docs or translated using Google or Bing translation service. I2OCR can’t read text from PDF File.

CustomOCR read text from scanned documents, webcam images of license plates or photo camera images of bank checks [13].

WeOCR is a platform for Web-enabled OCR (Optical Character Reader/Recognition) systems that enables people to use character recognition over networks [10].

2.2 GOCR Tool

GOCR is again a Command based tool same as Tesseract, it is also used as a backend to other programs. GOCR is an OCR

(Optical Character Recognition) program, developed under the GNU Public License. It converts scanned images of text back to text files. Joerg Schulenburg started the program, and now leads a team of developers [14]. GOCR program is written in C language and it runs on Linux, Windows and OS/2 platform.

GOCR 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 [15] [16].

GOCR claims it can handle single-column sans-serif fonts of 20–60 pixels in height. It reports trouble with serif fonts, overlapping characters, handwritten text, heterogeneous fonts, noisy images, large angles of skew, and text in anything other than a Latin alphabet. GOCR can also translate barcodes [17] [18].

2.3 Other Desktop OCR Tools

Simple OCR, ABBYY FineReader 11, Transym OCR (TOCR), Free OCR to word, Image to OCR Converter, A2ia are Desktop OCR tools.

Simple OCR is desktop tool. It is a proprietary optical character recognition application developed originally by Cyril Cambien of France under the title WOCAR (until v2.5 in 2001) [19]. Simple OCR offers Machine print and Hand written Recognition. Simple OCR uses its own OCR engine that is capable of learning the fonts in a particular document. Simple OCR gives facility to interactive correction with suggestions from dictionary. It includes both single file and batch of files processing mode. Simple OCR cannot convert PDF file into text file. Font and format detection is not possible in simple OCR [1].

ABBYY FineReader 11 efficiently converts an image file or PDF file into editable and searchable text file. ABBYY FineReader 11 detects any combination of 189 languages to help you expand your global capabilities. FineReader 11 supports a wide range of output formats. The results can also be sent directly to applications such as Microsoft Word, Excel and PowerPoint, Adobe Acrobat, Corel WordPerfect and OpenOffice.org Writer [20]**.** ABBYY FineReader is free for only 30 days and allows processing of up to 100 pages. It exports or saves maximum 3 pages to an external application at a time.

Transym OCR is one of the proprietary Optical Character Recognition tools, which provides the good amount of accuracy. Transym coverts the color images to gray scale images then does the OCR of these images. It reads broken, blurred and obscure characters. It allows to format output text.

Free OCR to Word tool is a simple and basic functionality OCR program. Free OCR to Word has a simple interface. So it is easy to use. It allows rotation of the image. Free OCR to word cannot recognize formatting. It does not support PDF and multi-page file. It supports only one language.

Image to OCR Converter is text recognition tool that can read text from BMP, PDF, TIF, JPG, GIF, PNG and all major image formats and saves the extracted text in WORD, DOC, PDF, HTML and TEXT formats with accurate text formatting and spacing. Complicated layout of legal documents, faxes, documents with tables, designs and photos captured with digital and phone cameras are accurately recognized and recreated in output formats. It recognizes more than 40 different languages. It provides security features such as password protection and watermark to the converted documents. It provides automatic detection and correction of rotated, skewed and tilted documents. Broken text and characters are also reconstructed to provide better accuracy and recognition [21]. Image to OCR converter allows converting only two pages in unregistered version at a time.
Security features such as password protection and watermark to the converted document are only provided in registration mode which is not free.

A2iA OCR tool recognizes isolated characters by distinguishing the individual shapes and sizes within each character classifier – identifying extracted such as curves, loops and lines- and organizing them in a logical and actionable manner according to the user’s information management specification [22].

2.4 Other Web OCR Tools

Google Docs, ABBYY FineReader, OCR Online, OnlineOCR.net are Web OCR tools.

Google Docs is a free, web-based office suite offered by Google within its Google Drive service. There are no limits on the number of files that you can process in a day. You can perform OCR on image and PDF files. It preserves most formatting. Documents, spreadsheets, presentations can be created with Google Docs, imported through the web interface, or sent via email. With Google Docs, you can perform OCR on images and PDFs as large as 2 MB. For PDF files, they only look at the first 10 pages when searching for text to extract. Text in some languages may not be recognized and processed.

ABBYY FineReader also provides online service. It supports maximum 30 MB file size. You can choose up to 3 recognition languages for a multilingual document. FineReader can convert multi-page document and also preserves formatting. FineReader can export your converted file to Google Docs, Ever note and Drop Box.

OCR Online is an online service for converting PDF and image file into editable and searchable text format. OCR Online has superior multilingual support and is capable of processing documents in 153 languages providing prime quality for a single language or any combination of the languages it supports [23]. Only 5 pages are allowed to convert in a week. If you want to convert more than 5 pages a week, you have to pay for it. It supports maximum 10 MB file size.

Table 1 Recognition Rate (In Percentage) of each Character and Digit using Character Recognition Tools

Character /Digit	Characte Recognition Tools						Highest Recognition Rate
	Simple OCR	Free OCR to Word	Image to OCR Converter	Free OCR	Custom OCR Online	PDF OCR X	
A	8.57	34.29	55.71	58.57	64.29	2.86	64.29
B	15.71	15.71	38.57	34.29	44.29	0	44.29
C	28.57	32.86	54.29	52.86	35.71	5.71	54.29
D	12.86	50	48.57	64.29	62.86	1.43	64.29
E	5.71	15.71	54.29	50	67.14	12.86	67.14
F	30	14.29	62.86	44.29	51.43	5.71	62.86
G	7.14	1.43	7.14	7.14	8.57	0	8.57
H	1.43	30	35.71	58.57	64.29	8.57	64.29
I	68.57	21.43	30	45.71	48.57	0	68.57
J	10	24.29	62.86	24.29	22.86	0	62.86
K	35.71	25.71	44.29	65.71	54.29	0	65.71
L	7.14	51.43	71.43	70	71.43	0	71.43
M	35.71	25.71	45.71	50	65.71	2.86	65.71
N	0	18.57	21.43	32.86	21.43	2.86	32.86
O	17.14	4.29	11.43	12.86	17.14	0	17.14
P	58.57	34.29	57.14	42.86	78.57	1.43	78.57
Q	5.71	62.86	45.71	54.29	64.29	15.71	64.29
R	10	21.43	42.86	35.71	44.29	10	44.29
S	50	8.57	41.43	21.43	15.71	7.14	50
T	11.43	22.86	55.71	47.14	65.71	11.43	65.71
U	50	62.86	81.43	54.29	75.71	32.86	81.43
V	50	48.57	88.57	25.71	35.71	31.43	88.57
W	40	41.43	44.29	44.29	62.86	32.86	62.86
X	51.43	24.29	32.86	40	24.29	18.57	51.43
Y	28.57	18.57	4.29	14.29	20	0	28.57
Z	34.29	35.71	30	58.57	74.29	0	74.29
1	37.14	25.71	22.86	28.57	44.29	0	44.29
2	47.14	14.29	17.14	12.86	14.29	0	47.14
3	68.57	20	58.57	30	37.14	2.86	68.57
4	17.14	7.14	22.86	11.43	31.43	7.14	31.43
5	64.29	21.43	48.57	32.86	65.71	4.29	65.71
6	32.86	5.71	27.14	27.14	37.14	2.86	37.14
7	71.43	17.14	22.86	22.86	55.71	5.71	71.43
8	51.43	5.71	1.43	1.43	5.71	0	51.43
9	68.57	10	17.14	5.71	8.57	0	68.57
0	28.57	12.86	65.71	12.86	18.57	0	65.71

Online OCR is a free web-based Optical Character Recognition tool that allows you to convert scanned PDF documents (including multipage files), faxes, photographs or digital camera captured images into editable and searchable electronic documents including Adobe PDF, Microsoft Word, Microsoft Excel, Rtf, Html and Txt [24].

3. EXPERIMENTAL RESULTS

Researchers have collected handwritten data samples for English capital alphabets and digits from seven persons of different ages. Each characters and digits are written 10 times by each person. Thus, Researchers have collected 70 samples of each characters and digits. Researchers have compared those 2520 samples with six different character recognition tools and found out the highest recognition rate which is shown in Table 1. Custom OCR Online gives the best result because it identifies 14 characters and 4 digits which is the highest recognition rate among the six evaluated tools.

As shown in Highest Recognition Rate column of Table 1, recognition rate of handwritten characters and digits in selected six tools is shown in Table 2 and Table 3.

Table 2 Highest Recognition Rate of Characters

Character Recognition Tool	**Characters**	**Total Characters**
Simple OCR	I, S, X, Y,	4
Free OCR to Word	-	0
Image to OCR Converter	C, F, J, L, U, V	6
Free OCR	D, K, N,	3
Custom OCR Online	A, B, E, G, H, L, M, O, P, Q, R, T, W, Z	14
PDF OCR X	-	0

Figure 2 represents column chart of Table 2.

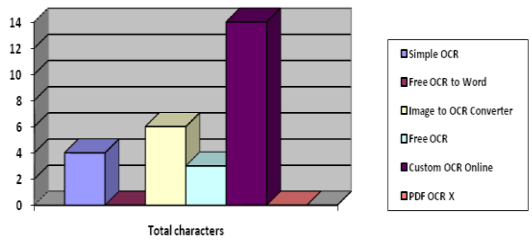

Figure 2 Column Chart representation of Table 2

As shown in Table 2 and Figure 2, Custom OCR Online gives highest recognition rate for 14 characters.

Table 3 Highest Recognition Rate of Digits

Character Recognition Tool	**Digits**	**Total Digits**
Simple OCR	2,3,7,8,9	5
Free OCR to Word	-	0
Image to OCR Converter	0	1
Free OCR	-	0
Custom OCR Online	1,4,5,6	4
PDF OCR X	-	0

Figure 3 represents column chart of Table 3.

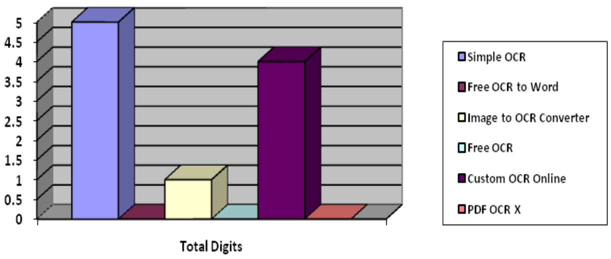

Figure 3 Column Chart representation of Table 3

As shown in Table 3 and Figure 3, Simple OCR gives the highest recognition rate for 5 digits.

4. CONCLUSION

In this paper, Optical character recognition tools are presented. Among them, Researchers have selected six open source and free tools for handwritten character recognition and presented comparative study of it. Due to variation in styles of writing of people, 100% accuracy is not possible in case of handwritten text. Although, Custom OCR Online gives better result for characters and Simple OCR gives better result for digits.

5. ACKNOWLEDGMENTS

Researchers are thankful to all the writers who have contributed by providing handwritten input for proposed experimental work.

6. REFERENCES

[1] http://www.freewaregenius.com/2011/11/01/how-to-extract-text-from-images-a-comparison-of-free-ocr-tools/

[2] http://en.wikipedia.org/wiki/Tesseract_(software)

[3] http://en.wikipedia.org/wiki/FreeOCR#User_interfaces

[4] http://www.comptalks.com/convert-images-to-text/

[5] http://symmetrica.net/cuneiform-linux/yagf-en.html

[6] http://sourceforge.net/projects/gimagereader/

[7] https://wiki.gnome.org/Apps/OCRFeeder

[8] https://code.google.com/p/lector/

[9] https://code.google.com/p/lime-ocr/

[10] https://code.google.com/p/tesseract-ocr/wiki/3rdParty

[11] http://www.free-ocr.com/

[12] http://www.i2ocr.com/

[13] http://www.customocr.com/

[14] http://jocr.sourceforge.net/

[15] Shivani Dhiman, A.J. Singh, “Tesseract vs Gocr A Comparative Study”, International Journal of Recent Technology and Engineering, Vol. 2, Issue 4, Sep. 2013.

[16] C.A.B Mello and R.D Lins, “A Comparative Study on OCR Tools.” Vision Interface '99, Trois-Rivières, Canada, 19-21 May.

[17] http://en.wikipedia.org/wiki/GOCR

[18] SfR Fresh (n.d.). "Member "gocr-0.45/README" of archive gocr-0.45.tar.gz"

[19] http://en.wikipedia.org/wiki/SimpleOCR

[20] http://abbyyindia.com/finereader/professional/features/

[21] http://products.softsolutionslimited.com/img2ocr/

[22] http://www.a2ia.com/

[23] http://www.ocronline.com/

[24] http://www.onlineocr.net/service/about