ISSN (Online) 2394-6849

International Journal of Engineering Research in Electronics and Communication Engineering (IJERECE)
Vol 3, Issue 11, November 2016

# A Review of the Optical Character Recognition Methodology in Image Processing Techniques

[1]Dr. T.C.Manjunath [2] Pavithra G [3] M.R. Prasad
[1] Ph.D. (IIT Bombay), FIETE, FIE, CE, SMIEEE
Professor & HOD, ECE Dept., Dayananda Sagar College of Engg., Bangalore.
[2] B.E., M.Tech., [Ph.D.-VTU-Pursuing] Research Scholar, VTU – RRC, Belagavi, Karnataka
[3] B.E., M.Tech., [Ph.D.-VTU-Pursuing] Research Scholar, VTU – RRC, Belagavi, Karnataka
& Asst. Prof., CSE Dept., JSSATE, Bangalore, Karnataka

***Abstract:* -- This review paper deals with a review of the optical character recognition using the Matlab software. OCR or optical character recognition is the process of converting scanned images of machine-printed or handwritten text (numerals, letters, and symbols) into a computer-processable format; also known as optical character recognition (OCR). A typical OCR system contains three logical components, viz., an image scanner, OCR software and hardware, and an output interface. OCR systems can largely be grouped into two categories: task-specific readers and general-purpose page readers. A task-specific reader handles only specific document types. Such as bank checks, letter mail, or credit-card slips. General-purpose page readers are designed to handle a broader range of documents such as business letters, technical writings, and newspapers. The relevant literature presented here will be of an extensive use to many of the readers who are pursuing a research in the design & development of optical character recognition systems**

***Keywords*:-- OCR, Scanner, document, task-specific readers, General-purpose page readers.**

## I. INTRODUCTION

Optical character recognition, usually abbreviated to OCR, is the mechanical or electronic translation of images of handwritten, typewritten or printed text (usually captured by a scanner) into machine-editable text. OCR is a field of research in pattern recognition, artificial intelligence and computer vision. Though academic research in the field continues, the focus on OCR has shifted to implementation of proven techniques.

Optical character recognition (using optical techniques such as mirrors and lenses) and digital character recognition (using scanners and computer algorithms) were originally considered separate fields. Because, very few applications survive that use true optical techniques, the OCR term has now been broadened to include digital image processing as well. Early systems required training (the provision of known samples of each character) to read a specific font. "Intelligent" systems with a high degree of recognition accuracy for most fonts are now common.

Some systems are even capable of reproducing formatted output that closely approximates the original scanned page including images, columns and other non-textual components. The concept of optical character recognition was introduced in 1929 by G. Tauschek who obtained a patent on OCR in Germany, followed by Handel who obtained a US patent on OCR in USA in 1933.In 1950, David Shepard, built a machine along with Harvey cook called Gismo.

Shepard then founded Intelligent Machines Research Corporation (IMR), which went on to deliver the world's first several OCR systems used in commercial operations. J. Rainbow developed a prototype machine in 1954 that was able to read uppercase typewritten output at the "fantastic" speed of one character per minute. IBM, Recognition Equipment, Inc., Farrington, control Data, and Optical Scanning Corp, marketed OCR systems by 1967.

NASA used imaging system to enhance and manipulate satellite images. The purpose of this OCR is to take English handwritten documents as input, recognize the text and modify the handwriting such that it is a beautified version of the input. Any optical character recognition system consists of two parts, viz., the handwriting recognition & the next one being the beautification part.

## II. HANDWRITING RECOGNITION

Handwriting recognition can be broken into a number of relatively independent modules. After going through several papers and web pages on handwritten word recognition, I thought of various strategies for each of these modules. I then considered the accuracy and efficiency of these strategies independently and as a whole & presented the reviews here. Also, I some ideas based on feature extraction and relative position matching with the help of directional graphs are presented. In our future work, we aim at implementing this idea keeping in mind the strategies which we considered best for overall accurate, efficient and scalable handwriting recognition software.

## III. HANDWRITING BEAUTIFICATION

Due to the individual differences in handwriting, a 100% accurate handwriting recognition software has not yet been developed. Human reading ability is dependent on various factors including the knowledge of grammar, context, etc. Hence, it is difficult to build artificial methods to recognize documents as accurately as humans. Also, even handwriting recognition software with dictionary and context-understanding capabilities cannot guarantee accuracy in recognizing proper nouns. Due to these limitations, a number of scheme can be developed which will be considered in the latter part of this review paper. The handwriting consists of slants, curvatures, jointing, etc. The output of this can be used by humans to read deformed document.

## IV. OCR: A REVIEW

The process of converting scanned images of machine printed or hand written text such as numerals, letters and symbols into a computer processable format (like machine editable text or encoding scheme such as ASCII or Unicode) is called as OCR. The basic components of OCR are

- Image Scanner :
  Has 4 components a detector, An illumination source, A scan lens and a document transport.
- OCR hardware/software :
  performs three operational steps:
  - Document analysis,
  - Character recognition
  - Contextual processing
- Output Interface:
  Allows character recognition results to be electronically transferred into the domain that uses the result

## V. METHODOLOGY EMPLOYED

OCR systems consist of four major stages:

Pre-processing
↓
Segmentation
↓
Feature extraction
↓
Post-processing

## VI. PRE-PROCESSING

The raw data is subjected to a number of preliminary processing steps to make it usable in the descriptive stages of character analysis. Pre-processing aims to produce data that are easy for the OCR systems to operate accurately.

The main objectives of pre-processing are:

- ♣ Binarization
- ♣ Noise reduction
- ♣ Stroke width normalization
- ♣ Skew correction
- ♣ Slant removal

Binarization : Document image binarization (thresholding) refers to the conversion of a Gray-scale image into a binary image.

***Two categories of thresholding:***

- Global, picks one threshold value for the entire document image which is often based on an estimation of the background level from the intensity histogram of the image.
- Adaptive (local), uses different values for each pixel according to the local area information

***Skew Correction*** : Skew Correction methods are used to align the paper document with the coordinate system of the scanner. Main approaches for skew detection include correlation, projection profiles, Hough transform, time-frequency distributions of Cohen's class to the horizontal projection profile. In this paper, we describe the existing methods already available in the literature such as the projection profile technique and the Cohen distributions.

***The steps of the OCR can be explained as follows:***

1. The document is rotated to the right and left in steps of $12^0$ within the range $\pm 90^0$. For each angle the horizontal projection profile is extracted.

2. The Cohen distribution is calculated for each projection, as well as the maximum intensity of the distribution.

3. The angle whose histogram presents the maximum intensity is selected and the document is rotated by this angle (angle1).

4. The procedure is repeated from step 1 to step 3 but this time the document is rotated within a smaller range(around angle1) by one degree at a time and a more exact angle is calculated (angle2).

5. The estimated skew angle is angle1+ angle2.

***Slant removal :***

After getting a thinned image, it is scanned from left to right and all the vertical/tilted lines in the whole image are detected.

The average of slants of these lines gives us the slant of the image.

## VII. SEGMENTATION

In this step, the words or characters are extracted by using 2 types of segmentations, Implicit Segmentation & explicit segmentation. In implicit approaches the words are recognized entirely without segmenting them into letters. This is most effective and viable only when the set of possible words is small and known in advance, such as the recognition of bank checks and postal address. In Explicit Segmentation, explicit approaches one tries to identify the smallest possible word segments (primitive segments) that may be smaller than letters, but surely cannot be segmented further. Later in the recognition process these primitive segments are assembled into letters based on input from the character recognizer. The advantage of the first strategy is that it is robust and quite straightforward, but is not very flexible.

## VI. FEATURE EXTRACTION AND CHARACTER RECOGNITION

***Features:*** These are a predefined set of elementary parts which, when combined in various fashions, constitute all the characters. Some examples of features are vertical line, horizontal line, left curve, right curve, etc. Each feature has an associated set of x and y co-ordinates w.r.t. the OCR.

***Characters*:** Depending on the relative position of features in it, a character can be viewed as a Directional Graph of Features called feature-graphs. In feature extraction stage, each character is represented as a feature vector, which becomes its identity. The major goal of feature extraction is to extract a set of features, which maximizes the recognition rate with the least amount of elements.

## VII. POST-PROCESSING

***Goal :*** the incorporation of context and shape information in all the stages of OCR systems is necessary for meaningful improvements in recognition rates. The simplest way of incorporating the context information is the utilization of a dictionary for correcting the minor mistakes.

## IX. APPLICATIONS

A special type of OCR, magnetic ink character recognition (MICR), is used by several industries, including banks. The enormous amount of paper in the form of checks, loans, and bank statements, combined with the need for accurate and quick processing, prompted the banking industry to seek new ways to manage the flow of paper. In 1956, the American Bankers Association recommended adopting magnetic ink for high-speed automatic character recognition, resulting in MICR. With MICR, data are recorded using a magnetic ink that is readable by either a scanning device or a person. On bank checks, which represent the most common use of MICR, characters in magnetic ink detail the bank’s identification number, the individual’s account number, and the check number.

Checks can be scanned and the data are quickly and accurately read into a computer for further processing. Another use of OCR allows printed documents—such as text, images, or photographs—to be stored in a computer. Either hand-held scanners or page scanners are used to convert physical documents into computer-readable forms. Page scanners are stationary. The page is typically placed face down on the glass plate of the scanner and then scanned. Hand-held scanners are manually moved over the document.

## VII CONCLUSION

A brief review of the optical character recognition techniques was presented in this paper.

## REFERENCES

[1]. Schantz, Herbert F. The History of OCR, Optical Character Recognition. Manchester Center, VT: Recognition Technologies Users Association, 1982.

[2]. Shelly, Gary B., and Thomas J. Cashman. Introduction to Computers and Data Processing. Brea, CA: Anaheim Publishing Company, 1980.

[3]. Stair, Ralph M., and George W. Reynolds. Principles of Information Systems: A Managerial Approach, 5th ed. Boston: Course Technology—ITP, 2001.

[4]. Asha Tarachandani and Pooja Nath, “Address block recognition & beautification software,” IIT Kanpur, India, April 2000.

[5]. http://en.wikipedia.org/wiki/Optical_character_recognition