Optical Character Recognition: A Review

[1]Ankit Kumar Singh, [2]Aman Gupta, [3]Aman Saxena

Department of Electronics & Instrumentation Engineering
Galgotias College of Engineering & Technology

Abstract— **The Optical Character Recognition is the electronic conversion of image of typewritten or printed text into machine-encoded text. It is common method of digitizing printed texts. Advantages being easy storage, edit ability, searching, etc. OCR is a field of research in pattern recognition, artificial intelligence and computer vision. In previous decades it has gain more importance due to feasibility with which it can convert image into editable document. The objective is to develop user friendly application which performs conversion of image into editable and searchable data. The OCR takes image as the input, gets text from that image and then converts it into editable document. This system can be useful in various applications like banking, legal industry, and other industries. It is mainly designed to save time and labor cost.**

Index Terms: **Image Processing, MATLAB, Optical Character Recognition (OCR), Segmentation, Template Matching and Correlation.**

1. INTRODUCTION

Character recognition is the set of process which together can classify the input character according to the predefined character layout or class. With the world moving towards digital revolution, there has been increasing demand of application through use of which all documents of importance can be converted into editable document and can be stored online. This may be done with the purpose of storing valuable data online, so that it cannot be lost, or for saving time and labor demanded in converting handwritten or typed documents into editable document. Therefore modern society needs the input text into computer readable form.

Character Recognition is a common method of digitizing printed texts so that it can be electronically edited, searched, stored more compactly, displayed on-line, and used in machine processes such as machine translation, text-to-speech, key data and text mining.

This method is a simple approach to convert the input text into computer readable form, which is derived from ideas of different researchers who has given their valuable contribution in developing algorithms for OCR. Some research for hand written characters has been also done by researchers with artificial neural networks.

Digital document processing is gaining popularity for application to office and library automation, bank and postal services, publishing houses and communication technology.

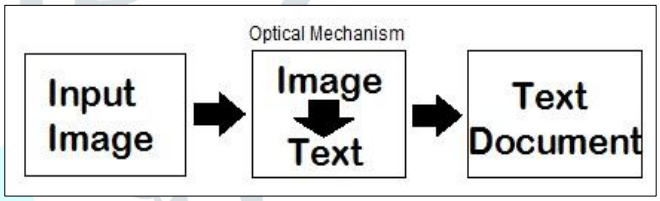
A. OCR

OCR is the acronym for Optical Character Recognition. This technology allows user to automatically recognize text through an optical mechanism. Eyes perform the optical Mechanism in case of human being. The image seen by eyes is input to brain. The ability of understanding these inputs varies from person to person in accordance with many factors. OCR is a technology that functions like human ability of reading. OCR helps to edit and search the words in the scanned documents like paper, PDF files and images. OCR is depicted from the human tendency of seeing and analyzing the characters but it is not matched with human reading capabilities.

Fig. 1 – OCR Process

B. MATLAB

MATLAB (Matrix Laboratory) is a numerical computing



environment and fourth generation programming language. MATLAB allows plotting of functions, matrix manipulation, implementation of algorithms, creation of user interface and many more functions. For this project, MATLAB R2010a version is used.

C. IMAGE PROCESSING

Image processing is processing of images by the use of mathematical operations with the help of any form of signal processing for which input is an image.

It can be done with three types of processing technique:

- Digital Image Processing
- Optical Image Processing
- Analog Image Processing

Image Processing techniques involve treating the image as a two-dimensional signal and applying standard signal processing techniques to it. Image can also be processed as three-dimensional signals.

2. LITERATURE SURVEY

The first OCR was designed in 1965 based on technology proposed primarily by the Jacob Rainbow which was used by the United States Postal Services. Then in 1970s, Dr. Sinha of Indian Institute of Technology, Kanpur made efforts to propose pattern analysis system. In 1974, Ray Kurzwell developed Omni-font OCR which could recognize text printed in virtually any form. In 2000s, OCR was made available online as a service (WebOCR).

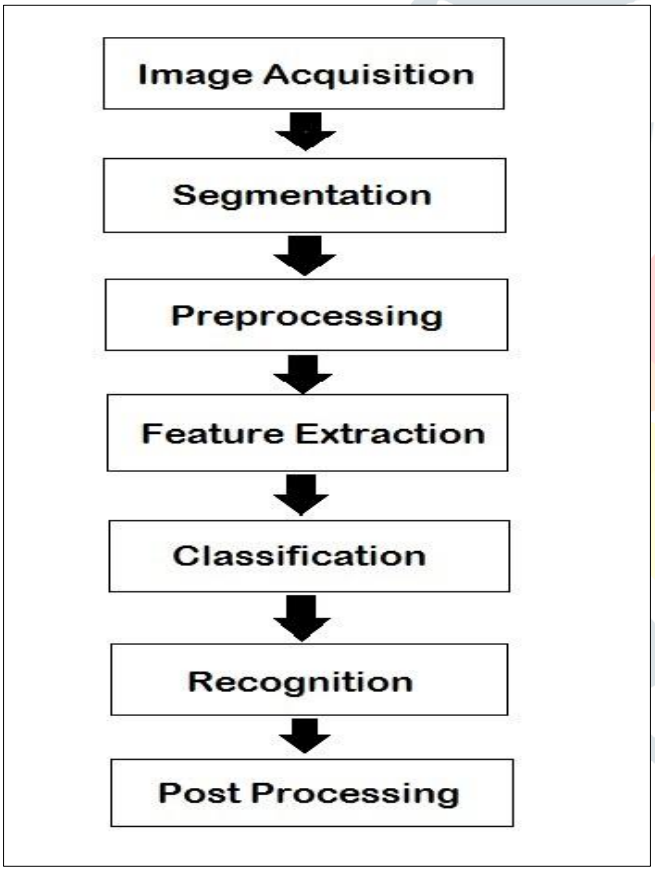
OCR system has been designed for most common writing system which includes Latin, Arabic, Indic, Bengali, Devanagri, Chinese, etc using most common programming languages MATLAB, ANN, LABVIEW, TESSERACT.

3. COMPONENTS OF OCR SYSTEM (Proposed OCR system)

OCR system is interconnection of several components which together perform the intended function. There are several steps involved in the process. The first step is to digitize the image or typed or handwritten document (analog document) using an optical scanner. In second step regions containing text are located, after that each symbol is extracted through a process of segmentation. In third step extracted symbols are processed to eliminate noise and to facilitate the extraction of features in the next step. In the upcoming step the identity of each symbol is found by comparing the extracted symbols with their features with defined descriptions of the symbol classes obtained through template collection process. At final, the recognized letters are used to form the words and numbers of the original text.

Fig.2 - General Offline Character Recognition System

A. IMAGE ACQUISITION



This step consists of optical scanning. In optical scanning, a digital image of typed or hand-written or image document is captured. To perform this process, optical scanners are used which are easily available in the market. Optical scanners convert light intensity into gray-levels. While performing OCR, the most common method of choice is to convert the multi -level image into bi-level image of black & white, which came to known as thresholding process. A fixed threshold is used for the above process. It is usually perform by the scanner to save memory space and computational effort. In thresholding image pixels which are lighter than threshold value (say 175) are converted into white pixels, i.e. 0, and the pixels which are darker than that are converted into dark pixels, i.e. 255. The process of thresholding makes it easy to convert the gray scale image into binary image.

Note: The above threshold values i.e. 175 & 255 are represented in octal system.

B. SEGMENTAION

It is a process of isolation of characters or words. This is done into several steps. Firstly, image is searched for the first dark pixel and then line segmentation takes place. In a line there can be several words, each containing several characters which are meant to be segmented. Secondly, words are searched in a segmented line and at last all the words in line are segmented. Thirdly, segmented words are searched for characters and each character is segmented. Hence, all the lines, words and characters in the image document is separated and saved for further processing.

This technique is easy to implement but faces problem whenever characters touch or where characters are fragmented and consists of several parts. Sometimes problems in segmentation can also arise from the font size, geometry of text and presence of unwanted marks.

C. PREPROCESSING

Preprocessing of image is done to improve the chances of successful recognition. In general noise filtering, smoothing and normalization is done in this step. Image which is scanned out in first step i.e acquisition may contain a certain amount of noise. This noise can lead to errors in the recognition process and should be removed or rectified. In most cases smoothing is considered as best option for this problem, smoothing implies a combination of filling and thinning process. Filling is the process of eliminating small breaks, gaps and holes in digitized or binarized characters. On the other hand 'Thinning' refers to the process in which width of line is reduced. Mostly preprocessing is done with help of some window (smoothing) filters.

Techniques of preprocessing include:

De-skew: aligns the document properly either few degrees in clockwise or counterclockwise direction so to make lines of text perfectly horizontal or vertical.

Despeckle: removing spots i.e. positive or negative spots and smoothing of edges.

Line removal: removes unnecessary lines.

D. FEATURE EXTRACTION

The step got its name because features of the characters that are crucial for classifying them at recognition stage are extracted here. To increase the recognition rate and reduce the misclassification, this stage needs to be effective.

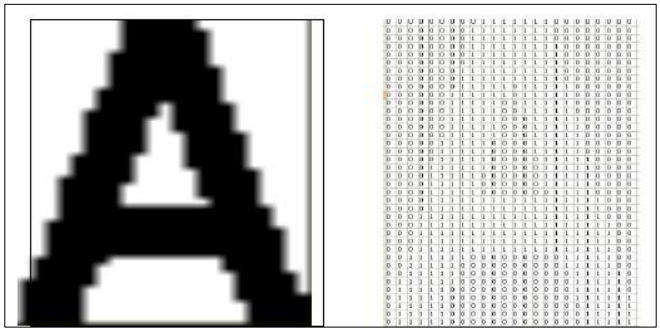


Fig. 3 - Character extraction in form of Matrix

MATLAB supports mat2cell command for extraction of image in form of a cell, so that it can be correlated with the saved templates.

E. CLASSIFICATION AND RECOGNITION

The decision making part of the recognition system is the classification stage. Here the matrix containing the image of input characters is directly matched with the set of prototype characters representing each possible class. The class of prototype giving the best match is assigned to the pattern. But this technique suffers from noise and style variations.

F. POST PROCESSING

Being final stage of system, it prints corresponding recognized characters in structured text form. From above processes we get set of individual symbols, but these symbols in themselves do not contain useful information. But once associated in the form of string or word they can convey a meaning. Therefore individual symbols are associated that belong to same string with each other, thus making up words and numbers, this is commonly known as grouping.

Even the best character recognition systems will not give 100 % correct results or identification of all the included characters, but some errors may be there which may b e detected and corrected further.

Note: Pre-processing and segmentation can be perform in reverse order, i.e. pre-processing then segmentation.

4. WORKING OF SYSTEM

A. DATABASE GENERATION (TEMPLATES)

Database for the OCR system are collected from various sources and are saved for matching with the characters which are to be recognized. Templates are actually images of characters, which are recognized by their class and structure and are used for matching and comparison.

B. CONVERTING COLOUR IMAGE TO GRAY SCALE IMAGE

In today's world, almost all scanning and image capturing devices use color. Color images matrices are labeled as red(R), green (G) and blue (B). Techniques provided up there in this proposed system are based on grey scale images and therefore, there is need of initially converting scanned or captured color images to grey scale.

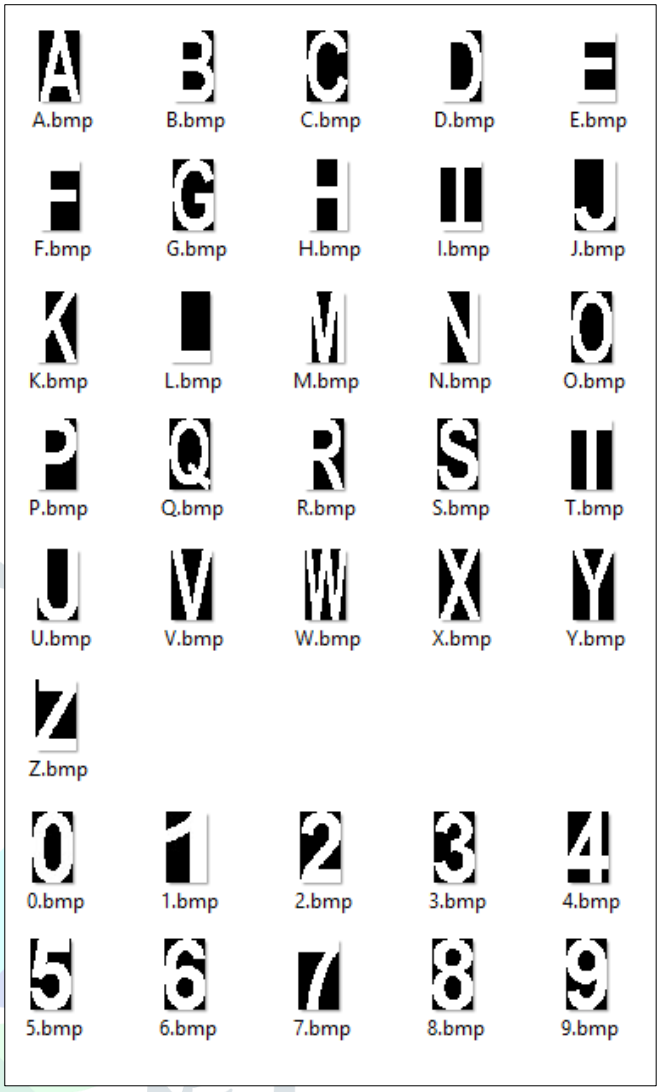


Fig. 4 – Training samples

C. BINARISATION

The process of converting a gray scale image (0 to 255 pixel values) into binary image (0 and 1pixel values) by the selection of threshold value in between 0 and 255 (here threshold value is 175).

D. CONVERTING IMAGE TO ARRAY

Image is converted into array with the help of mat2cell command, which form the array of 0 and 1's.

E. IMAGE SEGMENTATION

Three steps are involved in image segmentation process.

(i) LINE SEGMENTATION

It detects and clips off the line from the document, which is further used for word detection and segmentation.

(ii) WORD SEGMENTATION

It detects and clips off the word from the line, which is further used for character detection and segmentation.

(iii) CHARACTER SEGMENTATION

It detects and clips off the characters from the word, which is further used for matching and recognition.

5. RESULTS

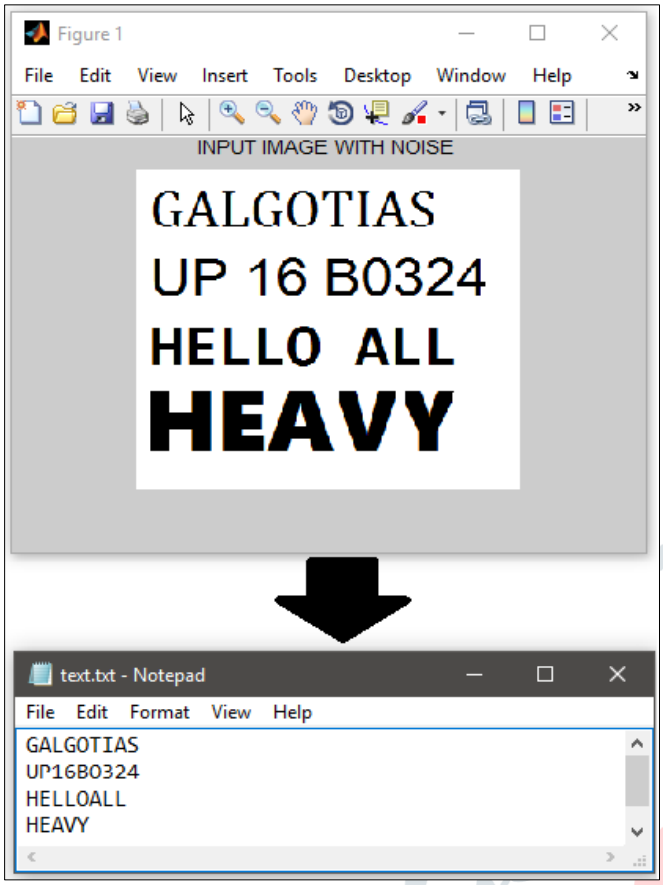


Fig. 5 – Text image sample and its output

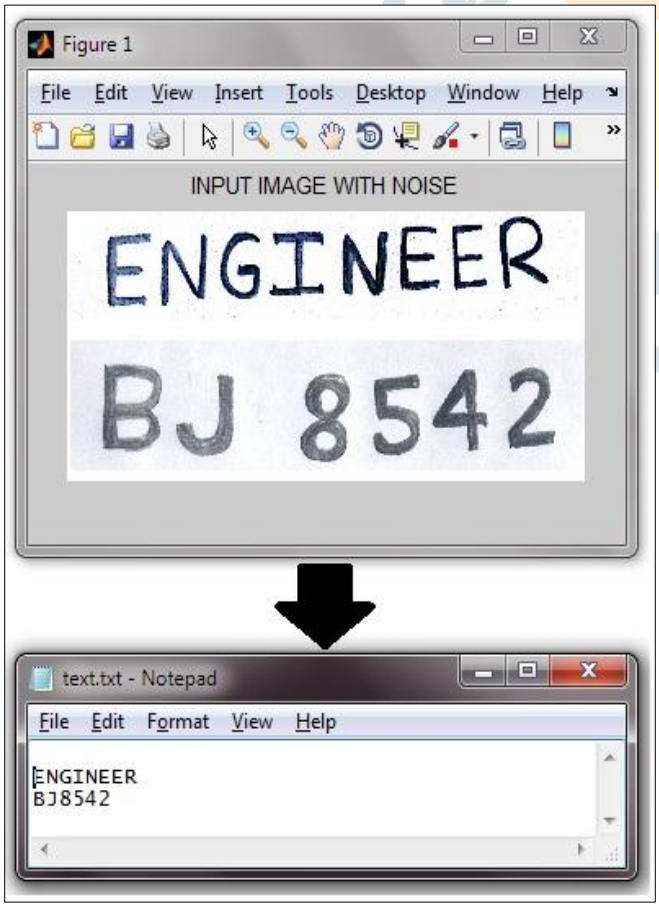


Fig. 6 – Handwritten image sample and its output

6. CONCLUSION

Optical Character recognition is the process of recognition of optically processed characters. With OCR it has become easier to interpret texts from real-world photos through continuously developing OCR engines and system which are easily available and open source.

A. ADVANTAGES

- The approach uses Mathworks MATLAB.
- MATLAB (Matrix Laboratory) is paid but economic, less costly platform, which is multitasking and very convenient to use.
- Early OCR systems requires expensive scanners with special purpose electronic and optical hardware, but this approach uses simple scanners or cameras either special purpose or that come with mobile phones.
- It can be also framed with GUI interface using MATLAB.

B. LIMITATIONS

- Accuracy of OCR systems is subject to or directly dependent on the quality of the input document.
- As image input to OCR is Noisy and garbled in most cases, the application will perform some post processing.
- The spaces and punctuation marks are not recognized by the proposed system.

C. FUTURE SCOPE

OCR has seen continuous development in previous decades with successful development of OCR engines for most known languages in the world i.e. for Latin, Devanagri, Nepali, Chinese, etc. The upcoming work may include a application which can perform combined character recognition and translation to other language. So that it can prove helpful for people from other countries that face difficulties understanding the local language.

7. APPENDIX

- OCR- **O**ptical **C**haracter **R**ecognition.
- GUI- Graphical **U**ser **I**nterface.
- LABVIEW - **LAB**oratory **V**irtual **I**nstrumentation **E**ngineering **W**orkbench.
- ANN- **A**rtificial **N**eural **N**etwork.

ACKNOWLEDGMENT

It is a matter of great pleasure and privilege to publish this review paper on "OPTICAL CHARACTER RECOGNITION" under the valuable guidance of Prof. Ankit Sharma. We would like to express my deep sense of gratitude to my guides for this valuable guidance, advice and constant aspiration to our work. F.A. Author thanks to our guides, who has helped us a lot on the completion for our project work.

REFERENCES

[1] Sheetal A. Nirve and G. S. Sable, "Optical Character Recognition for printed text in Devanagari using ANFIS" published under International Journal of Scientific & Engineering Research (IJSER), Vol. 4, Issue 10, October 2013, ISSN 2229-5518.

[2] Jesse Hansen, "A MATLAB project in Optical Character Recognition (OCR)".

[3] Jagruti Chandarana, Mayank Kapadia, "Optical Character Recognition" published under International Journal of Emerging Technology and Advanced Engineering (IJETAE), Vol. 4, Issue 5, May 2014, ISSN 2250-2459.

[4] Kauleshwar Prasad, Devvrat C. Nigam, Ashmika Lakhotiya and Dheeren Umre, "Character Recognition Using Matlab's Neural Network Toolbox", published under International Journal of u- and e- Service, Science and Technology, Vol. 6, Issue 1, February 2013.

[5] Sandeep Tiwari, Shivangi Mishra, Priyank Bhatia, Praveen Km. Yadav, "Optical Character Recognition using MATLAB", published under International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE), Vol. 2, Issue 5, May 2013, ISSN: 2278 – 909X.

[6] Neetu Bhatia, "Optical Character Recognition Techniques: A Review" published under International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), Vol. 4, Issue 5, May 2014, ISSN: 2277 128X.

[7] Ravina Mithe, Supriya Indalkar, Nilam Divekar, "Optical character recognition" published under International Journal of Recent Technology and Engineering (IJRTE), Vol. 2, Issue 1, March 2013, ISSN: 2277-3878.