A Review on OCR Technology

Jay Dilipbhai Thanki, Priyank Dineshbhai Davda, Dr. Priya Swaminarayan

Student, Department of MCA, Faculty of IT & Computer Science, Parul University, Vadodara, Gujarat, India

Student, Department of MCA, Faculty of IT & Computer Science, Parul University, Vadodara, Gujarat, India

Dean, Department of MCA, Faculty of IT & Computer Science, Parul University, Vadodara, Gujarat, India

Abstract: Optical Character Recognition (OCR), is that the process of conversion of image text or handwritten text into machine understandable form. Simply OCR means conversion of characters that is recognized and convert it into computer readable form. It is widely used as a kind of data entry from original paper data sources such as banking papers or consultation papers, whether passport documents, invoices, statement, receipts, card, mail or any number of printed records. It is a standard method of digitizing printed texts in order that they will be electronically edited, searched, and stored more compactly. OCR is the field of research in Pattern Recognition, Artificial Intelligence and Computer Vision. OCR is that the electronic translation of handwritten, type written or printed text into machine translated images. It is widely used to recognize and search text from documents or to publish the text on a website. This document represents review of Optical Character Recognition methods such as Correlation of Character Recognition, Pre-Processing, Segmentation, Neural Network and Structural Components Extraction and discuss their advantages and disadvantages. Through Neural Network method, Optical Character Recognition typing error can be solved which may increase the efficiency of conversion rate. This paper represents all the techniques and algorithms that is used to find accurate results for OCR.

I. Introduction

Optical Character Recognition is apart from two different terms or words that is Optical and Character Recognition. Optical Means which is refers to anything that will relate to light or a vision. A Computer Mouse is a best Example of an optical Device that uses Optical technology. It uses LED and photodiodes to work out the direction that mouse is moving across the surface. Secondly, Optical Character Recognition is that the mechanical term and electronic conversion of scanned or photograph images of typewritten or printed text into machine readable text.

Optical Character Recognition is active research area that attracts every human brain which has the capability to recognize text or patterns very easily from images. These days there is a huge trend and demand for Cloud storage and digitalization for storing sensitive information to a computer storage disk or at cloud to later re-utilize the things by means of computers. One simple way to store information at cloud is to scan and saved, but it would be more time consuming and not affordable to every organization. So, OCR comes with the concept of same algorithm that we followed manually since last so many decades. OCR helps us in such a way that it not only just scanned image or documents, but it has the capability to store on cloud platform as well as it can convert everything to computer understandable format so that in future, if we need any information from anywhere, by giving very few commands we can get our information back for our purpose.

Since the OCR research is an active and very important field in general pattern recognition problems, because of its fast progress and comprehensive fields of reviews are needed on a regular basis to keep track of the new advancements [7].

This paper will introduce the reader to elaborate various kind of studies by providing a comprehensive literature review of optical character recognition research. It humanizes the reader with major challenges and main phase of optical character recognition, applications, and techniques of OCR.

II. Literature Review

Optical Character Recognition is not a new problem that we research on, but its root elements can be traced back to the systems before the inventions of computers. The earliest and oldest OCR system were not that much capable as latest generations computers but mechanical devices that were able to recognize the characters, but very slow speed and low accuracy.

In paper [1], the author has discussed various issues that were considered as a major issue while checking performance of character recognition system. Also, this paper has more focused on Neural Network that is basically the main root of OCR System that we follow now, and they have also mentioned how NN will help people to resolve their problems and improves accuracy rate of result.

In paper [2], the author has drawn the attention towards the application areas where OCR applied. Some of the major application areas are CAPTCHA, Institutional Repository and Optical Music Character Recognition. In this paper, they mentioned only those applications which is only used to recognize text.

In paper [3], the author has presented a very interesting and unique problem to identify text from the documents which is OCR for cursive handwriting. To recognize text from it, author has divided all the task into different segments and give them a purpose accordingly. The paper describes how OCR will first normalize the text and the give accurate result from cursive handwritten document or printed document as well.

In paper [4], the author has presented brief study on Car Parking Control System using OCR technology or device. Nowadays in many countries there are lack of parking space availability or people must have to wait for their space. So, to overcome that problem Car Park Control System developed using OCR that can display the parking information on LED or LCD screen. This paper also explains in brief all the segments that how this OCR will work in this real-life problem.

In paper [5], the author has explained in brief about OCR terminology. The paper described in detail how OCR was developed gradually and how its applications are growing day by day. OCR can recognize characters through online or offline methodology. Proposed OCR system in this paper is based on grid infrastructure that supports specific set of languages.

In paper [7], the author has tried to explain in brief and summarizing the research so far done in the field of OCR. It provides a summary of various aspects of OCR and discuss corresponding proposals aimed to resolve issue in OCR system. As this paper has also drawn the attention towards the fundamentals and technology behind OCR's work. Author's main purpose was to represent various technologies that are already using in OCR, so they have represented it in a best possible way. Author has also mentioned some of the major real-life applications that OCR used in like, Number Plate Recognition, Smart Libraries, and various other real time applications too. Despite of the significant amount of research in OCR, recognition of the characters for languages like, Arabic, Sindhi and Urdu remains open challenge till now. In this paper, author also covers the methods to recognize characters from documents or images i.e., Image Analysis.

In paper [8], the author has tries to explain about Optical Character Recognition more deeply that it is the process of taking an image of letters or typed text and converting it into data that computer can understands. According to author, OCR can be used widely in the field of Pattern Recognition and Artificial Intelligence. This Paper also describes detailed methodology in the field of Character and Text Recognition. Author also tries to explain some of the markable and notable point in such a way that if anyone is not coming from Information Technology field, then He/She can also understand the terms that author has used. I read about OCR techniques that how it will work: Pre-Processing, Character Recognition, and Post Processing. Author has also explained the term called Matrix matching which means 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.

.

III. APPLICATION AREAS

Some of the Applications areas in which Optical Character Recognition can take place and which is also currently under research and development are as follows:

I. **Data Entry [4]:**

covers technologies for entering large amount of restricted data. Initially such machines were used for banking applications. The system is characterized by reading only limited set of printed characters usually numerals and special symbols. They are designed to read data like account numbers, customer's identification, article numbers, amount of money etc [4].

II. **Aid for Blind [9]:**

In the youth before digital computers and requirements for input of huge amounts of knowledge emerged this was an imagined application area for reading machines. Along with speech synthesis systems such reader enables blind to understand printed documents [9].

III. **Automatic Cartography [4]:**

Optical recognition from maps presents special problems within character recognition. The symbols are intermixed with graphics, text is printed at different angles and characters are of several fonts or even handwritten [4].

IV. **Form Readers [4]:**

Many systems can read specially designed forms. In such forms all irrelevant information to reading machines is printed in color invisible to scanning device. The characters are in printed or handwritten Uppercase letters or in specified boxes. The processing speed is dependent on amount of data on each form but maybe few forms per minute [4].

V. **Signature Verification and Identification [9]:**

This application is useful for banking environment. Nowadays, online transaction is taking place of offline transection because of digital India concepts. So, parallelly online fraud can also be taken place while transecting something. So, to reduce that type of fraud we can verify the user or verify their signature through OCR. As soon as the customer put digital signature, it will verify through the OCR. The signature is simply considered as pattern which is matched with signatures stored in database [9].

VI. **Legal Industry [9]:**

Legal Industry is likewise one of the recipients of the OCR innovation. OCR is utilized to digitize documents, and to specifically enter PC database. Legitimate experts can further search documents required from tremendous databases by basically writing few keywords [9].

VII. **CAPTCHA [8]:**

A CAPTCHA is a system that can create and grade tests that human can pass yet current software technology cannot. In other words, in CAPTCHA, a picture comprising an arrangement of letters and numbers is produced with variety of size and textual styles, highlights and noise so that text cannot be read via OCR. Current OCR frameworks are often utilized to evacuate the noise and portion the image to form the image tractable by such malicious users [8].

VIII. **ATMA (Android Travel Mate Application) [10]:**

ATMA: android travel mate application is proposed by Mishra, Nitin, and C. Patwardhan, that it empowers Tourists and Travelers to effortlessly catch the native signboards, nation dialect Books pages, hotel menus, banners and so on. Unicode text format was obtained from content embedded within the caught image by an implicit OCR. With the goal that travelers can translate native Dialect Unicode content into their own nation dialect, it likewise gives translation features [10].

IV. TYPES OF OPTICAL CHARACTER RECOGNITION SYSTEM

There have been tons of directions in which research and improvements on OCR that is already being carried out since many years. The reason behind taking this section highlighted is to discuss different types of OCR systems that have emerged because of this research. To identify these systems and to analyse those systems in a quickie way, we can categorize these systems based on Image acquisition mode, Character connectivity, font-restrictions etc. the following Fig.1 categorizes the Character Recognition System Types.

Based on type of input, the OCR can be further divided mainly into two segments formerly, first is Handwriting Recognition and Machine Printed Character Recognition.

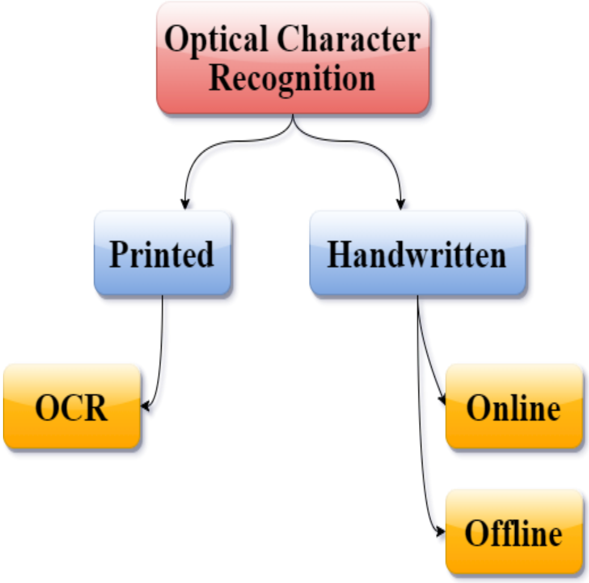

Figure 1: Types of OCR

The above figure indicates types of OCR Systems that usually uniforms that how OCR will be considered while converting via online or offline mode, and the positions of characters on the page can be predicted.

Handwriting character recognition sometimes turn to be very tough job due to different writing style of users as well as different pen movements of the user for some typical character. So, we have separated handwriting into online and offline systems to mainly checked performance in real time while the users writing one by one character. The online mode of written document is less complex as they can capture the temporal or time-based information. The offline recognition systems operate on static data i.e., the input is electronic image. Therefore, it is very difficult to perform recognition process [6].

V. ADVANTAGES OF OPTICAL CHARACTER RECOGNITION SYSTEM

There are several reasons for choosing OCR System over other Method for Scanning. It is as follows:

- **Make your Immutable files Searchable [11]:**
  When you have lots of PDF files and textual electronic images, the Information they contain is not searchable and editable. That means you have lots of frozen text. This kind of information's affects your ability to search anything. The OCR technology enables you to convert frozen text into machine-readable data so that it can be searchable [11].
- **Make your Edits Easy [12]:**
  OCR is a tool that makes your business operations more adaptable to changes. Convert those unalterable files into text documents that are easily be edited. You need to have OCR to convert those documents into editable text documents to make it easier for editing [12].
- **Prevent Human Errors [11]:**
  It is common that every Human may make mistakes, but you should be able to make changes. OCR not only enables you to edit and search inalterable documents but also detects the incorrect or misprinted information in document. So, all human error can be proactively resolved using OCR technology [11].
- **Save Time & Money [13]:**
  Most business still have a lot of documents in hard form. OCR will significantly reduce the time and money spent on manually entering the data into the computer. You can simply use the OCR to scan printed document or image to obtain editable document [13].
- **Save Space [12]:**
  You will also save a lot of your office space that has been occupied by piles of paper documents. You can digitalize all your paper documents with the help of OCR and free some space in your office. So, less paper translates into more order and space in your workspace [12].

VI. USES OF OPTICAL CHARACTER RECOGNITION SYSTEM

OCR engines have been developed into many kinds of domain-specific OCR Application, Such as Receipt OCR, Invoice OCR, Check OCR, Legal Billing Document OCR. Some of the ways in which Optical Character Recognition (OCR) can take place are as follows:

- Data Entry for Business documents, E.g., Cheque, Passport, Invoice, Bank Statements and Receipts [9]
- Automatic Number Plate Recognition [10]
- In Airports, for Passport Recognition and Information Extraction [11]
- Automatic Insurance Documents Key Information Extraction [12]
- Traffic Sign Recognition [8]
- Extracting Business Card information into a Contact List [10]
- More quickly make textual versions of printed documents E.g., Book Scanning [13]
- Make electronic Images of printed documents searchable E.g., Google Books [11]

VII. MODULES OF OPTICAL CHARACTER RECOGNITION SYSTEM

An OCR Module is a unit containing OCR software along with other features. An OCR module can be a simple software which can convert paper documents into searchable electronic files or can be a complex software with additional functionality [11]. The Modules that were identified in the Optical Character Recognition System are as follows:

- Image Acquisition
- Pre-Processing
- Segmentation
- Feature Extraction
- Training a Neural Network
- Post Processing

Graphical Representation of OCR Modules are as follows:

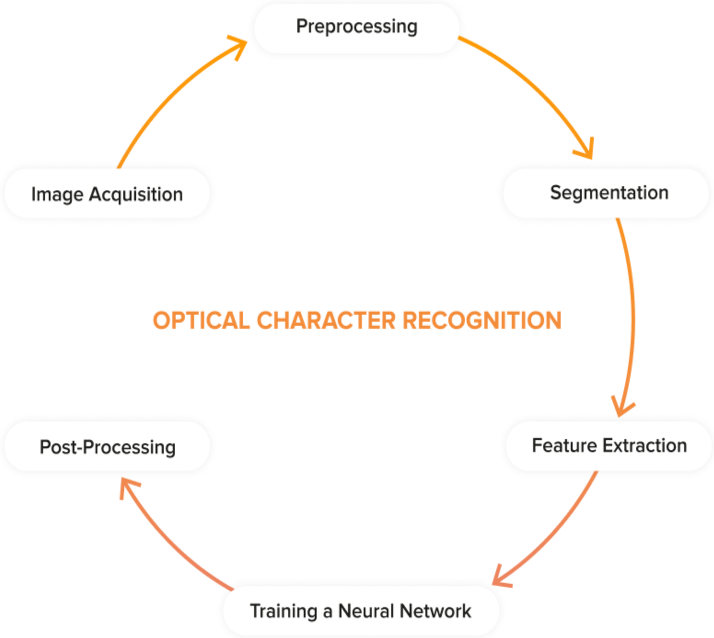

Figure 2: Modules of OCR

1. **Image Acquisition:**
   Image acquisition is the initial step of OCR that comprises obtaining a digital image and converting it into suitable form that can be easily processed by computer. This can involve quantization as well as compression of image [8]. In most of the case, the binary image suffices to characterize the image suffices to characterize the image. So, Image acquisition is only converting the Image or document into digital format and store it into computer.
2. **Pre-Processing:**
   Next to image acquisition is pre-processing that aims to enhance the quality of image. One of the pre-processing techniques is thresholding that aims to binaries the image based on some threshold value [9]. The Approaches for pre-processing are to Noise removal, skew removal, and thinning. With the help of this techniques, image or document will only improve the quality and arrange all the data into qualitative manner. So that it will be understandable by human and computer both.
3. **Segmentation:**
   In this step, the image is segmented into characters before being passed to classification phase. The segmentation can be performed explicitly or implicitly as a by-product of classification phase [11]. In addition, other phase of OCR can help in providing contextual information useful for segmentation of image. In addition, character or image is identified into multiple segments and will passed it to the next stage i.e., features extraction.
4. **Feature Extraction:**
   In this stage, various features of characters are extracted. These features uniquely identified characters. The selection of the right features and the total number of features to be used is an important research question [11]. Different types of features such as the image itself, geometrical features and statistical features can be used. Finally, various techniques such as principal component analysis can be used to reduce the dimensionality of the image [11].
5. **Training a Neural Network:**
   Neural Network is a wonderful tool that can help to resolve OCR type problem. For training the neural networks we use the ‘vector’ generated by the ‘Database Templates’ using the above-mentioned features extraction techniques. It may be noted that neural networks use Backpropagation algorithm for learning. The ‘Target’ values are specified by the system programmer to accommodate for small recognition errors, which may be changed to application to application [13].
   The neural network was trained for 1000 iterations, which took around 21 seconds to complete. But system needs to calculate the effect of joint errors in all the parameters, rather than overall error. An error goal of 0.0001 or 0.01% was achieved by the Neural Network technique [13].

6. Post Processing:

Once the character is classified under neural network, there are various approaches that can be used to improve the accuracy of OCR result. To improve OCR result, contextual analysis can be performed. The geometrical and document context of the image can help in reducing the chance of errors. Another method to improve the errorless information is to verify through various stages like, digitalize the image, convert it into binary format or ASCII Code, then arrange it into different segment and then if that image or document contains error it will be resolve through the Neural Network techniques.

CONCLUSION:

We all are known with the fact that India is growing fast to transforming into digitalization with the government's slogan 'Digital India' Initiative by Hon. Prime Minister of India Shree Narendra Modi. Almost all the government procedures and transaction are paper driven, which may create major concerns of security and safety to store larger amount of information with respect to susceptibility of human prone errors etc. To overcome all these things, Indian government are working on new digital era which is 'Paper Free' digital India. So, OCR is one of the easier and most safe way to achieve that feat. It is very useful and popular method of transforming text or images into digital form.

In this paper, we try to study several papers, some are review papers while other are practical based papers. The whole paper focuses on Methodology that OCR follows and how it works to produce accurate output. Also, we have seen types that OCR follows at the very basic stage. Although it has some limitations, but overall OCR has its own wide range of applications too that can give accurate output for given input.

REFERENCES:

Journal Papers:

1. Shyla Afroge, Boshir Ahmed, Firoz Mahmud "OCR Using back Propagation and Neural Network", IEEE (ICECTE) 2016.
2. 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.
3. Nafiz Arica, Fatos T. Yarman – Vural, "Optical Character Recognition for Cursive Handwriting", IEEE (Transaction on Pattern Analysis) 2020.
4. Anton Satria Prabuwano, Ariff Idris, "A Study on Car Park Control System using Optical Character Recognition", IEEE (CoVisBot Lab.) 2018.
5. Najib Ali Mohamed Isheaway 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. Qadri, M.T., & Asif, M, 2009, "Automatic Number Plate Recognition System for Vehicle Identification Using Optical Character Recognition" presented at International Conference on Education Technology and Computer, Singapore, 2009, Singapore: IEEE.
7. Noman Islam, Zeeshan Islam, Nazia Noor, "A Survey on Optical Character Recognition System" Journal of Information & Communication Technology – JICT IEEE, Vol. 10 Issue. 2, ISSN: 2409-6520, December 2016.
8. Prabhojan B Pashte, Rajendra Mane, Pankaj T Bait, "A Review on Optical Character Recognition Methodology", IJSRDV5I20835, Volume: 5, Issue: 2, 01/05/2017, Pages: 1049-1050.
9. Abin M Sabu, Anto Sahaya Das "A Survey on various Optical Character Recognition Techniques", IEEE Conf (ICEDSS)2018, DOI: 03.03.17/ICEDSS.2018.9781538634790/18, Page 152-155.
10. Nallasamy Mani & Paul Voumard "An Optical Character Recognition using Artificial Neural Network", IEEE 1996, DOI: 0-7803-3280-6/96, Page 2244-2247.
11. Jianhong Xie "Optical Character Recognition Based on Least Square Support Vector Machine", IEEE Conf (IITA)2019, DOI: 10.1109/IITA.2009.327, Page 626-629.
12. P.A. Khaustov, V.G. Spistyn, E.I. Maksimova "Algorithms for OCR Based on Structural Component Extraction", IEEE Conf (IFOST)2016, DOI: 978-1-5090-0855-1/16, Page 355-358.
13. Tan Chiang Wei, U. U. Sheikh, Ab Al-Hadi Ab Rahman "Improved Optical Character Recognition with Deep Neural Network", IEEE Conf (CSPA)2018, DOI: 978-1-5386-0389-5/18, Page 245-249.

Web References:

"OCR Introduction", Available: http://www.dataid.com/aboutocr.htm [Accessed: October 21, 2020].