A Review on Optical Character Recognition

Archit Singhal[1],and Bhoomika[2]

1 Department of Computer Science, The NorthCap University, Gurugram, Haryana 122017

singhal97.archit@gmail.com, wbhoomikaw@gmail.com

Abstract--- Nowadays, using a keyboard for entering data is the most common way but sometime it becomes more time consuming and need lots of energy. So, a technique was invented named Optical Character Recognition abbreviated as OCR that transfigures printed as well as handwritten text into machine encoded text by electronic means. OCR has been a topic for research for more than half a century. It electronically and mechanically converts the scanned images which can be handwritten, typewritten or printed text. In general, to figure out the characters of page, OCR compares each scanned letter pixel by pixel to a known database of fonts and decides onto the closest match.

Index Terms - optical character recognition, processed, pixel, scanned document, machine **encoded text.**

I. INTRODUCTION

Optical Character Recognition is a simple way of digitizing machine-encoded text that can be searched through and processed by a machine. It is amongst the greatest topic of research in the field of Artificial Intelligence, Pattern Recognition, Machine Vision and Signal Processing. Character Recognition techniques associate a symbolic identity within the image of character. It extracts the significant information and directly enters it to the database instead of using accustomed methods of manual data insertion.

This technique was firstly introduced for two main reasons i.e., expanding telegraphy and helping blinds to get education. Emanuel Goldberg and Edmund Fourier d'Albe were first to work on this technique in 1914. They built a machine that firstly scan the characters and later convert them into standard telegraph code and another device named Optophone that produced specific tone around specific letters or characters. These machines were patented in 1931 and now they are acquired by IBM.

OCR in general is classified into two types: Off-line and On-line. This technique of Off-line recognition is used for automated conversion of text into codes of letter which are usable by computer and applications developed for text processing. But, it is more difficult, as different people have different handwriting font. Whereas, On-line recognition deals with a continuous input of data stream that comes from a transducer when the user types or writes.

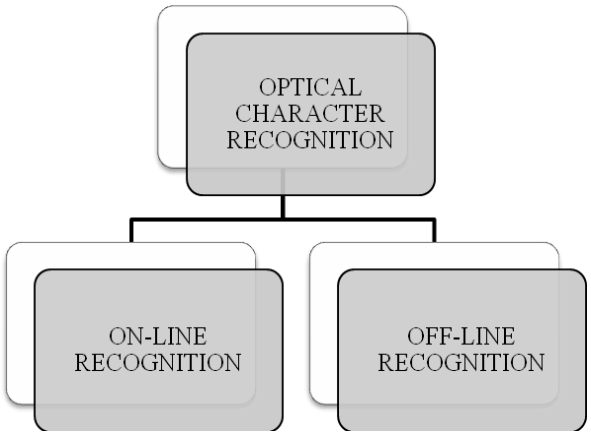

Fig. 1. Types of Optical Character Recognition

II. LITERATURE REVIEW

Research Paper Statement: A technique named Optical Character Recognition abbreviated as OCR which is in its development stage has proven to be much beneficial for transfiguring any kind of handwritten material to digitized form.

This paper reviews the work done by various authors in the field of exploring Optical Character Recognition. Prior studies have identified various steps involved from pre-processing the image to give the final Digitized output. Also, the paper has depicted various fields where this technology has been efficiently implemented. But as it is in its development stage, it also faces few challenges in giving the best required output. Integrating the concept and theories provided in paper to various

other fields with more advanced development will show much better results surpassing 99%.

Additionally, material learned in paper can be applied to benefit the community through a variety of tangible services

III. PHASES OF OCR

The whole procedure of transfiguring the handwritten as well as printed text into machine encoded text is broadly divided into four simple phases:

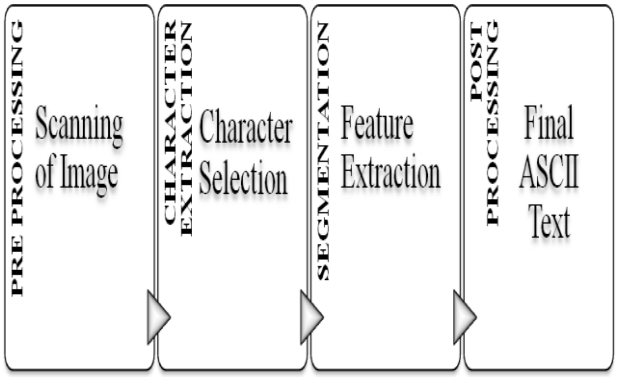

Fig. 2. Phases of Optical Character Recognition

A. Pre Processing

In this phase, the image is scanned starting from the top to the end & converted into gray level image, which is then converted into digital binary image. This process is sometimes termed as Digitization of image or Binarization. We use various scanners for this phase and last digital image then goes to the next step.

B. Character Extraction

The pre-processed image of the previous step serves as the input of the following step. In this step, each single character of the image is recognized. Also, the image is converted into the window size from the normalized form in this step.

C. Segmentation

This is the most important step of the whole process as it removes most of the noises from the images for more understandable form. It segments different characters into various zones i.e., upper, middle and lower zone.

Segmenting is difficult in offline recognition because of variability in paragraph, words of line and character's shape (slant, skew, curve, etc). Sometimes this difficulty arises due to overlapping of one or more characters also.

D. Post Processing

In this phase, features of every character is enhanced and extracted. In this phase, we can classify every character in a unique way. Feature of individual character is enhanced. Also, if there are some unrecognized characters found, they are also given some meaning. Extra templates can also be added in this phase for providing a wide range of compatibility checking in the systems database.

IV. APPLICATIONS OF OCR

Optical Character Recognition transfigures the scanned documents into more than an image file; rather, turn it into a readable as well as searchable text-file that can be processed by computers.

OCR is a field with enormous application in number of industries such as legal, healthcare, banking, education, etc.

A. Banking

In banks, cheques are processed using OCR without any kind of involvement by humans. The inserted cheque in the device is searched & scanned for the writing in various fields and the amount is transferred to the following payer. This whole process reduces the overall cheque process time.

B. Healthcare

The use of OCR technology has also been increased in Healthcare industry to process paperwork. In the healthcare industry, they deal with the huge amount of forms like patient details, medical-history, insurance forms, etc., so, in order to reduce energy and time, this technology is used.

C. Legal Industry

Documents are scanned; information is extracted and automatically entered into the database to save space. The time consuming task that requires the need to search for information through boxes is also eliminated. This helps in locating any of the specific text/document easily. It has also helped legal industry to have easy, fast and readily available access to a huge library of documents.

D. Invoice Imaging

It is important to maintain a track of financial records to avoid any piles of backlog payments. Among other processes, OCR helps in simplifying collection and analysis of large sets of data. It is also used to decrypt the large amount of information stored in the Digital code like Bar & QR codes.

E. Other Fields

OCR is extensively used in many other different areas also, like:

CAPTCHA- to prevent hacking;

Digital Libraries- sharing of digital teaching material;

Optical Music Recognition- to extract information from images;

Automatic Number Recognition- to identify vehicle registration plates;

Handwriting Recognition; Education;

Maestro Recognition Server; Trapeze.

V. CHALLENGES OF OCR

The techniques of OCR require images of high resolution which have basic structural property differentiating text and background to get high accuracy in character recognition. Image generation plays an important role in determining the accuracy and successful recognition. The image generated by scanners gives high performance and accuracy while images generated by cameras have numerous errors due to surroundings & factors related to camera.

These errors are clarified as follows

A. Tilting

The image of documents obtained by scanners is parallel and in line to plane of sensor which is not observed in image taken by hand-held devices. The text nearer to camera seems a little large while the text distant appears smaller which causes perspective distortion resulting in tilted pictures. The perspective intolerant recognizer causes lower recognition rate

B. Scene Complexity

The image taken by portable devices generally involves various no of artificial objects such as building, symbols, cars etc. considering a regular environment. The object detected makes the text recognition in processed image very challenging as the appearances and structure of these objects is comparative to the text present around it. Text itself is easily present in any form to encourage decipherability making the scene of segregating text from non-text very intricate

C. Conditions of uneven lighting

The major challenge for OCR is degradation of text quality due to uneven lighting and shadows when images are taken in a natural environment. This results in poor detection, recognition and detection of text. This case of shadows and uneven lightning differentiates between images taken by the camera and scanners. The lack of proper lighting makes scanned images more preferred than images processed by camera for their better text and characteristics quality. But these problems of lighting can be solved by using flash in camera which also lead to some new challenges.

D. Skewness

In OCR technique, the POV for the image used as input might change when the image is taken from camera or any hand-held devices which is not applicable for scanner image input. As a result, change of point of view leads to skewing which provides a great degree of poor results when image is processed. To overcome this problem, many deskew techniques are available such as RAST algorithm, Methods of Fourier transformation, projectile profile etc.

E. Aspect ratio

The image of documents obtained by scanners is parallel and in line to plane of sensor which is not observed in image taken by hand-held devices. The text nearer to camera seems a little large while the text distant appears smaller which causes perspective distortion resulting in tilted pictures. The perspective intolerant recognizer causes lower recognition rate and accuracy. The new latest cell phones can easily recognize if the portable device is tilted and then can prohibit clients to click images. This all detection and prohibition is done with the help of orientation sensors which also allows camera to align the text in plane of form resulting in greater degree of evenness

F. Wraping

One character on another can be another challenge for OCR to be precise. This situation arises when images are scanned using flatbed scanners which procured text on picture of the twisted text.

For panacea, a technique called dewraping was introduced by Ulges et al, which treat these texts the same way as they are equally distant and parallel to each other.

G. Multilingual Environments

Latin language contains a large number of symbols, character classes as it is composed of many other languages like, Japanese, Korean and Chinese. Arabic languages have characters with different writing shapes. Hindi language contains syllables which are made up of combining different shapes. Therefore, multilingual become a primary problem in OCR.

H. Fonts

Using different styles and fonts for different characters can make them overlap with each other and thus making OCR difficult. It is difficult to perform precisely accurate recognition due to various within subclasses variations and forming pattern sub-spaces.

VI. SCOPE OF OCR

Nowadays, a diverse collection of OCR systems are available but still we face many problems therein. The collections of OCR systems were earlier categorized into two groups. The first group includes machines that are specially designed to recognize specific set of problems which are mostly hard-wired so become little expensive and also decreases throughput rates. The second kind of group includes all software based techniques which involve computer or low cost scanner. Due to advancements in recent technologies, the second group of OCR systems is much more cost effective with high throughput; however, there are few limitations in these systems regarding speed and reading set of characters. They read the data line-by-line and transfer it to the OCR software systems. OCR systems are now categorized into five different groups based on character sets, namely, fixed-font, multi-font, omni-font, constraint handwriting and scripts.

Imperfections and irregularities in OCR systems are mainly due to problems occurred during scanning phase which usually result in inappropriate text or character. These irregularities often result in the misinterpretation among text and graphics or among text and noise. Perfectly scanned character can also cause imperfections due to characters with the same shapes and features, which makes the system difficult to exactly recognize the character. With this we can conclude that precision of OCR totally depend on the quality of input it takes.

Although we have seen a lot of improvement and advancement in OCR in recent years, from reading only a limited set of characters to reading characters with different fonts and styles and further reading handwritten text. In the coming years, seeing advancement in technology, one can predict that OCR can have much more potential and recognition in following years.

VII. CONCLUSION

This paper tells about a field in Artificial Intelligence i.e., Optical Character Recognition; its types, its whole process and its applications in different areas.

Optical Character Recognition or OCR has made scanned documents to become more than an image file, rather, turning them into a fully searchable, readable as well as editable text file that can be processed by computers.

The research in this area has been going from more than half of the century and the aftermath have been striking with successful recognition rates surpass 99% with notable advancement accomplishing for cursive handwritten character recognition. Further, the research in this area aims for more improvement and scope.

REFERENCES

[1] G.Vamvakas, B.Gatos, N. Stamatopoulos, and S.J.Perantonis: A Complete Optical Character Recognition Methodology for Historical Documents 2007.

[2] Karez Abdulwahhab Hamad, Mehmet Kaya: A Detailed Analysis of Optical Character Recognition Technology. International Journal of Applied Mathematics, Electronics and

Computers Advanced Technology and Science ISSN: 2147-8228.

[3] Combination of Document Image Binarization Techniques 2011.

[4] International Conference on Document Analysis and Recognition 2015.

[5] D-Lib Magazine: How Good Can It Get? Analysing and Improvising of OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs.

[6] Raghuraj Singh1 , C. S. Yadav2 , Prabhat Verma3 , Vibhash Yadav4: Optical Character Recognition (OCR) for Printed Devnagari Script Using Artificial Neural Network

[7] B. B. Chaudhary and U. Pal: OCR Error Detection and Correction of an Inflectional Indian Language Script, Pattern Recognition, IEEE Proceeding of 13th International Conference on Image Processing 2002.

[8] Agia Paraskevi, Athens: Institute of Informatics and Telecommunications, National Center for Scientific Research: Demokritos, GR-153 10.

[9] Optical Character Recognition System Using BP Algorithm: Department of Industrial Systems and Information Engineering, Korea University, Sungbuk-gu Anam-dong 5 Ga 1, Seoul 136-701, South Korea.

[10] Dholakia, K., A Survey on Handwritten Character Recognition Techniques for various Indian

[11] Languages, International Journal of Computer Applications, 115(1), pp 17–21, 2015.

[12] Yu, F. T. S., Jutamulia, S. (Editors): Optical Pattern Recognition, Cambridge University Press, 1998.

[13] Mantas, J.: An Overview of Character Recognition Methodologies, Pattern Recognition, 19(6), pp 425–430, 1986.

[14] Pradeep J, Srinivasan E, Himavathi S.: Diagonal based feature extraction for handwritten character recognition system using neural network. InElectronics Computer Technology (ICECT), 2011 3rd International Conference on 2011 Apr 8 (Vol. 4, pp. 364-368). IEEE.

[15] Bishnu A, Bhattacharya BB, Kundu MK, Murthy CA, Acharya T.: A pipeline architecture for computing the Euler number of a binary image. Journal of Systems Architecture. 2005 Aug 31;51(8):470-87.

[16] Verma R, Ali DJ. A-Survey of Feature Extraction and Classification Techniques in OCR Systems. International Journal of Computer Applications & Information Technology. 2012 Nov;1(3).

[17] Md. Anwar Hossain, Optical Character Recognition based on Template Matching(2018)

[18] C. Vasantha Lakshmi1 and C. Patvardhan "An optical character recognition system for printed

[19] Telugu text , Pattern Analysis & Applications", Category, Theoretical Advances, Volume 7, Number 2 / July, 2004 Pages 190-204