A study of optical character recognition techniques

[1] Ms. Puja Suresh Deokate, [2] Prof. Yuvraj Vijay Parkale

[1]P.G.Student, E&TC Departement, [2]Assistant Professor, E&TC Departement

S.V.P.M's College of Engineering Malegaon (BK) Baramati, India SPPU, Pune

Abstract

Document text recognition uses a concept called OCR (optical character recognition), There is an identity card written by a computer or written text description. This includes scanning documents with text and converting characters into their digital form by character. Thus, it is defined as the process of digitizing a document image into its constituent characters. Equipment used to obtain clearer images for analysis are cameras and flatbed scanners. Although OCR technology is still out of the world since 1870, OCR technology still can not reach fulfillment. This form of demand for optical character recognition has been done by various researchers, industry and technology enthusiasts to draw attention to this area. In recent years, there has been significant increase in the number of people who invested in research and time and efforts of companies in this field. This research summarizes the progress, different aspects, and various issues related to this area. It is intended to present a strange look at various proposals, progress and discussion with the aim of resolving various problems that arise in the traditional OCR.

Keywords- Optical character recognition(OCR), OCR techniques, Text to speech(TTS).

I. Introduction

Optical Character Recognition (OCR) is a process by which computers can access text character input by providing computers with images. The computer uses an OCR engine - presents an image (computer-identifiable) image (human identifiable) with a specific task to predict a computer program. Optical Character Recognition (OCR) is a part of the algorithm in which the text containing the text and the printed message changes in digital form which can be handled by the machine. The human neurological system has been amazed that it has the ability to understand the meaning of samples mentioned. This particular field is always an initiative, especially when compared to the machine, which trumps the brain in mathematical and generally general compositions, which have traditionally been a binary system. When the brain recognizes the photograph and its contents with any handwritten or printed material, the images are not able to evaluate the information provided in the image.

OCR reads out damaged or under-quality codes and best guesses what the code is about. Widely used in the form of print documents, passport documents, invoices, bank statements, business cards, mails, computerized receipts, printout of static data or any appropriate documentation. As a result of this, great discipline and many attempts have been

made to bring a major hindrance in the images of typewritten or handwritten character into inefficiency of the cave to transform the image of the document seen by the machine.

If using Projection Profile-based method, text in the text makes it easy to distinguish the contents of text in the language, description, text, from document images separately. Different methods are used at each central level of the OCR. Text segmentation is done using the broadcast profile method. They proposed algorithms to correct the skin corners of the text document [4]. Blur is an important factor in damaging OCR purity. This paper offers a prediction forecast based on local obscure predictions. The relationship between blur effect and character size is checked which is useful for classifier [5].

The grading system is used to evaluate the performance of printed text using different quality measures. The identification result showed high identification rates because the system was capable of enabling the approval rate of 98.69%, including the sensitivity of 0.9857 precision and 1's sensitivity [6]. This paper offers a complete OCR system for capturing embedded text documents in capturing images / graphics in the camera[7]. It described skew detection and correction of scanned document image written in assamese language using vertical and horizontal projection profile analysis[8].

II. CHARACTER RECOGNITION TECHNIQUES

OCR translates machine editable format. character recognition is the process of classifying input letters according to a predefined alphabetical order. due to the increasing interest of computer applications , modern communities need to read text to the computer. this text may be scanned by handwritten documents or text typed in various fonts or in conjuction with both. The valid identity methods helps in communicating between human and computer.

2.1 Types of character recognition

Printed- OCR typewritten text targets one character at a time. This is accomplished using pattern matching and feature analysis. ICR on the other hand may also target handwritten text and makes use of machine learning techniques.

Handwritten- Convenient handwriting input from sources like paper documents, photos, touch screen, and other devices has the potential for computer capability and meaning. The image of the written text may be sensed 'off line' from a piece of paper by optical scanning.

Offline mode is the process of static document, where as the online version is more advanced and uses handwriting movement analysis. Instead of using the commonly used pattern learning algorithm, online mode allows you to capture motion, e.g segment drawn sequence and their direction of stoke. Irrespective of the type of character recognition, the documents are converted to a grayscale format. since this basic grayscale conversion isn't enough various other processing techniques are used to allow better recognition of characters and letters. Identifying words and characters handled by OCR when devices with optical media receive data, they contain devices like cameras and scanners. OCR ideally deals with pixelated data of two types and those types can be divided on the basis of whether the text is handwritten OCR or printed OCR. The symbols and signs can be of any size and orientation as a handheld camera/device won't provide the accuracy of flatbed scanner and handwritten text varies from person to person. Considering the variety available in handwriting types and various educational styles around the world, it is easy to prove that it is challenging to apply a handwritten OCR in comparison to the printed OCR, in which the image is custom-made to be processed, according to the fonts it records. They present a homogeneous pattern as compared to the former.
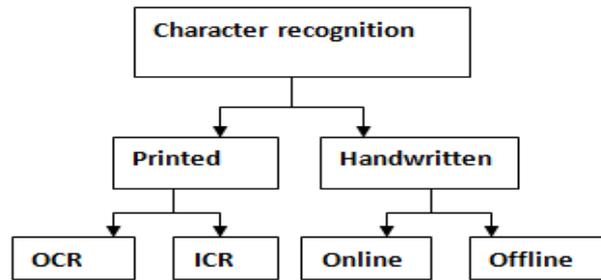
Fig.1.Types of character recognition

Drawbacks of OCR systems summarized below:

1. Considering the variety available in handwriting types and the various educational styles around the world, it is easy to prove that it is challenging to apply a handwritten OCR in comparison to the printed OCR, in which the image is custom-made to be processed, according to the fonts it records.

2. Due to various difficulties like shadows, skiing, blurred lines, expensive pre-processing algorithms are essential for this kind of transformation, which can cause any kind of bug before the model can apply to any type of character conception or approval phase.

3. The identities they have made have been made on the already existing data. They are not dynamic or in real time.

4. When the height of the characters is less than 20 pixels, the accuracy of popular systems such as Tesseract and OCR space drop is much less.

III   Proposed system

Optical Character Recognition (OCR) is a process for converting a printed document or scanned page into ASCII characters that can detect computers. A well equipped computer system with such OCR system improves the speed of input operations, reduces some human error and enables compact storage, faster recovery and other file handling. Postal code identification in the category of applications, large administrative systems include automatic data entry, banking, automatic cartography and reading aids for blind. Accuracy, flexibility and speed are the main features that describe the good OCR system. Many algorithms have been developed to identify feature-based characters.
Some of them are commercially viable and have gone into production like Omni Page, Word Scan, Type Reader etc. Dependencies on font, size, and orientation have reduced the performance of the system. The dependency rate in this algorithm depends on the choice of features.
Most existing algorithms contain a detailed process on the computer before the properties increase, so the computer duration increases. In this paper we discuss a methodology that identifies a characteristic, which will effectively reduce the time when processing the image while maintaining efficiency and versatility. The parallel computing capability of neural networks

ensures high speed recognition for commercial conditions. The main components included in the implementation are: Best choice of features which explains details of characters, number of features and less image processing time.
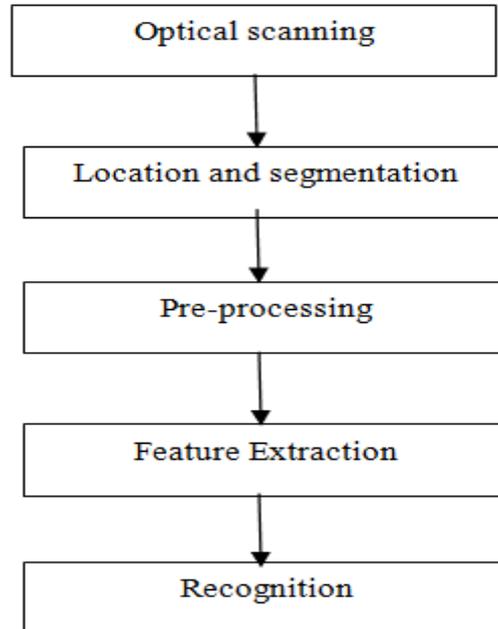

Fig.2. OCR process

1. Optical Scanning-

The optical scanning process involves obtaining a digital image of the original document. Converts light intensity to Gray-level to the OCR optical scanners used. This process is known as thresholding. Thresholding converts multilevel image to black-and-white bit-level image.

2. Location and segmentation-

It involves isolation of character and word. Segmentation requires the separation of letters or words. The location of the text is done using pixels with the x and y index.

3. Pre-processing-

The resulting image can have some noise in the scanning process. Depending on the scanner resolution the characters may be broken or memorized. That's why preprocessing can be done to paste the digitized character. In addition to smoothing preprocessing also involves normalization of the characters. The objective of post-processing is to correct errors or resolve ambiguities in OCR results by using contextual information. There are a number of levels at which context may be operative. It can be at the word level, at the sentence level and at the level of semantics. The most common post-processing technique which operates at the word level is the dictionary look-up method. The output of the OCR is compared to the system's built-in dictionary (lexicon) and candidates are generated. According to the output of OCR output and dictionary lookup, the number indicating the level of confidence in the correct classification has been improved.

4. Feature extraction-

This technique is used for capturing the essential characteristics of the symbols. Feature extraction is done by matching the matrix containing the input character with a set of prototype characters that represent each possible class. The objective of feature extraction is to capture the

essential characteristics of the symbols, and it is generally accepted that this is one of the most difficult problems of pattern recognition. The most direct way to describe a character is the actual raster image. Another feature is to remove some features that still show signs but leave important features.

5. Recognition-

Identity is the process of identifying each letter and assigning it to the appropriate character range.

IV   Text to speech synthesizer

A Text-to-Speech Synthesizer is an application that converts text into speech language and converts text using natural language processing (NLP) and then converts it into synthetic text synthesized speech using digital signal processing (DSP) technology. There are two main steps in the text-to-speech (TTS) synthesis process. The first is text analysis, where input text is written in a phonetic or other lingual form and creates another speech waveform, where the output is created from phonetic and potential data. In these two steps, it is generally called high and low-level synthesis. Speech sound is finally created with low-level synthesizer via high-level synthesizer information
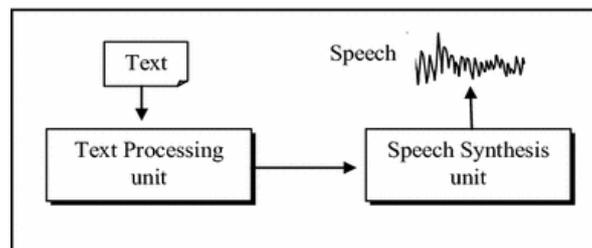

Fig.3. Text to speech synthesizer

V   Applications

1. For Blind People-

We develop a protocol that helps the visually impaired people who read paper-printed text effectively and effectively. optical character recognition is useful for visually impaired people who cannot read text document, but need to access the content of the text documents.

2. Data entry-

This area includes technologies for entering large amounts of restricted data. Initially this would be document reading machine has been used for banking applications. The system is described Just read a very limited set of printed characters, in degrees And some special signs. They are designed to read account numbers, customer-like data Identity, article number, amount of money etc. The documents are limited With a fixed line of limited number to read per document.

3. Automatic number-plate readers-

There are some systems for automated readability of car number plates. As opposed to others OCR application, input image is not a natural Believe Image and it must be caught Very fast

camera Despite character it creates these particular problems. The set is limited and the syntax is restricted.

4. Automatic cartography

Character Recognition in Maps presents special problems under Recency Character Recognition. Symbols are combined with graphics, text can be printed at different angles and Characters can have many fonts or even handwriting.

5. Signature verification and identification.

This is an application specially useful for the banking environment. Such a system establishes the identity of the writer without attempting to read the handwriting.

VI Conclusion

This paper elaborated survey of disparate techniques of OCR has been studied. Though it is not an atomic process it comprises various phases of recognition such as optical scanning, pre-processing, classification, feature extraction and recognition. We demonstrate how appropriate pre-processing and post processing techniques are required for OCR on digital camera acquired images. Despite the vast researching in the field of Optical Character Recognition, there are various challenges that still exist such as recognition of characters in various languages, real-time recognition etc. Finally, the use of OCR in real world applications remains an active area of research.

References

[1] Annapoorani, A.Kanchana, S. Aditi, "Book Reader Using Raspberry Pi for Visually Impaired", International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 05 ,2018

[2] Amal Jojie , Ashbin George, Dhanya Dhanalal Nayana , "Book Reader for Blind" IOSR Journal of Engineering (IOSRJEN) ISSN (e): 2250-3021, ISSN (p): 2278-8719 ,Emerging Research Trends in Electrical Engineering-2018 (ERTEE'18)

[3] Ajay Roy, Almaria Thomas K et.al, " Text Recognition And Face Detection Aid For Visually Impaired Person Using Raspberry Pi", International Conference on circuits Power and Computing Technologies [ICCPCT], 2017

[4] A. S. Sawant, "Script Independent Text Pre-processing and Segmentation for OCR," Int. Conf: Electr. Electron. Signals, Commun. Optim. - 2015, pp. 1-5, 2015.

[5] V. Kieu, F. Cloppet, and N. Vincent, "OCR Accuracy Prediction Method Based on Blur Estimation," 20/6 12th IAPR Work. Doc. Anal. Syst., pp. 317-322, 2016.

[6] J. B. Pedersen, K. Nasrollahi, and T. B. Moeslund, "Quality Inspection of Printed Texts," JWSSP 2016- 23rd Int. Conf: Syst. Image Process. 23-25 May 2016, Bratislava, Slovakia, pp. 6-9, 2016.

[7] A. F. Mollah, N. Majumder, S. Basu, and M. Nasipuri, "Design of an Optical Character Recognition System for Camera- based Handheld Devices," L/CSI, vol. 8, no. 4, pp. 283-289, 2011.

[8] B. Jain and M. Borah, "A Comparison Paper on Skew Detection of Scanned Document Images Based on Horizontal and Vertical," L/SRP, vol. 4, 2014

[9] E. N. Bhatia, "Optical Character Recognition Techniques: A Review," IJARCSSE, vol. 4, no. 5, pp. 1219-1223, 2014.

[10] Ani R, Effy Maria, J Jameema Joyce, Sakkaravarthy V, Dr. M. A. Raja   " Voice Assisted Text Reading system for Visually Impaired Persons Using TTS Method", IEEE International Conference on Innovations in Green Energy and Healthcare Technologies(ICIGEHT'17)

[11] A. Sravanthi, V. Aishwarya, S. Shrinidhi, P Jayashree, "Design and development of Personal Assistive Device for Elderly", IEEE Eighth International Conference on Advanced Computing (ICoAC),2016

[12] Kai Wang, Jianming Jin, Qingren Wang, " High Performance Chinese/English Mixed OCR with Character Level Language Identification", 10th International Conference on Document Analysis and Recognition,2009

[13] Shanu Sharma, Pooja Sharma, " Image Processing based Degraded Camera Captured Document Enhancement for Improved OCR Accuracy", IEEE 6th International Conference -Cloud System and Big Data Engineering (Confluence),2016

[14] A vinav Sharan , Arpit Agarwal, " Data Extraction from Exam Answer Sheets using OCR with Adaptive Calibration of Environmental Threshold Parameters", IEEE on computational research,2013

[15] Sajad Shirali-Shahreza, M. T. Manzuri-Shalmani, M. Hassan Shirali-Shahreza " Preparing Persian/Arabic Scanned Images for OCR", 0-7803-9521 IEEE,2006