A Review on OCR Methodology

Mr. Prabhojan Pashte[1] Mr. Samir Kerawdekar[2] Mr. Pankaj Bait[3] Prof. P. G. Magdum[4]
[1,2,3]Student [4]Associate Professor
[1,2,3,4]Department of Computer Engineering
[1,2,3,4]Rajendra Mane College of Engineering and Technology, Ambav, Ratnagiri, India

Abstract— Optical Character Recognition or OCR is the process of taking an image of letters or typed text and converting it into data that computer understands. The exact mechanism that allows humans to capture an image of a paper document after which the text is extracted from that image. Hence, paper documents are easily converted into editable computer files. OCR is widely used in the field of pattern recognition and artificial intelligence. The paper describes the detailed methodology in the field of Optical Character Recognition.

Key words: Optical Character Recognition, OCR, Methodology, digitization

I. INTRODUCTION

Optical Character Recognition is a technology that involves photo scanning of the text character-by character, analyze it, and then the transformation of the character image into character codes. A good example is the constructor of a building, taking physical copies of blueprints as an input and using OCR technique put them onto computers for further work. At the present moment, OCR is the best method for digitizing typed pages of text.

Optical Character Recognition is broadly divided into two parts, offline recognition and online recognition. Offline recognition deal with the system where input is either an image or a scanned form of the document whereas in online recognition the successive points are represented as a function of time and the order of strokes are also available [9]. In this paper we are dealing only with offline recognition technique.

There are three basic techniques on which OCR works: Pre-processing, Character Recognition, and Post-processing. Before going directly to character recognition, it is needed that image we are going to used must be error free. So that accurate recognition can possible. Pre-processing implies changes if the image is not properly aligned, edges are not smooth, detects line and character, converts colour images into black-and-white images, etc. After compiling all changes to input image file, we can move to character recognition.

OCR works on two methods. One of the methods is Matrix Matching. The technique compares what the OCR scanner sees as a character with prescribed character templates. When an image matches one of these library templates within a given level of similarity, the computer marks that image as the corresponding ASCII character.

Feature extraction, another method where OCR act without strict matching to library templates. This method deviates by how much computer intelligence is applied by the manufacturer. The computer looks for general features such as disconnecting edges, perfectly closed shapes, diagonal lines, intersecting lines, etc. This method is much more versatile than matrix matching. Matrix matching works best when the OCR discovers library templates of characters with whom character get to compare. Whereas if there are no library characters, feature extraction is higher-ranking [1].

After character recognition stage, if there any unrecognized character found, that character get meaning in post-processing stage. Post-processing takes care of few things such as knowledge of the grammar for allowing great accuracy. The Levenshtein Distance algorithm has also been used in OCR post-processing to further optimize results from an OCR API [2].

II. FLOW OF OCR SYSTEM

In OCR processing, an input image or file is first to browse in the computer by scanning. Scanning speed will be decided by the quality of the scanner machines, paper quality, cleanness, proper setting of the OCR system.

After scanning, it reads that file, decides the threshold value, analyzed for light and dark areas in order to identify each alphabetic letter or numeric digit. When a character is recognized, it is converted into an ASCII code. The recognizing process is to interpret images. The library templates and configuration threshold will determine the accuracy of interpretation of the OCR.

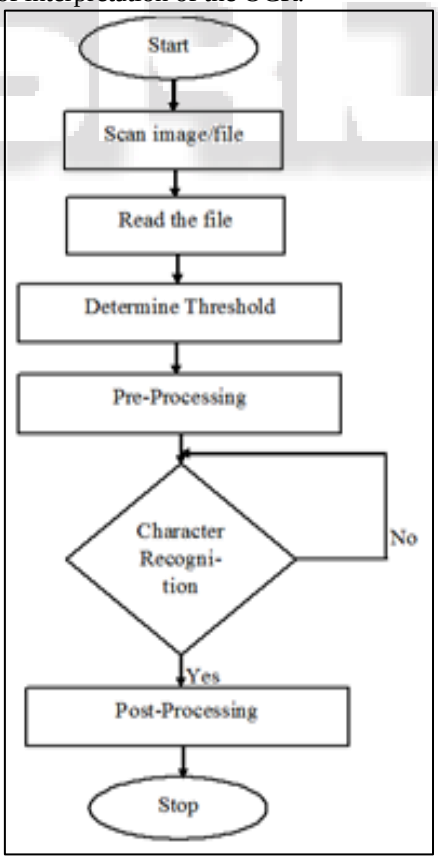

Fig. 1: Flowchart of OCR System

III. LITERATURE REVIEW

In paper [3], the author has discussed various applications of OCR and presents the experimentation for three applications such as Captcha, Institutional Repository, and Optical Music Character Recognition.

In paper [4], the author has presented a complete methodology for handwritten historical documents without having much knowledge about fonts. Nowadays since everything becomes digital it is necessary that all that we have today in handwritten form convert it into digital form for safety purpose as well as editing purpose.

In paper [5], the author has presented brief study on OCR terminology. The paper describes how OCR was evolved and how its applications growing day by day. OCR can recognize characters by online or offline processing. Proposed OCR system is based on grid infrastructure, grid infrastructure is that infrastructure which supports specific set of languages.

In paper [6], the author has presented a way to construct OCR which can read any document that has fixed font size and style or handwritten style. Simultaneously this approach will also be simple, efficient, and less costly. To achieve efficiency and less computational cost, OCR in this paper uses database to recognize English characters which makes this OCR very simple to manage.

In paper [7], the author has proposed a special kind of OCR system which converts images that contain Arabic text to a format that can be edited. The OCR is capable to produce exact output for different sizes of Arabic text.

In paper [8], the author has presented brief study on the open source OCR tool Tesseract. It also provides comparatively study of Tesseract with Transym, commercial OCR tool. Basically, Tesseract is command based tool where input is provided in the form of image having text in it. It works well and gives accurate and expected results while scanning greyscale images whereas in case of complex color images, output is not as accurate as expected.

IV. CONCLUSION

We all are aware about the fact that India is rapidly transforming into digitization with governments 'Digital India' initiative. Almost all of the government procedures are paper-driven, which creates major concerns of security, inability to store huge information, susceptibility of human-prone errors, etc. To remove all these, Indian government plans to establish new paper-free 'digital' government. Optical Character Recognition helps to achieve that feat. It is very useful and popular method of transforming text-images into digital form.

In this paper we have studied several papers, some are review papers while other are application-based papers. The whole paper focuses on methodology of OCR, how it works to produce accurate output. No doubt, OCR is growing industry and has its own benefits and drawbacks such as complex color text images cannot produce expected output. Though it has some limitations, it has wide range of applications too.

ACKNOWLEDGEMENTS

We, authors gratefully thankful to Dept. of Computer Engineering, RMCET and extend gratitude towards our project guide and coordinator, Prof. Magdum sir for his valuable advice and showing faith in us throughout the process. We are also thankful to Prof. M. M. Gadkari (HOD, Dept. of Computer Engineering, RMCET) for his support and motivation throughout the year.

REFERENCES

[1] "OCR Introduction", Available: http://www.dataid.com/aboutocr.htm. [Accessed: March 31, 2017].
[2] "How to optimize results from the OCR API when extracting text from an image?", Available: https://community.havenondemand.com/t5/Wiki/How-to-optimize-results-from-the-OCR-API-when-extracting-text/ta-p/1656. [Accessed: March 31, 2017].
[3] Amarjot Singh, Ketan Bacchuwar, and Akshay Bhasin, "A Survey of OCR Applications", International Journal of Machine Learning and Computing, Vol. 2, No. 3, June 2012.
[4] G. Vamvakas, B. Gatos, N. Stamatopoulos, and S.J.Perantonis, "A Complete Optical Character Recognition Methodology for Historical Documents", 8th International Workshop on Document Analysis System (DAS'08), pp. 525-532, Nara, Japan, September 2008.
[5] Najib Ali Mohamed Isheawy And Habibul Hasan,"Optical Character Recognition (OCR) System" IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 17, Issue 2, Ver. II (Mar – Apr. 2015), PP 22-26.
[6] Shalin A. Chopra, Amit A. Ghadge, Onkar A. Padwal, Karan S. Punjabi, Prof. Gandhali S. Gurjar," Optical Character Recognition", International Journal of Advanced Research in Computer and Communication.
[7] Abdelwadood Mesleh, Ahmed Sharadqh, Jamil Al-Azzeh, MazenAbu-Zaher, Nawal Al-Zabin, Tasneem Jaber, Aroob Odeh and Myssa'a Hasn," An Optical Character Recognition", Contemporary Engineering Sciences, Vol. 5, 2012, no. 11, 521 – 529.
[8] 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, October 2012.
[9] "Document Analysis and Recognition", 2005. Eighth International Conference on 29 Aug.-1 Sept. 2005, Alon

