7th International Conference on Recent Innovations in Science, Engineering and Management
The Institutions of Engineers, Delhi State Center (India)
16th September 2016, www.conferenceworld.in
RISEM - 16
ISBN : 978-93-86171-07-8

OCR-OPTICAL CHARACTER RECOGNITION

Abhishek Verma[1], Suket Arora[2], Preeti Verma[3]

[1] *B.Tech (CSE), ACET, Amritsar,*

[2] *Research Scholar, IKJPTU, Kapurthala,*

[3] *Research Scholar, IKJPTU, Kapurthala,*

ABSTRACT

Optical Character Recognition or OCR is the electronic translation of handwritten, typewritten or printed text into machine translated images. It is widely used to recognize and search text from electronic documents or to publish the text on a website. OCR is the machine replication of human reading and has been the subject of intensive research for more than three decades. OCR can be described as mechanical or electronic conversion of scanned images where images can be handwritten, typewritten or printed text. It is a method of digitizing printed texts so that they can be electronically searched and used in machine processes. It converts the images into machine-encoded text that can be used in machine translation, text-to-speech and text mining. This paper presents a simple, efficient, and less costly approach to construct OCR for reading any document that has fix font size and style or handwritten style. 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. So this research paper is based on the construction, working and applications of OCR. Paper will also discuss different stages of OCR like optical scanning , location segmentation ,preprocessing ,feature extraction and recognition post processing.

Key Words: Feature Extraction, Location Segmentation, Machine-Encoded Text, Optical Character Reader, Optical Scanning, Preprocessing, Recognition Post Processing.

I. INTRODUCTION

Until a few decades ago, research in the field of Optical Character Recognition (OCR) was limited to document images acquired with flatbed desktop scanners. The usability of such systems is limited as they are not portable because of large size of the scanners and the need of a computing system. Moreover, the shot speed of a scanner is slower than that of a digital camera. But now, with the advancement of processing speed and high internal memory of hand -held mobile devices such as high- end cell -phones, Personal Digital assistants (PDA), smart phones, iPhones, iPods, etc. has given a new direction to OCRs.An app called 'Google Translate' has been launched for android phones which is one of the best app based on OCR.

Researchers have researched a lot about desktop OCRs. However, computing under handheld devices involves a number of challenges. Because of the non-contact nature of digital cameras attached to some handheld devices, acquired images very often suffer from skew and perspective distortion. In addition to that, manual involvement in the capturing process, uneven and insufficient illumination, and unavailability of sophisticated focusing

ISBN : 978-93-86171-07-8

system in some former devices yield poor quality images. The processing speed and memory size of former handheld devices was not yet sufficient enough so as to run desktop based OCR algorithms that are computationally expensive and require high amount of memory. The processing speeds of mobile devices with built-in camera start with as low as few MHz to as high as 624 MHz but now , the new generation handheld devices have been launched with high performance cameras with auto-focusing system.

II. CONSTRUCTION AND THE DIFFERTENT COMPONENTS OF OCR

OCR is a simple machine with different parts that works according to the programming. At first, a typical OCR consists of an optical scanner that scans the analog document .Then each symbol and character is extracted through a location segmentation process. The next stage is of preprocessing of scanned image. After that in the next stage i.e feature extraction, the characters are further extracted according to the required dimensions. At the end in the last stage of recognition post processing, the characters are reconstructed with reduced noise according to the original text. Fig.1 represents the stages of an OCR.

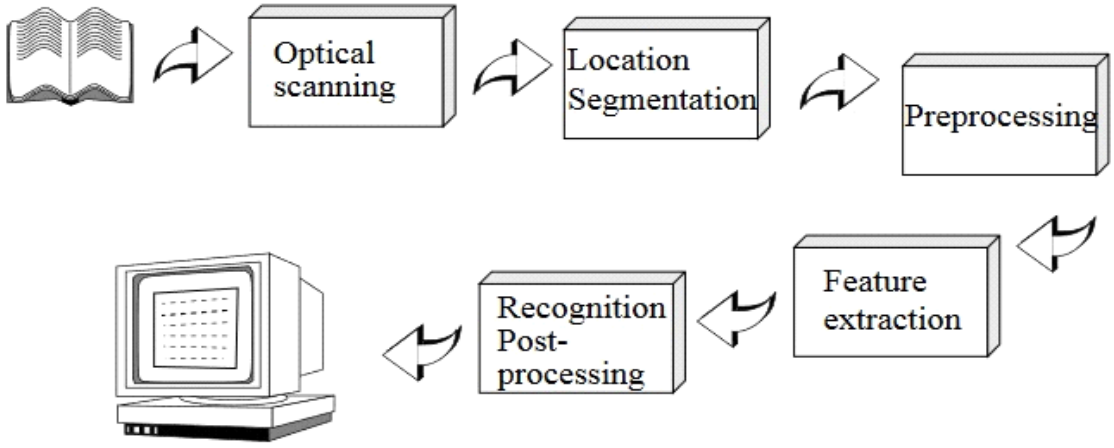

Fig.1: Stages Of An OCR

III. WORKING OF OCR

The working of OCR is done step by step according to different stages as mentioned above in the construction of OCR.

3.1 Optical Scanning

Through the scanning process a digital image of the original document is captured with the help of scanner or camera. In OCR optical scanners are used, which generally consist of a transport mechanism and a sensing device that converts light intensity into gray-levels. Printed documents usually consist of black print on a white background. Hence, when performing **OCR**, it is common practice to convert the multilevel image into a bilevel image of black and white. Often this process is known as thresholding. It is performed on the scanner to save memory space and computational effort. The thresholding process is important as the results of the following recognition are totally dependent of the quality of the bilevel image. Still, the thresholding performed on the

scanner is usually very simple. A fixed threshold is used, where gray-levels below this threshold is said to be black and levels above are said to be white. For a high-contrast document with uniform background, a fixed threshold can be sufficient. However, a lot of documents encountered in practice have a rather large range in contrast. In these cases more sophisticated methods for thresholding are required to obtain a good result. The best methods for thresholding are usually those which are able to vary the threshold over the document adapting to the local properties as contrast and brightness. However, such methods usually depend upon a multilevel scanning of the document which require more memory and computational capacity. Therefore such techniques are seldom used in connection with **OCR** systems, although they result in better images.

3.2 Location Segmentation

Segmentation is a process that determines the constituents of an image. It is necessary to locate the regions of the document where data have been printed and distinguish them from figures and graphics. For instance, when performing automatic mail-sorting, the address must be located and separated from other print on the envelope like stamps and company logos, prior to recognition. Applied to text, segmentation is the isolation of characters or words. The majority of optical character recognition algorithms segment the words into isolated characters which are recognized individually. Usually this segmentation is performed by isolating each connected component, that is each connected black area. This technique is easy to implement, but problems occur if characters touch or if characters are fragmented and consist of several parts. The main problems in segmentation may be divided into four groups:

3.2.1 Extraction of touching and fragmented characters

Such distortions may lead to several joint characters being interpreted as one single character, or that a piece of a character is believed to be an entire symbol. Joints will occur if the document is a dark photocopy or if it is scanned at a low threshold. Also joints are common if the fonts are serifed. The characters may be split if the document stems from a light photocopy or is scanned at a high threshold.

3.2.2 Distinguishing noise from text

Dots and accents may be mistaken for noise, and vice versa.

3.2.3 Mistaking graphics or geometry for text

This leads to non text being sent to recognition.

3.2.4 Mistaking text for graphics or geometry

This often happens if characters are connected to graphics.

3.3 Preprocessing

The image resulting from the scanning process may contain a certain amount of noise. Depending on the resolution of the scanner and the success of the applied technique for thresholding, the characters may be smeared or broken. Some of these defects, which may later cause poor recognition rates, can be eliminated by using a preprocessor to smooth the digitized characters. The smoothing implies both filling and thinning. Filling eliminates small breaks, gaps and holes in the digitized characters, while thinning reduces the width of the line. w. In addition to smoothing, preprocessing usually includes normalization. The normalization is applied to obtain characters of uniform size, slant and rotation. To be able to correct for rotation, the angle of rotation must

ISBN : 978-93-86171-07-8

be found. For rotated pages and lines of text, variants of Hough transform are commonly used for detecting skew. However, to find the rotation angle of a single symbol is not possible until after the symbol has been recognized. Therefore , preprocessing leads to the symmetry and alignment of characters in a scanned image.

3.4 Feature Extraction

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 straight forward way of describing a character is by the actual raster image. Another approach is to extract certain features that still characterize the symbols, but leaves out the unimportant attributes. The techniques for extraction of such features are often divided into three main groups, where the features are found from:

1. The distribution of points.
2. Transformations and series expansions.
3. Structural analysis.

The following Fig.2 represents the technique of feature extraction.

Feature extraction technique	Robustness 1 2 3 4 5	Practical use 1 2 3
Template matching	● ● ○ ○ ○	○ ● ○
Transformations	○ ● ● ● ●	○ ○ ●
Distribution of points: Zoning	○ ● ○ ○ ●	● ● ○
Moments	● ● ○ ● ●	○ ● ○
n-tuple	● ○ ● ○ ●	● ● ●
Characteristic loci	○ ● ● ● ●	● ● ○
Crossings	○ ● ● ● ●	● ● ○
Structural features	○ ● ● ● ●	● ○ ●

● High or easy ● Medium ○ Low or difficult

Fig.2: Feature Extraction Technique

The different groups of features may be evaluated according to their sensitivity to noise and deformation and the ease of implementation and use. The criteria used in this evaluation of these groups are as following:

3.4.1 Robustness

- **Noise:** Sensitivity to disconnected line segments, bumps, gaps, filled loops etc.
- **Distortions:** Sensitivity to local variations like rounded corners, dilations and shrinkage.
- **Style variation**: Sensitivity to variation in style like the use of different shapes to represent character or the use of serifs slants etc.
- **Translation:** Sensitivity to movement of the whole character or its components.
- **Rotation:** Sensitivity to change in orientation of the characters.

ISBN : 978-93-86171-07-8

3.4.2 Practical use

- Speed of recognition.
- Complexity of implementation.
- Independence.

3.5 Recognition Post Processing

These techniques are different from the others , in that no features are actually extracted. Instead the matrix containing the image of the input character is directly matched with a set of prototype characters representing each possible class. The distance between the pattern and each prototype character is computed, and the class of the prototype giving the best match is assigned to the pattern. The technique is simple and easy to implement in hardware and has been used in many commercial OCR machines. However, this technique is sensitive to noise and style variations and has no way of handling rotated characters.

At the end of the above processes the image is converted to the text with all corrections ,therefore a clear and original image of text is produced. The present system of Optical Character Recognition is presented in the Fig.3. .

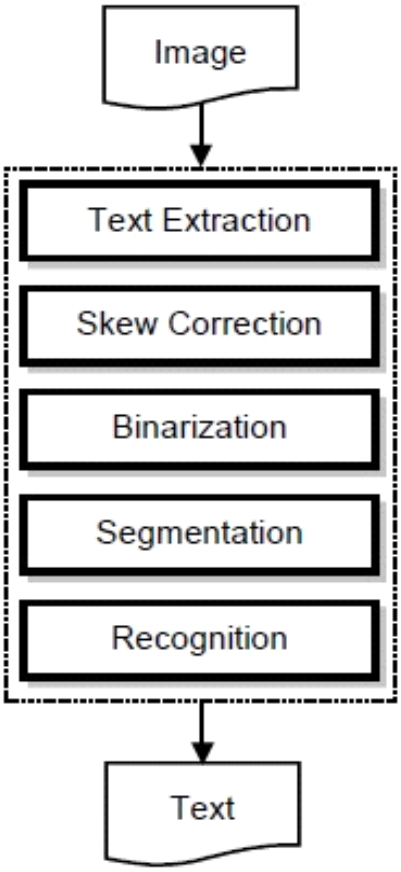

Fig 3: Block Diagram Of Present System

IV. APPLICATIONS OF OCR

ISBN : 978-93-86171-07-8

OCR itself is an application of computers and it further has many applications. OCR engines have been developed into many kinds of object-oriented OCR applications, such as receipt OCR**,** invoice OCR, check OCR, legal billing document OCR. Now a days, hand held OCRs are used for converting a language say English to other language say Punjabi with the help of scanning the image captured by camera. This is possible with the help of some OCR programmed apps like Google Translate. OCRs can be used for the following main purposes:

4.1 Banking

Another important application of OCR is in banking , where it is used to process cheques without human involvement. A cheque can be inserted into a machine where the system scans the amount to be issued and the correct amount of money is transferred. This technology has nearly been perfected for printed checks, and is fairly accurate for handwritten checks as well reducing the waiting time in banks. Fig.4 cheques by OCR

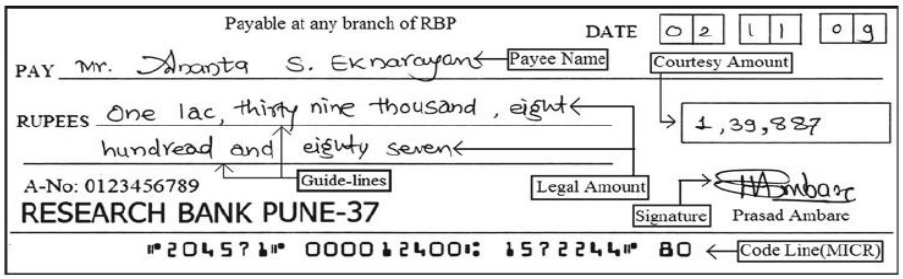

Fig.4: Processing of cheques by an OCR

4.2 Invoice Imaging

ISBN : 978-93-86171-07-8

Invoice imaging is widely used in many businesses applications to keep track of financial records and prevent a backlog of payments from piling up. In government agencies and independent organizations, OCR simplifies data collectioSSSSn and analysis, among other processes. As the technology continues to develop, more and more applications are found for OCR technology, including increased use of handwriting recognition. Furthermore, other technologies related to OCR, such as barcode recognition, are used daily in retail and other industries.

4.3 Legal Industry

Legal industry is also one of the beneficiaries of the OCR technology. OCR is used to digitize documents, and directly entered to computer database. Legal professionals can further search documents required from huge databases by simply typing a few keywords.

4.4 Captcha

A CAPTCHA is a program that can generate and grade tests that human can pass but current computers programmers' cannot. Hacking is a serious threat to internet usage. Now a day's most of the human activities like economic transactions, admission for education, registrations, travel bookings etc are carried out through internet and all this requires a password which is misused by hackers. They create programs to like dictionary attacks and automatic false enrolments which lead to waste of memory and resources of website. Dictionary attack is attack against password authenticated systems where a hacker writes a program to repeatedly try different passwords like from a dictionary of most common passwords. In CAPTCHA, an image consisting of series of letters of number is generated which is obscured by image distortion techniques, size and font variation, distracting backgrounds, random segments, highlights and noise in the image. This system can be used to remove this noise and segment the image to make the image tractable for the OCR (Optical Character Recognition) systems. Fig.5 represents the segmentation result of captcha.

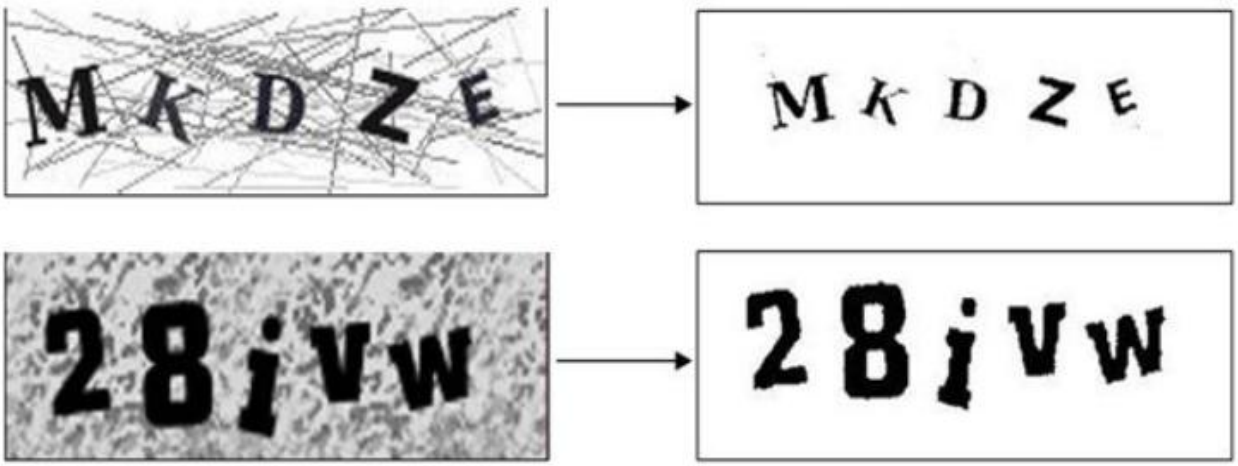

Fig.5: Segmentation result for captcha

4.5 Institutional Repositories and Digital Libraries

Institutional repositories are digital collections of the outputs created within a university or research institution. It is an online locale of intellectual data of an institution, especially a research institution where it is collected, preserved and aired. It helps to open up the outputs of an institution and give it visibility and more impact on

worldwide level. Enables and encourages interdisciplinary approaches to research and facilitates the development and sharing of digital teaching materials and aids. It is basically a collection of peer reviewed journal articles, conference proceedings, research data, monographs, books, theses and dissertations and presentations. Their first role is to provide the Open Access literature. Practical implementation of this includes setting up a system which consists of scanner which scans the documents. This scanned document is then fed as an input to an Optical Character Recognition system where information is acquired and retained in digitized form. Fig.6 represents the segmentation result for an institutional repository.

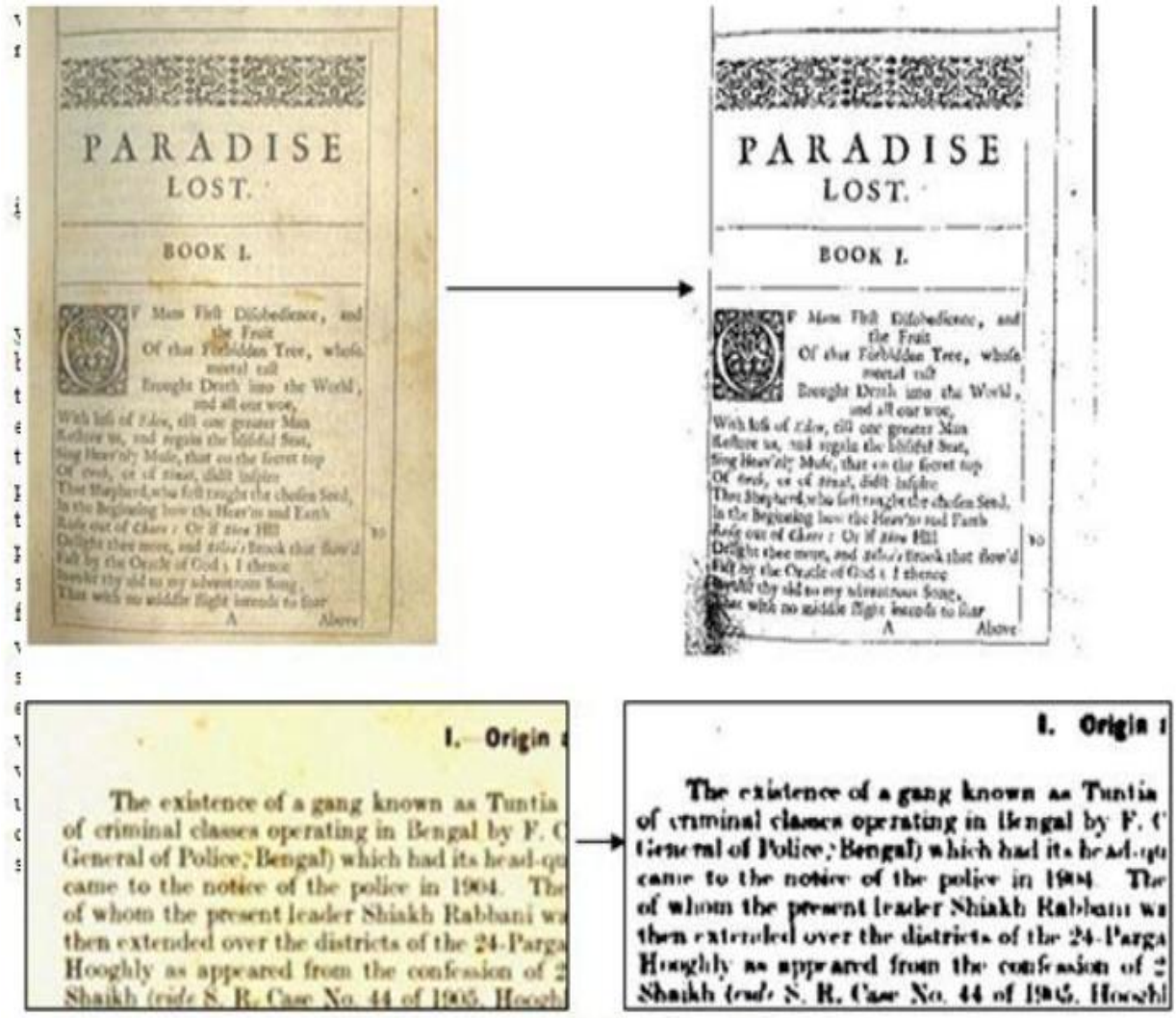

Fig.6: Segmentation result for institutional repository

.4.6 Healthcare

Healthcare has also seen an increase in the use of OCR technology to process paperwork. Healthcare professionals always have to deal with large volumes of forms for each patient, including insurance forms as well as general health forms. To keep up with all of this information, it is useful to input relevant data into an electronic database that can be accessed as necessary. Form processing tools, powered by OCR**,** are able to extract information from forms and put it into databases, so that every patient's data is promptly recorded.

4.7 Handwriting Recognition

ISBN : 978-93-86171-07-8

Handwriting recognition is the ability of a computer to receive and interpret intelligible handwritten input from sources such as paper documents, photographs, touch-screens and other devices. The image of the written text may be sensed "off line" from a piece of paper by optical scanning (optical character recognition) or intelligent word recognition. Alternatively, the movements of the pen tip may be sensed "online", for example by a pen-based

computer screen surface. Various OCR based android and iso apps has been also designed for recognition of hand written documents and papers. The books in various languages are also published with the help of handwriting recognition by an OCR. Fig.7 represents the handwriting recognition by an OCR.

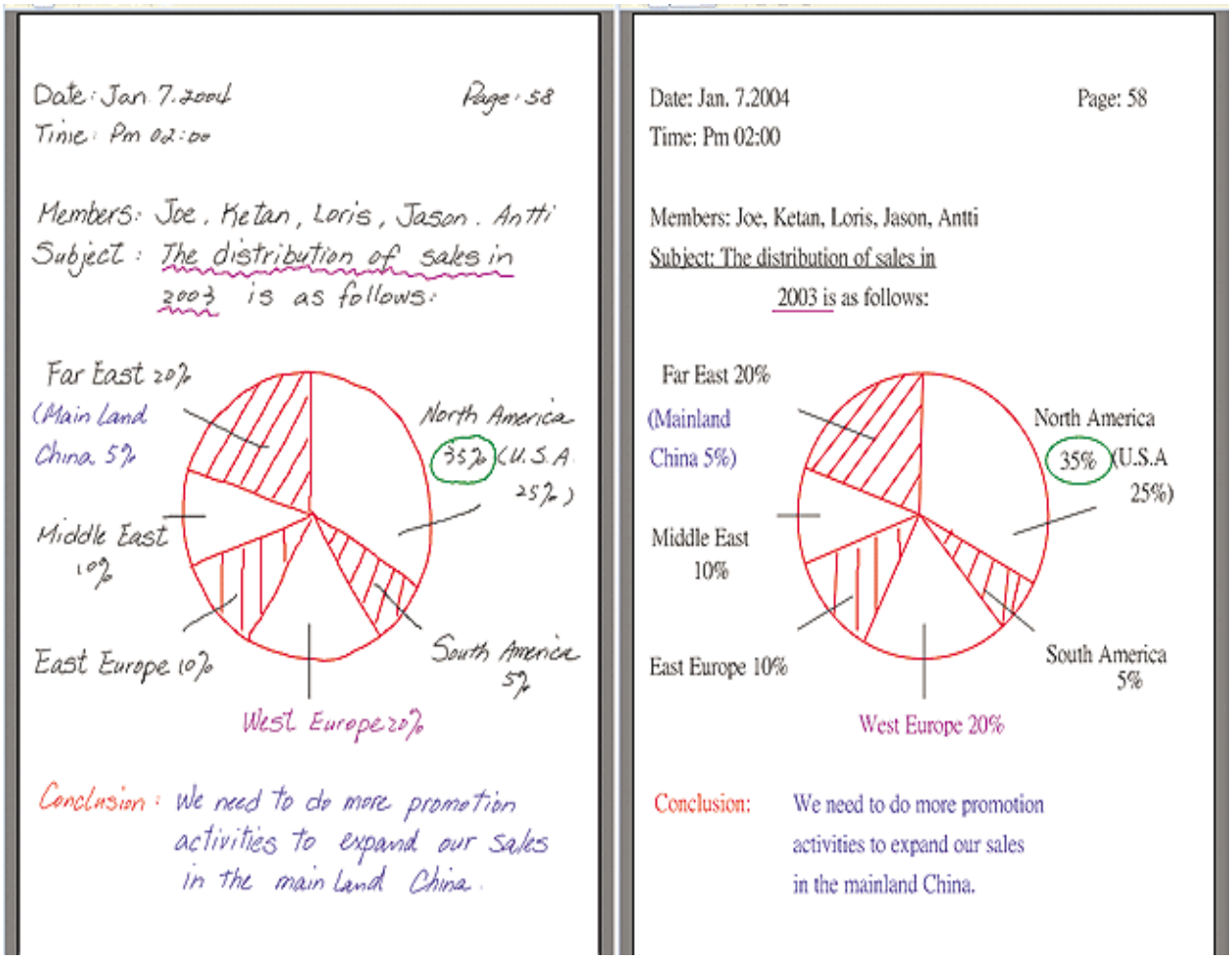

Fig.7: Handwriting recognition by an OCR

4.8 Optical Music Recognition

Automated learning system extract information from images and is part of major researches. Optical music recognition (OMR) born in 1950's is a developed field and initially was aimed towards recognizing printed sheets which can be edited into playable form with the help of electronic and electrochemical methods. An OMR system has many applications like processing of different classes of music, large scale digitization of musical data and also it can be used for diversity in musical notation. Image enhancement and segmentation is the basic step and hence the paper focuses on it. Fig.8 represents the segmentation result of OMR.

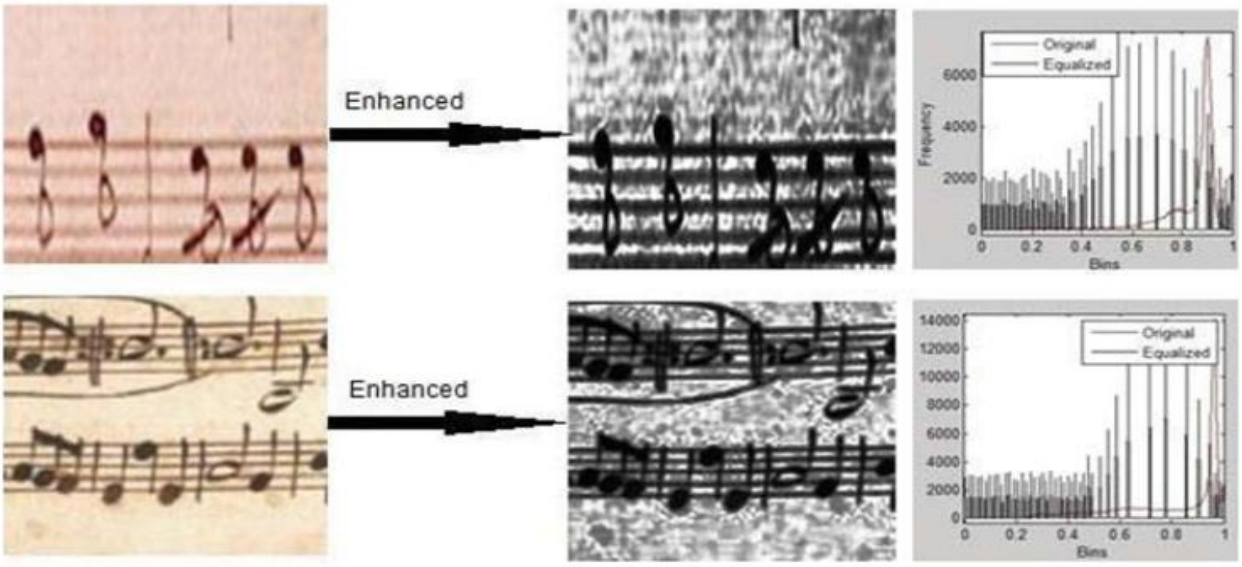

Fig.8: Segmentation result for OMR

4.9 Automatic Number Recognition

Automatic number plate recognition is used as a mass surveillance technique making use of optical character recognition on images to identify vehicle registration plates. ANPR has also been made to store the images captured by the cameras including the numbers captured from license plate. ANPR technology own to plate variation from place to place as it is a region specific technology. They are used by various police forces and as a method of electronic toll collection on pay-per-use roads and cataloging the movements of traffic or individuals. Automatic number plate recognition by an OCR also helps the police to control rash driving in accidents prone roads and areas. Therefore, this application of OCR is very useful worldwide. Fig,9 represents the segmentation result of automatic number plate recognition.

Fig.9: Segmentation result of automatic number plate recognition

V. CONCLUSION

This is detailed discussion about optical character recognition techniques and includes its use in different area of character recognition by using OCR. From study of various papers I have seen that selection of relevant technique plays an important role in performance of character recognition rate. This material serves as a helpful

guide and update for readers working in the Optical Character Recognition area. In the nutshell , researches in the field of OCRs is never ending because still OCR is a developing branch of computers , and a vast number of things about OCRs are still there to find out.

REFERENCES

[1] https://en.wikipedia.org/wiki/Optical_character_recognition

[2] Amarjot Singh, Ketan Bacchuwar, and Akshay Bhasin "A survey of OCR Applications" innternational Journal of Machine Learning and Computing Vol.2,No.3,June 2012.

[3] Sarika Pansare,Dhanshree Joshi"A Survey on optical character recognition techniques"in International Journal of Science and Research **(IJSR)**

[4] Shalin A.Chopra,Amit A. Ghadge, Onkar A. Padwal Karan S. Punjabi, Prof. Gandhali S.Gurjar," Optical Character Recognition "International Journal ofAdvanced Research in Computer and Communication Engineering Vol.3, Issue 1, January 2014.

[5] Ayatullah Faruk Mollah , Nabamita Majumder, Subhadip Basu and Mita Nasipuri "Design of an Journal of Computer Science Issues, Vol. 8, Issue 4, No 1, July 2011