IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Volume 5, PP 70-74

A Review on Ocr Systems

Shantanu Wagh[1], Prasad Shetty[2], Kunal Sonawane[3], Vivek Sharma[4], Varsha Salunkhe[5],

[1](Student, Information Technology, Atharva College of Engineering, Mumbai University, India)
[2](Student, Information Technology, Atharva College of Engineering, Mumbai University, India)
[3](Student, Information Technology, Atharva College of Engineering, Mumbai University, India)
[4](Student, Information Technology, Atharva College of Engineering, Mumbai University, India)
[5] (Asst. Professor, Information Technology, Atharva College of Engineering, Mumbai University, India)

Abstract: *OCR (optical character recognition) started long back and has evolved step by step to a much better and efficient system today. Several different approaches were used to recognize the patterns of the characters. These approaches were later developed and a new OCR machine emerged with each development promising more accuracy and speed with capacity of extracting different types of characters and shapes at a time. There are many articles published on OCR from the past few decades and many commercially available OCR are available too. In this review we are going to highlight the history and various techniques used throughout the entire course of OCR evolution and development.*

Keywords: *Binarization, Dots Per Inches, Grayscale, OCR Training, Optical Character Recognition.*

I. Introduction

During the early ages OCR was considered as an aid for the blind people. But later on, it developed into a vast field of research and development. Reading the characters printed on a physical paper is being started many decades ago even before the advent of computers. In 1929 Tausheck obtained a patent on OCR in Germany and in 1933 Handel did the same in the U.S. These are the first concepts of the idea of OCR as far as we know. At that time certain people dreamed of a machine which could read characters and numerals. This remained a dream until the age of computers arrived, in the 1950's. The principle is template/mask matching. This reflects the technology at that time, which used optical and mechanical template matching. Light passed through mechanical masks is captured by a photo detector and is scanned mechanically. When an exact match occurs, light fails to reach the detector and so the machine recognizes the characters printed on paper. Mathematically speaking, the principle is the axiom of superposition, which was first described by Euclid as the seventh axiom in the first volume of Elements.Our paper aims to introduce a very new and highly developed application of typical OCR ideology.OCR technology can also be used to convert any textual patterns in an image into actual text. There are basic drawbacks of OCR engines and overcoming them is our major objective.OCR engines are not reliable interms of text extraction. Normally an OCR engine has an accuracy rate of up to 85% for an unfiltered image [1]. This basic drawback of OCR‘s makes them usable for very small quantity of text extraction and often used where few errors can be neglected. We can surpass this limitation by filtering the images through various noise removal processes.

1.1 History of OCR

Tauschek was the first to file a patent for OCR in Germany 1929 and later he received a US patent in the year 1935. Handel also received patent for OCR from US in the year 1933. both the machines used a circular disc with template symbols cut out of it so that light shines through it. The image to be recognized is held in front of the disc and is then illuminated. The light reflecting off a portion of it is then focused through the template hole and detected by a photo sensor. Based on efficiency, robustness and versatility the OCR system can be classified into 5 generations.

1.1.1 First generation of OCR

The first generation was employed in 1960 in which the OCR machines could only read selected fonts and shapes of the characters. NCR 420 and Farrington 3010 were typically used during those times. IBM 1418 was used logical template matching and was widely commercialized. It could read only some specific fonts of IBM. The method which IBM used in implementing it was logical template matching. In it, the positional relationship was used extensively. Apart from this, IBM also launched its other OCR based systems which were as follows IBM 1428, IBM 1285.IBM was not the only one who was developing First Generation OCR systems. Many other Japanese companies were also developing it. Prominent OCR systems among them were Facom 6300A which was developed by Fujitsu company and H-852 made by Hitachi whereas N240D-1 was built by

NEC.[2] The first two systems mentioned above used stroke width transform method whereas the last one used logical template matching.

1.1.2 Second generation of OCR

The second generation of OCRs was available during the middle of 1960 to early 1970. they were more capable and could recognize both characters printed by machine as well as written by hand. Initially during this generation researchers were focusing on recognition of numerical characters only. But later they developed systems which could even recognize other types of characters. IBM 1287 was the first OCR system of second generation. It was a hybrid system that combined both digital and analog technology. It was also showcased during the 1965 New York Fair [2]. But this was not the only technology developed during this era. Another technology developed during this era was automatic Letter Sorting Machine which was developed by Toshiba. It was basically developed using Structural analysis method. Hitachi also made some efforts during this time and was able to develop H-8959 OCR system.

1.1.3 Third generation of OCR

These systems were poplar during the period of 1975-1985. They can operate on large set of characters than the earlier versions and could also read poor quality characters. During this era more importance was given to performance and cost of OCR systems. As the world was demanding more performance from these systems researchers were also keen to increase quality of OCR systems.OCR systems were expected to even recognize images with bad quality and poor printing. Hand written characters were varying a lot. So, OCR had to recognize every variation so it had to be trained properly.

1.1.4 Fourth generation of OCR

The fourth generation OCRs are capable of scanning and recognizing characters from complex documents intermixed with texts, tables and mathematical symbols and also low quality noisy documents such as fax and photo copies, unconstrained handwritten characters and color documents. During this time, commercial OCR software packages to be run on computers were introduced in market. Some of the major OCR applications developed during this era were: Postal address readers and Self Reading software for the Blind [2]. Postal address readers were able to sort handwritten addresses. Self-reading software was brought into market by company called as XeroxKurzwel.It used OCR to read characters and spell it out to listeners in English.

1.1.5 Fifth generation of OCR

The Fifth generation of OCR is present Generation of OCR.Today we have developed OCR systems which can even recognize old and ancient scriptures of different languages. Today OCR systems have been trained to understand Roman, Egyptian, Greek, Chinese, Sanskrit, Japanese and many other languages. Old scriptures don’t remain in original condition. In technical terms we can say that their images may need a lot of filterations. But today’s OCR systems are capable to do that. Apart from this today OCR is extensively used even for security purposes, and in traffic regulation systems.

Table: OCR Generations

GENERATION	ERA OR TIME PERIOD	TYPE OF CHARACTERS READ BY OCR
FIRST	EARLY 1960s	LIMITED LETTERS AND SHAPES
SECOND	MID 1960 TO EARLY 1970	HAND-WRITTEN AND MACHINE PRINTED CHARACTERS
THIRD	1975-1985	POOR QUALITY DOCUMENTS
FOURTH	1985-1995	COMPLEX DOCUMENTS WITH GRAPHICS, SYMBOLS, ETC
FIFTH	1995-PRESENT	ANCIENT SCRIPTURES

2.Approaches used for OCR

Over the years various approaches are used for designing OCR systems. They are as follows: -

2.1 Matrix Matching

It is used for converting every character in the image into some pattern inside a matrix. The next step involves comparing the above-mentioned pattern with some index of known characters. This type of recognition is best when we do it for monotype and uniform column pages.

2.2 Fuzzy Logic

Fuzzy logic is a multi-valued logic in which the variables have some truth values and that truth value is a real number which could be between 0 and 1. It is usually done when answers dont have values like yes/no, true/false, black/white etc. It is a part of Artificial Intelligence in which computers are made to think like humans. It involves Logical Reasoning. The term was coined as Fuzzy Logic in 1965 by Lotfi Zadeh. [3] Fuzzy Logic was earlier considered as infinite valued logic.

2.3 Template Matching Method

This approach was taken by Kelner and Glauberman in the year 1956. The reduction in complexity was achieved by projecting two-dimensional information onto one using a magnetic shift register. Template matching is done by taking the total sum of the differences between each sampled value and the corresponding template value, each of which is normalized [6].

An appropriately placed input character is scanned vertically from top to bottom by a slit through which the reflected light on the printed input paper is transmitted to a photo detector. It has one weakness in registration which can be covered by autocorrelation and other is based on moment method.

2.4 Structure Analysis Method

Handwritten characters have many variations of curves and shapes and hence it is very difficult to create templates for the same. Hence a so-called structure analysis method is applied to the handwritten characters. In case of structural analysis there is no mathematical principle.

A structure can be broken into parts, the parts itself are able to define the structure by their features and relationships between the parts. So, the main problem arises is how to choose the features and relationships between the parts so that each character can be clearly identified. Feature extraction has therefore become the key in pattern recognition research.

2.5 Neural Networks

Neural networks are a widely used method which has been extensively used not only for pattern recognition in OCR Systems but also for time series prediction, function approximation, clustering, etc. It works like the human neural system. It creates a sample of pixels of text character extracted from image and then matches those pixels with known index of pixel patterns of a particular character [4]. Its ability to abstract the characters helps a lot in case of damaged text or unfiltered text. It can represent complicated input and output relationships. When we say that it works like human nervous system, it means it resembles human brain. It does this by acquiring useful knowledge by learning from its environment. This knowledge is stored within its synaptic weights [5]. Synaptic weights are the strength of the connection between the nodes.

II. Proposed System

The overall accuracy of the engine depends on various factors namely:

- Quality of image
- Size of text in image
- Binarized image
- Properly rotated and skewed Image
- Trained data set

Our project does not directly pass the Image to the OCR engine, instead it performs several image filtering procedures on the image to get a proper finalized image.

This includes:

- Checking if image is rotated or skewed and correct it
- Scaling the image if it is smaller than 300dpi (Dots per Inch)
- Removing the unusable borders from the image to improvise speed and efficiency
- Binarizing the image i.e. converting it to grey scale

These steps are proven to increase OCR accuracy by 50%. This filtered and processed image is then passed to the OCR engine which will be highly trained so that it will recognize the characters even better than normal OCR engine.

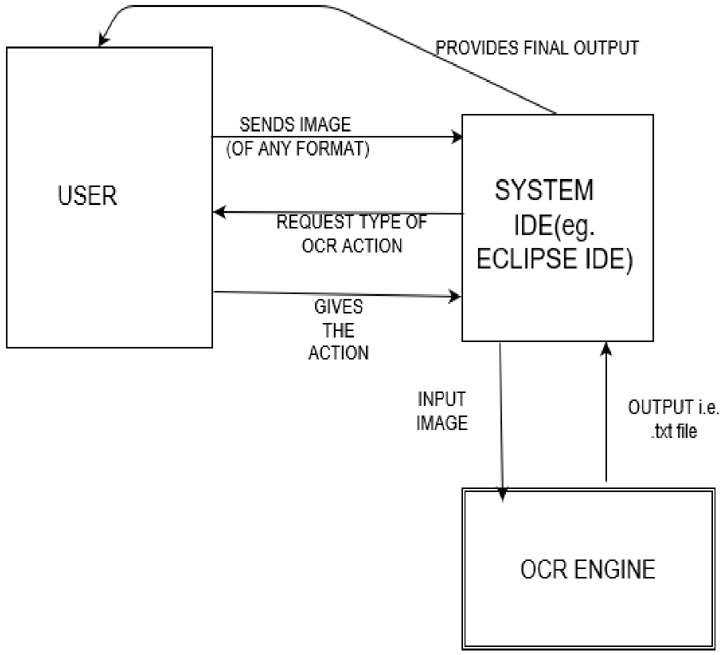

Fig.1: block diagram of proposed system

III. Methodology

4.1 Receiving Image Input

The user is responsible for sending the image of any format that is .jpg, .png, .tiff, .pdf, .bmp etc. It captures the image with the help of a camera, it then receives a request from the system module that whether the image given is an image is of a passage or program. It selects the appropriate option and accordingly receives the output from the system IDE. The system IDE is the environment in which we are going to perform Image preprocessing. For e.g.: Eclipse IDE. The user tries to provide the best possible image input but ultimately it depends on the condition of the material whose image is to be taken.

4.2 Image Pre-processing

As we already know that the image given to us is not in its best form which means it needs some amount of filterations due to different types of noise present in it, we have to perform Image preprocessing. Image preprocessing basically involves skewing, deskewing, binarization, rotation, Changing DPI to 300 etc.

For Binarization we plan to use Otsu Thresholding Algorithm. This algorithm uses a threshold value for converting image into its equivalent gray scale image. The Threshold Value it uses is approximately between 130 to 150. For Skewing and deskewing we intent to use Canny’s Edge detection algorithm which would help us in identifying edges of image and find out the skewing angle.

4.3 Character Recognition

Our OCR engine performs this task. It is the heart of the entire system it carries out the Optical Character Recognition process and extracts each and every alphabet, number or symbol from the image. The extraction of text depends upon the accuracy of the OCR engine. The OCR engine provides accuracy as best as possible so that there are least possible errors in the extracted text. To make OCR Engine more perfect we train it for different fonts through which OCR engine would be able to recognize the characters of the text properly. In our paper we plan to use Neural Network Algorithm to perform character recognition. In neural network algorithm the position of the pixels of character we are going to recognize becomes our input to algorithm. This algorithm partitions its data set into trained set and validation set. The trained set is used for matching it with the input set whereas the validation dataset is to make predictions about the character and compute its accuracy.

IV. Conclusion

Though OCR is not a new technology its precision is something that the world demands. The OCR technique is being developed throughout the years and thus is able to achieve higher accuracy which can still be elevated through series of pre-processing and post processing procedures to improve accuracy along with training the OCR system.

References

[1] Miss. Pooja Chavre, Dr Archana Gotkar, *Scene Text Extraction using Stroke Width Transform for Tourist Translator on Android Platform, International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT) International Institute of Information Technology (I²IT), Pune, 9-10 Sept. 2016*

[2] Hiral Modi, M. C. Parikh, *A Review on Optical Character Recognition Techniques, International Journal of Computer Applications (0975 – 8887) Volume 160 – No 6, February 2017*

[3] Sukhpreet Singh, *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 2, Issue 6, June 2013.*

[4] N. Venkata Rao, Dr. A.S.C.S. Sastry, A.S.N. Chakravarthy, *Journal of Theoretical and Applied Information Technology 20th January 2016, Vol.83. No.2*

[5] Ms. Sonali. B. Maind, Ms. Priyanka Wankar, Research Paper on Basic of Artificial Neural Network, *International Journal on Recent and Innovation Trends in Computing and Communication, ISSN: 2321-8169 Volume: 2 Issue: 1.*

[6] Swapnil Desai, Ashima Singh, *Optical character recognition using template matching and back propagation algorithm , Inventive Computation Technologies (ICICT), International Conference on 26-27 Aug. 2016*