Text Extraction from Tamil and Hindi Document Images using Open Source Optical Character Recognition tools

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

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

Abstract- *Optical Character Recognition (OCR) is a technique, which is used to extract the text from document images and converted into text format. This kind of information retrieval is called as recognition based retrieval hence that it can be edited, searched, stored more efficiently. OCR is used for many applications such as library, organization, bank cheques, number plate recognition, historical book analysis and many others applications. Various OCR tools are available for converting document images in different types of languages. The primary objective of this work is to compare the performance analysis of the three different OCR tools for extracting the text information from Tamil and Hindi document images. The OCR tools considered in this analysis are Google Docs, Free Online OCR and i2OCR. Based on the conversion accuracy it is observed that the performance of Free Online OCR is better than other OCR tools.*

Keywords- *Optical Character Recognition, OCR architecture for Tamil and Hindi document images, Google Docs, Free Online OCR, i2OCR.*

I. INTRODUCTION

Scanned and Captured document images are becoming more popular in today's world and it used in paperless offices and digital libraries [22]. Electronic equipment like scanners, digital cameras and mobile phones are used to convert paper documents into their image format. If the documents are stored in image formats, information extraction from these images is challenging problem as it compared with digital texts [17].Information retrieval from document images and converted into their text format by using Optical Character Recognition (OCR) [22].

OCR is a technique, which is used to identify the characters, words and different fonts from document images, then converts these images into text format [4]. OCR analyzes the text from document images and translate character images into character codes, therefore it should be altered, searched, stored more in notepad and word document more efficiently [19]. This technique provides a full alphanumeric recognition of the text from document images [4] [5] [6] [19].It supports different types of image formats like JPG, PNG, BMP, GIF, TIFF and multi-page PDF files. Many different types of OCR tools are available today, but only few of them are open source and free [19]. Achieving 100% accuracy result is not possible, but it is better to have something rather than nothing [7] [8]. To improve accuracy most of the OCR tools use dictionaries, recognizing individual characters then it tries to recognize entire words that exist in the selected dictionary [8]. Sometimes it is very difficult to extract text because different font size, style, symbols, dark background and poor quality of document image often prohibit complete conversion using OCR [8]. If we are using high resolution documents the OCR tools will produce better results [4][22].Hence the output text result of a tools are based on the type of input image.

The remaining portion of this paper is discussed as follows. Section II describes overview of OCR architecture; Section III gives three different types of OCR tools. Section IV provides the conversion accuracy of the OCR Tools. Section IV discusses the performance analysis and conclusion is given in Section V.

II OVERVIEW OF OCR ARCHITECTURE

The block diagram of OCR system consists of various states as shown in Figure 1. They are Tamil and Hindi input document images, pre-processing, segmentation, feature extraction, classification and recognition text.

1. Tamil and Hindi Input Document Images

Tamil and Hindi document image is taken from scanner and captured using camera and mobile phones. It can be stored in different image formats like JPG, PNG, BMP, GIF and TIFF.

2. Pre-Processing

Pre-processing is the most important and essential for better performing OCR system.

***Noise Removal*:** Scanner and camera can produce noise for input document image; hence noise removal is an important factor. Various filtering techniques are available for removing noise from document image. These techniques are used to add or remove noises from the images, maintaining the correct contrast of the image, background removal which contains any scenes or watermarks. These are applied into images which enhance the image quality[12].

Binarization: Input RGB image converted into 0 and 1 form. The place where part of image is shown or it is black is displayed as 1 and the part where image is not show nor it is white is displayed as 0 [5].

Skew detection and Correction: Document images are getting skewed with either left or right orientation. The function for skew detection checks for an angle of orientation between ±15degrees and if detected then a simple image rotation is carried out till the lines match with the true horizontal axis, which produces a skew corrected image.(pdf10)

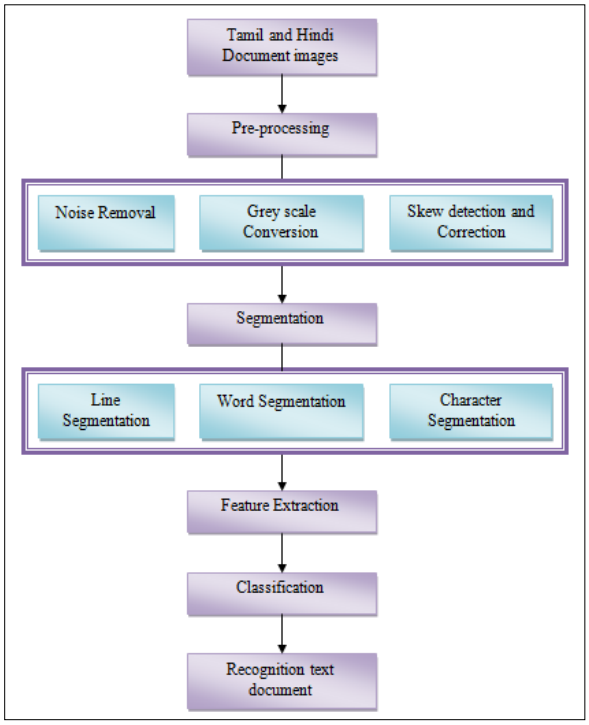

Figure 1 Block diagram of OCR architecture

3. Segmentation

After preprocessing step, output document image is considered as the input of this segmentation step. Segmentation leads to more compact image representations by partitioning an image into a set of disjoint segments to represent image structures. OCR technology using both line, word and character segmentation.

Line Segmentation: Each line in the document images may not be perfectly horizontal, they will not have so much of skew such that there is no inter line gap. Hence the lines are aligning horizontally using line segmentation. For the document image, line segmentation has calculated the number of black pixels in the each row. As both the font size and inter line gap is variable over a large range, there are less chances of the projection profile to have any kind of pattern. Thus, irrespective of number, all the consecutive rows that have a high value at the projection profile are grouped as line rows. All the rows that have a near zero value at the projection profile is considered as inter line gap.

Word Segmentation: Word segmentation is the process of dividing written text into meaningful units, such as words, sentences and topics. A segmented line is selected and split the text region into the units and then finally to words.

Character Segmentation: It is very essential for OCR to recognize the characters in the input document image. Character segmentation is that seeks to decompose an image of a sequence of characters into sub images of individual symbols. If there is a good segmentation of characters, the recognition accuracy will also be high.

4. Feature Extraction

Feature extraction involves extracting meaningful information about an object or a group of object from the document images, features is extracted using character recognition which helps to identify the different characters or words in document images. It used to analyze an input image then recognize characters and words in document image, then translates character images into character codes. Character codes are matched with the existing template, before that exiting templates are stored in a database.

5. Classification

Classification is performed after feature extraction step is done, which is performed as to which class the character belongs to. These features are analyzed using the set of rules and labelled as belonging to different

classes. Various Classification algorithms are performed this task. Classification algorithms are Support Vector Machine (SVM), K-nearest neighbour, neural network, Decision tree classification, etc. After classification the character codes are displayed as a text document.

III Online OCR tools Comparison

The main aim of this work is to extract the text from Tamil and Hindi document images, we taken two different kinds of document images one is Tamil document image another one is Hindi document image, table 1 shows the input images, they are tested with three different types of Online OCR tools. They are Google Docs, Free Online OCR, i2OCR. The extracted Tamil and Hindi document results of different tools are displayed in table 2.

A. 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 including Tamil and Hindi document images. In performance result Tamil image not convert properly, it makes some mistakes and Hindi document image converted perfectly.

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]. It supports both Tamil and Hindi document images, it makes lots of mistakes in Tamil document image and it give accurate conversion in Hindi document image.

C. i2OCR

Converting text from images usingi2OCR,it's free online Optical Character Recognition software. After converting text can be edited, formatted, indexed, searched, or translated [19]. Input image types are TIF, JPEG, PNG, BMP, GIF, PBM, PGM and PPM [21]. It takes unlimited uploads and supports more than 60Languages. It also supports both Tamil and Hindi document images, Tamil image not convert properly; it makes some mistakes and Hindi document image converted perfectly.

Table 1Tamil and Hindi Input document images

	Tamil document image	Hindi document image
Input image	தேடிச் சோறு நிதந்தின்று - பல சின்னஞ் சிறுகதைகள் பேசி - மனம் வாடித் துன்பமிக உழன்று - பிறர் வாடப் பலசெயல்கள் செய்து - நரை கூடிக் கிழப்பருவம் எய்தி - கொடுங் கூற்றுக் கிரையெனப்பின் மாயும் - பல வேடிக்கை மனிதரைப் போலே - நான் வீழ்வே னென்று நினைத்தாயோ ! - மாகவி பாரதியார்	कुछ करने की इच्छा वाले व्यक्ति के लिए इस दुनिया में कुछ भी असंभव नहीं है।

Table 2 Google Docs, Free Online OCR and i2OCR

OCR Tools	Tamil document image	Hindi document image
Google Docs	பிதஷ்பிசாறுதித்தின்று - பல சின்னஞ் சிறுகைதைகள்பெரி - மனம் வரஷத்துக்உயிகூழுன்று - பிறர் வரடப் பலஇசயல்கள்இசய்து - நலர சஷடிக்கிழப்பருவம்எய்தி - இகரடுய் டெற்றுக்கினரஇயர்ப்பின்மரயும் - பல சேபடிக்லகமனிதவரப்பிபாலே ~ நான் ஷசேகினன்றுநினைத்தாபியர 1 - மரகவிபாரதியார்	कुछकरनेकीइच्छकचवालेव्यक्तिकेलिए इसदुनियामेंकुछभीअसंभवनहींहै
Free Online OCR	தேஷச்சேறுகெந்நின்று சின்னஞ்சிறுகைதைகள்டுபெரி . மனி வரனைத்துள்பமிகூழுன்று ~ பிறர் வரடப்பலஇசயல்கள்இஸ்து - நர வநக்கிழப்பருவம்எய்தி . இகரடுய் கூற்றுக்ணெஷணப்பின்மணம் E மை டுவடிக்னகணிதவரப்டுபாலே ... நான் விழ்லேகினன்றுநினைத்தாடியா மரகவிபாரதியார்	कुछकरनेकीइच्छकचवालेटयक्तिकेलिए इसदुनियामेंकुछभीअसंभवनहींहै
i2OCR	பிதஷ்பிசாறுதித்தின்று - பல சின்னஞ் சிறுகைதைகள்பெரி - மனம் வரஷத்துக்உயிகூழுன்று - பிறர் வரடப் பலஇசயல்கள்இசய்து - நலர சஷடிக்கிழப்பருவம்எய்தி - இகரடுய் டெற்றுக்கினரஇயர்ப்பின்மரயும் - பல சேபடிக்லகமனிதவரப்பிபாலே ~ நான் ஷசேகினன்றுநினைத்தாபியர 1 - மரகவிபாரதியார்	कुछकरनेकीइच्छकचवालेटयक्तिकेलिए इसदुनियामेंकुछभीअसंभवनहींहै

IV Performance analysis of Tamil document image

Table 3 shows the Accuracy measures of different three OCR tools. Figure 2 displays the Accuracy and Error rate of different OCR tools. Table 4 presents conversion mistakes in document images.

Table 3 Accuracy measures

S. No	OCR Tools	Character Accuracy (%)	Character Error Rate (%)
1.	**Google Docs**	**6.66**	**93.34**
2.	**Free Online OCR**	**0**	**100**
3.	**i2OCR**	**6.66**	**93.34**

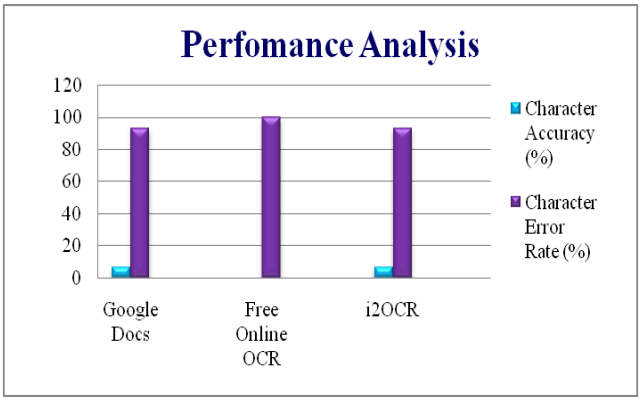

Figure 2 Performance analysis

TABLE 4 Google Docs, Free Online OCR, i2OCR

S. No	Original Text	Google Docs	Free Online OCR	i2OCR
1	தேடிச் சோறு			
2	நிதந்தின்று பல	✔		✔
3	சின்னஞ் சிறுகதைகள்			
4	பேசி மனம்			
5	வாடித் துன்பமிக			
6	உழன்று பிறர்	✔		✔
7	வாடப் பலசெயல்கள்			
8	செய்து நரர			
9	கூடிக் கிழப்பருவம்			
10	எய்தி கொடுங்			
11	கூற்றுக் கிரைபெனப்பின்			
12	மாயும் பல			
13	வேடிக்கை மனிதரைப்			
14	போல நான்			
15	வீழுவே னென்று			
16	நினைத்தாயோ !			
17	-மாகவி பாரதியார்			

V CONCLUSION

Optical Character Recognition (OCR) is a very important in the field of information retrieval. It retrieves the information from the document images and it converted in to text format. Now OCR system can upgraded to support different types language. In this paper discussed the essential characteristics of OCR, architectures, techniques and discuss about Online OCR tools like Google docs, Free Online OCR and i2OCR. The three Online OCR tools to support both Tamil and Hindi document images, hence we analyzed these tools. From this analysis, Google docs and i2OCR are little better than Free online OCR tool, these tools are converted the Tamil and Hindi document image are not properly. In Future, these issues are to be handled by developing new techniques and algorithms.

REFERENCES

1. Amit Choudhary, Rahul Rishi and Savita Ahlawat, "A New Character Segmentation Approach for Off-Line Cursive "Handwritten Words", International Conference on Information Technology and Quantitative Management (ITQM2013).
2. Archana A. Shinde and D.G.Chougule, "Text Pre-processing and Text Segmentation for OCR", IJCSET, Vol2.
3. Chirag Patel, Atul Patel and 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
4. 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.
5. "Feature Extraction Technique", International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 8, August 2015.
6. http://en.wikipedia.org/wiki/Optical_character_recognition
7. http://www.i2ocr.com/
8. http://www.labnol.org/software/convert-images-to-text-with-ocr/17418/
9. http://www.newocr.com/
10. http://www.slideshare.net/ijujournal/performance-comparison-of-ocr-tools
11. https://docs.google.com/
12. Nisha Goyal and Er. Shilpa Jain, "Optimized Hindi Script Recognition using OCR
13. Oivind due trier, Anil K.Jain, TorfinnTaxt, "Future extraction methods for character recognition A survey".
14. Pranob K Charles, V.Harish, M.Swathi, CH. Deepthi "A Review on the Various Techniques used for Optical Character Recognition", International Journal of Engineering Research and Applications (IJERA), ISSN: 2248-9622, Vol. 2, Issue 1,Jan-Feb 2012.
15. 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.
16. 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.
17. Seethalakshmi R, Sreeranjani T.R., Balachandar T., "Optical Character Recognition for printed Tamil text using Unicode", Journal of Zhejiang University SCIENCE, ISSN 1009-3095.
18. Shalin A. Chopra, Amit A. Ghadge, Onkar A. Padwal, Karan S. Punjabi and Gandhali S. Gurjar "Optical Character Recognition",International Journal of Advanced Research in Computer and Communication Engineering, ISSN (Online) : 2278-1021 ISSN (Print) : 2319-5940, Vol. 3, Issue 1, January 2014
19. Shivani Dhiman and A.J Singh, "TesseractVsGocr A Comparative Study",International Journal of Recent Technology and Engineering, ISSN: 2277-3878, Volume -2, Issue-4.
20. Yasser Alginahi, "Preprocessing Techniques in Character Recognition"
21. Youssef Bassil and 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.
22. Dr. S.Vijayarani and Ms. A.Sakila, "A Survey on Word Spotting Techniques for Document Image Retrieval", International Journal of Engineering Applied Sciences and Technology, July 2015.