ISSN: 2278 – 909X
International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE)
Volume 2, Issue 5, May 2013

# Optical Character Recognition using MATLAB

Sandeep Tiwari, Shivangi Mishra, Priyank Bhatia, Praveen Km. Yadav
Electronics & communication Department
Kanpur Institute Of Technology, Kanpur

**Abstract --** Character recognition techniques associate a symbolic identity with the image of character. In a typical OCR systems input characters are digitized by an optical scanner. Each character is then located and segmented, and the resulting character image is fed into a pre-processor for noise reduction and normalization. Certain characteristics are the extracted from the character for classification. The feature extraction is critical and many different techniques exist, each having its strengths and weaknesses. After classification the identified characters are grouped to reconstruct the original symbol strings, and context may then be applied to detect and correct errors.

***Index Terms*—corr2, feature extraction, mat2cell, pixels, segmentation.**

## I. INTRODUCTION

Machine replication of human functions, like reading, is an ancient dream. However, over the last five decades, machine reading has grown from a dream to reality. Optical character recognition has become one of the most successful applications of technology in the field of pattern recognition and artificial intelligence. Many commercial systems for performing OCR exist for a variety of applications, although the machines are still not able to compete with human reading capabilities. Optical Character Recognition deals with the problem of recognizing optically processed characters. Optical recognition is performed off-line after the writing or printing has been completed, as opposed to on-line recognition where the computer recognizes the characters as they are drawn. Both hand printed and printed characters may be recognized, but the performance is directly dependent upon the quality of the input documents. Progress in OCR has been steady if not spectacular since its commercial introduction at the Reader's Digest in the mid-fifties. After specially-designed typefaces, such as OCR-A, OCR-B, and Farrington 14B came support for elite and pica (fixed-pitch) typescripts, then "omnifont" typeset text. In the last decade the acceptance rates of form readers on hand-printed digits and constrained alphanumeric fields has risen significantly (form readers usually run at a high reject/error ratio). Many researchers now view off-line and on-line cursive writing as the next challenge or turn to multi-lingual recognition in a variety of scripts. Character classification is also a favourite testing ground for new ideas in pattern recognition, but since most of the resulting experiments are conducted on isolated characters, the results are not necessarily immediately relevant to OCR. Perhaps more striking than the improvement of the scope and accuracy in classification methods has been the decrease in cost. The early OCR devices all required expensive scanners and special-purpose electronic or optical hardware: the IBM 1975 Optical Page Reader for reading typed earnings reports at the Social Security Administration cost over three million dollars (it displaced several dozen keypunch operators). The almost simultaneous advent about 1980 of microprocessors for personal computers and of charge-coupled array scanners resulted in a huge cost decrease that paralleled that of general-purpose computers .Today, shrink-wrapped OCR software is often an add-on to desktop scanners that cost about the same as a printer or facsimile machine. Our purpose is to examine in some detail examples of the errors committed by current OCR systems and to speculate about their cause and possible remedy.

## II. COMPONENTS OF AN OCR SYSTEM

A typical OCR system consists of several components. In figure 1. a common setup is illustrated. The first step in the process is to digitize the analog document using an optical scanner. When the regions containing text are located, each symbol is extracted through a segmentation process. The extracted symbols may then be preprocessed, eliminating noise, to facilitate the extraction of features in the next step. The identity of each symbol is found by comparing the extracted features with descriptions of the symbol classes obtained through a previous learning phase. Finally contextual information is used to reconstruct the words and numbers of the original text.

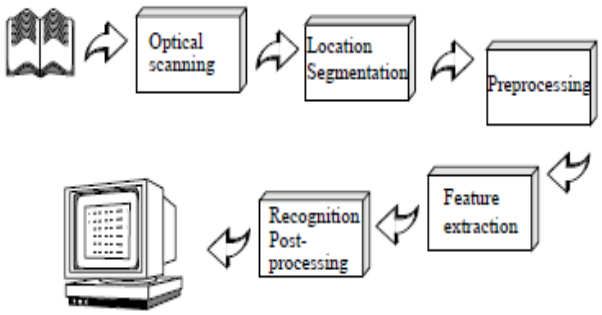


Fig.1. Components of an OCR System

### A. *Optical scanning*

Through the scanning process a digital image of the original document is captured. In OCR optical scanners are used, which generally consist of a transport mechanism plus 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, known as thresholding, is performed on the scanner to save memory space and computational effort. The thresholding process is important as the results of the following recognition is

ISSN: 2278 – 909X
International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE)
Volume 2, Issue 5, May 2013

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 prechosen fixed threshold can be sufficient. However, a lot of documents encountered in practice have a rather large range in contrast.

### *B. Location and segmentation*

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:

- *Extraction of touching and fragmented characters.*
- *Distinguishing noise from text.*
- *Mistaking graphics or geometry for text.*
- *Mistaking text for graphics or geometry.*

### *C. Preprocessing*

The image resulting from the scanning process may contain a certain amount of noise. 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. The most common techniques for smoothing, moves a window across the binary image of the character, applying certain rules to the contents of the window. 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 be found. For rotated pages and lines of text, variants of Hough transform are commonly used for detecting skew.

### *D. Feature Extraction*

The techniques for extraction of such features are often divided into three main groups, where the features are found from:

• The distribution of points.
• Transformations and series expansions.
• Structural analysis.

In MATLAB mat2cell command is used for the extraction of image in form of a cell for correlating with the saved templates.Fig.2 shows extraction of character in Matrix form.

### *E. Template-matching and correlation techniques*

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 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.

Fig.2.(a) Character extraction in form of Matrix

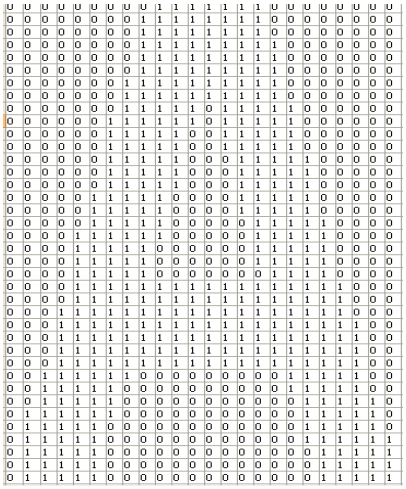

Fig.2.(b) Character extraction in form of Matrix

### *F. Post Processing*

It encompasses grouping, error detection and correction techniques. The result of plain symbol recognition on a document, is a set of individual symbols. However, these symbols in themselves do usually not contain enough information. Instead we would like to associate the individual symbols that belong to the same string with each other, making up words and numbers. The process of performing this association of symbols into strings, is commonly referred to as grouping. The grouping of the symbols into strings is based on the symbols location in the document. Symbols that are found to be sufficiently close are grouped together Up until the grouping each character has been treated separately, and the context in which each character appears has usually not been exploited. However, in advanced optical text recognition problems, a system

ISSN: 2278 – 909X
International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE)
Volume 2, Issue 5, May 2013

consisting only of single-character recognition will not be sufficient. Even the best recognition systems will not give 100% percent correct identification of all characters, but some of these errors may be detected or even corrected by the use of context.

## III. WHY MATLAB?

MATLAB stands for MATrixLABoratory. Here you play around with matrices. Hence, an image (or any other data like sound, etc.) can be converted to a matrix and then various operations can be performed on it to get the desired results and values. Image processing is quite a vast field to deal with. We can identify colors, intensity, edges, texture or pattern in an image. In this tutorial, we would be restricting ourselves to detecting colours (using RGB values) only. Using MATLAB you can solve technical computing problems faster than with traditional programming language, such as C, C++, JAVA, FORTRAN. There is a wide range of applications, including signal and image processing, image accusation, Neural Network, etc.

## IV. OCR PERFORMANCE EVALUATION

No standardized test sets exist for character recognition, and as the performance of an OCR system is highly dependent on the quality of the input, this makes it difficult to evaluate and compare different systems. Still, recognition rates are often given, and usually presented as the percentage of characters correctly classified. However, this does not say anything about the errors committed. Therefore in evaluation of OCR system, three different performance rates are investigated:

• ***Recognition rate.***

The proportion of correctly classified characters.

• ***Rejection rate***.

The proportion of characters which the system were unable to recognize. Rejected characters can be flagged by the OCR-system, and are therefore easily retraceable for manual correction.

• ***Error rate.***

The proportion of characters erroneously classified. Misclassified characters go by undetected by the system, and manual inspection of the recognized text is necessaryto detect and correct these errors. There is usually a tradeoff between the different recognition rates. A low error rate may lead to a higher rejection rate and a lower recognition rate. Because of the time required to detect and correct OCR errors, the error rate is the most important when evaluating whether an OCR system is cost-effective or not. The rejection rate is less critical. An example from barcode reading may illustrate this. Here a rejection while reading a barcoded price tag will only lead to rescanning of the code or manual entry, while a misdecoded price tag might result in the customer being charged for the wrong amount. In the barcode industry the error rates are therefore as low as one in a million labels, while a rejection rate of one in a hundred is acceptable. In view of this, it is apparent that it is not sufficient to look solely on the recognition rates of a system. A correct recognition rate of 99%, might imply an error rate of 1%. In the case of text recognition on a printed page, which on average contains about 2000 characters, an error rate of 1% means 20 undetected errors per page. In postal applications for mail sorting, where an address contains about 50 characters, an error rate of 1% implies an error on every other piece of mail.

## V. RESULTS

To illustrate the accuracy of proposed English handwritten and sample text images OCR algorithm by using MATLAB, performance was measured using the samples. Figure 3 and 4 shows the sample document scanned from HP deskjet scanner at 300 dpi. The images were then filtered, binarized, clipped and resized. Lines of text were then extracted from the images. The font size was identified; segmentation was performed on each line to segment characters taking in consideration the characteristics of English Verdana fonts templates. MATLAB (R2012.a/64-bit) is used to implement the proposed OCR algorithm. The recognition accuracy was 85% to 90% due to improper hand written characters. The templates of all Characters and numbers are of 24X42 pixels.

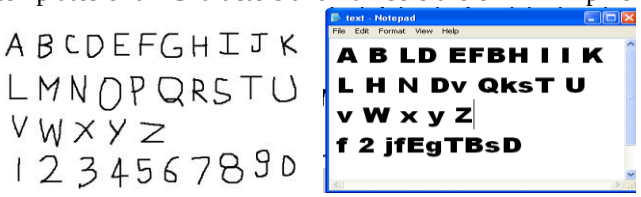


Fig.3. Handwritten Sample And its output

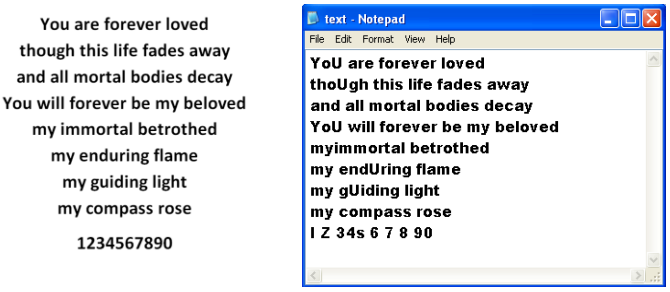


Fig.4. Text Image Sample And its output

## VI. FUTURE SCOPE

New methods for character recognition are still expected to appear, as the computer technology develops and decreasing computational restrictions open up for new approaches. There might for instance be a potential in performing character recognition directly on grey level images. However, the greatest potential seems to lie within the exploitation of existing methods, by mixing methodologies and making more use of context. Integration of segmentation and contextual analysis can improve recognition of joined and split characters. Also, higher level contextual analysis which look at the semantics of entire sentences may be useful. Generally there is a potential in using context to a larger extent than what is done today. In addition, combinations of multiple independent feature sets and classifiers, where the weakness of one method is compensated by the strength of

another, may improve the recognition of individual characters. The frontiers of research within character recognition have now moved towards the recognition of cursive script, that is handwritten connected or calligraphic characters. Promising techniques within this area, deal with the recognition of entire words instead of individual characters.

## VII. CONCLUSION

Today optical character recognition is most successful for constrained material, that is documents produced under some control. However, in the future it seems that the need for constrained OCR will be decreasing. The reason for this is that control of the production process usually means that the document is produced from material already stored on a computer. Hence, if a computer readable version is already available, this means that data may be exchanged electronically or printed in a more computer readable form, for instance barcodes. The applications for future OCR-systems lie in the recognition of documents where control over the production process is impossible. This may be material where the recipient is cut off from an electronic version and has no control of the production process or older material which at production time could not be generated electronically. This means that future OCR-systems intended for reading printed text must be omnifont. Another important area for OCR is the recognition of manually produced documents. Within postal applications for instance, OCR must focus on reading of addresses on mail produced by people without access to computer technology. Already, it is not unusual for companies etc., with access to computer technology to mark mail with barcodes. The relative importance of handwritten text recognition is therefore expected to increase.

## ACKNOWLEGEMENT

It's a pleasure and a great blessing of GOD for working on the project named "Optical Character Recognition Using MATLAB". Wherein we gained knowledge by working under the able leadership of our **Head of department Mr. Vaibhav Purwar** who helped and supported us in every sphere of our project. We all thank our **Project in- charge Mr. Asheesh Gupta** who well supported us and provided us with his precious time and support. It would we gracious to thank our **supervisor Mr. Gaurav Porwal** for his valuable advice and help which he provided us throughout the whole duration of this project. Besides that we thank whole of the E.C. department for their appreciation and kind support.

## REFRENCES

[1] H.S. Baird & R. Fossey.A 100-Font Classifier.Proceedings ICDAR-91, Vol. 1, p. 332-340, 1991.
[2] R. Bradford & T. Nartker.Error Correlation in Contemporary OCR Systems.Proceedings ICDAR-91, Vol. 2, p. 516-524, 1991.
[3] J-P. Caillot.Review of OCR Techniques. NR-note, BILD/08/087.
[4] R. G. Casey & K. Y. Wong.Document-Analysis Systems and Techniques.Image Analysisi Applications, eds: R. Kasturi& M. Tivedi, p. 1-36.
[5] Product help: http://www.mathworks.com/pl_homepage

## AUTHORS

Sandeep Tiwari is currently pursuing his B.Tech (Final year) in Electronics and Communication from Kanpur Institute of Technology (G.B.T.U). His main areas of Interest are MATLAB, Electronic devices, Optical Communications.

Shivangi Mishra is currently pursuing her B.Tech (Final year) in Electronics and Communication from Kanpur Institute of Technology (G.B.T.U). Her areas of interest are wireless Networks, MATLAB.

Priyank Bhatia is currently pursuing his B.Tech (Final year) in Electronics and Communication from Kanpur Institute of Technology (G.B.T.U). His areas of interest are wireless Networks, MATLAB and sensors.

Praveen Km. yadav is currently pursuing his B.Tech (Final year) in Electronics and Communication from Kanpur Institute of Technology (G.B.T.U). His areas of interest are Basic Electronics, MATLAB.