IJESRT
INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

# A DETAILED STUDY AND ANALYSIS OF OCR USING MATLAB

**Dhiraj Kumar Jasrotia*[1] & Aarti Malik[2]**
*[1]M.TECH., ECE, Electronics And Communication Engineering, Gurukul, Banur
[2] ASSISTANT PROFESSOR, ECE, Electronics And Communication Engineering, Gurukul, Banur

**DOI**: Will get Assigned by IJESRT Team

## ABSTRACT

This paper presents detailed review in the field of Optical Character Recognition. Various techniques are determine that have been proposed to realize the center of character recognition in an optical character recognition system. Even though, sufficient studies and papers are describes the techniques for converting textual content from a paper document into machine readable form. Optical character recognition is a process where the computer understands automatically the image of handwritten script and transfer into classify character. This material use as a guide and update for readers working in the Character Recognition area. Selection of a relevant feature extraction method is probably the single most important factor in achieving high character recognition with much better accuracy in character recognition systems without any variation.
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.

**Keywords:** Neural Network, Feature extraction, Classification, OCR , Feature extraction, Segmentation.

## I. INRTODUCTION

Optical Character Recognition (OCR) is the process of translation of images of typewritten or handwritten text into machine editable text. A number of techniques of Pattern Recognition such as Template Matching, Neural Networks, Syntactical Analysis, Hidden Markov Models, Bayesian Theory, etc have been explored to develop robust OCRs for different languages. At present we have efficient and inexpensive OCR packages which are commercially available for the recognition of printed documents. Among those we have enough facilities for languages such as English, Chinese etc. Even handwritten document recognition facilities are widely available for these languages. When considering the Indian languages many attempts are made to develop the OCR system for Devanagari, Oriya, Tamil [1], Telugu [2], and Kannada [3] etc. But this area needs further more developments and the researches are going in this field. The important practical applications of OCRs are seen in converting documents into electronic format, library catalogue, postal mail system, bank cheques and reading aid for the blind.

A lot of work has been carried out in the area of recognition of Roman, Chinese and Japanese language, while very few has been reported for Indian language recognition especially for the South Indian scripts.In order to make Indian language OCR familiar, Government agencies are providing much assistance and nowadays a lot of institutions are come forward to encourage the research work in this area. The TDIL (Technology Development for Indian Language) Programme, RCILTS (Resource for Indian Language Technology Solutions) and Ministry of Communications and Information Technology, Government of India are the main agencies which stress the need of efficient OCR in Indian languages. They set up many research centers and funding many projects in this area.

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. LITERATURE REVIEW

Many methods have been proposed for character recognition. But they are often subjected to substantial constraints due to unexpected difficulties. Historically character recognition system has evolved in three ages [2], namely the periods cited denoting as **1900-1980 (early ages) –** The history of character recognition can be traced as early as 1900. When the Russian Scientist Tyering attempted to develop an aid for visually handicapped. The first character recognizers appeared in the middle of 1940s with the development of digital computers. The early work on the automatic recognition of characters has been concentrated either upon machine printed text or upon small set of well distinguished hand written text or symbols. The commercial character recognizers were available in 1950s. **1980-1990 Developments –** The studies until 1980 suffered from the tack of powerful computer hardware and data acquisition derives. However, the character recognition research was focused on basically the shape recognition techniques without using any semantic information.
**After 1990 advancements –** The real progress on character recognition system is achieved during this period, using the new development tools and methodologies, which are empowered by continuously growing information technologies. In the early nineties, Image processing and Pattern recognition techniques are efficiently combined with the Artificial Intelligence methodologies. Nowadays in addition to the more powerful computers and more accurate electronic equipments such as scanners, cameras and electronic tablets, we have efficient, modern use of methodologies such as neural .

Many methods have been proposed for character recognition. But they are often subjected to substantial constraints due to unexpected difficulties. Historically character recognition system has evolved in three ages [2], namely the periods cited denoting as **1900-1980 (early ages) –** The history of character recognition can be traced as early as 1900. When the Russian Scientist Tyering attempted to develop an aid for visually handicapped. The first character recognizers appeared in the middle of 1940s with the development of digital computers. The early work on the automatic recognition of characters has been concentrated either upon machine printed text or upon small set of well distinguished hand written text or symbols. The commercial character recognizers were available in 1950s.

**1980-1990 Developments –** The studies until 1980 suffered from the tack of powerful computer hardware and data acquisition derives. However, the character recognition research was focused on basically the shape recognition techniques without using any semantic information.

**After 1990 advancements –** The real progress on character recognition system is achieved during this period, using the new development tools and methodologies, which are empowered by continuously growing information technologies. In the early nineties, Image processing and Pattern recognition techniques are

efficiently combined with the Artificial Intelligence methodologies. Nowadays in addition to the more powerful computers and more accurate electronic equipments such as scanners, cameras and electronic tablets, we have efficient, modern use of methodologies such as neural networks, Hidden Markov models; Fuzzy set reasoning and Natural language processing. Character recognition system is the base for many different types of applications in various fields, many of which we use in our daily lives. Post offices, banks, security systems, number plate recognition system and even in the field of robotics use this system as the base of their operations. **Character Recognition Approaches** Character recognition systems extensively use the methodologies of pattern recognition, which assigns an unknown sample to a predefined class. Many techniques for character recognition are investigated by the researchers and character recognition approaches can be classified as [3] Template matching, Statistical techniques, Syntactic or structural, Neural network, Hybrid or Combination approaches. **Template matching approach** This is the simplest way of character recognition, based on matching the stored data against the character to be recognized. The matching operation determines the degree of similarity between two vectors i.e. group of pixels, shapes curvature etc. a gray level or binary input character is compared to a standard set of stored data set. According to similarity measure (e.g. Euclidean, Yule similarity measures etc.), a template matcher can combine multiple information sources, including match strength and k-nearest neighbor measurements from different matrices. The recognition rate of this method is very sensitive to noise and image deformation. **Statistical Techniques** Statistical decision theory is concerned with statistical decision functions and a set of optimality criteria, which maximizes the probability of the observed pattern given the model of a certain class. Statistical techniques are based on the assumptions such as Distribution of the feature set, statistics available for each class, collection of images to extract a set of features which represents each distinct class of patterns. The measurements taken from n-features of each word unit can be thought to represent an n-dimensional vector space. The major statistical methods applied in the character recognition field are Nearest Neighbor Likelihood or Bayes classifier, clustering Analysis, Hidden Markov Modeling, Fuzzy Set Reasoning, Quadratic classifier etc. **Syntactic or Structural Approach**

In Syntactic Pattern recognition a formal analogy is drawn between the structure of pattern and syntax of a language. Structural pattern recognition is intuitively appealing because in addition to classification, this approach also provides a description of how the given path constructed from the primitives. Flexible structural matching is proposed for identification of alphanumeric characters. **Neural Networks** Various types of neural networks are used for character recognition classification. A neural network is a computing architecture that consists of massively parallel interconnection of adaptive neural processors. Because of its parallel nature, it can perform computations at a higher rate compared to classical techniques. Because of its adaptive nature, it can adapt to changes in the data and learn the characteristics of input signal. Output from one node is fed to another one in the network and final decision depends on the complex interaction of all nodes. Several approaches exist for training of neural networks like error correction, Boltzman, Hebbian and competitive learning. Neural network architectures can be classified as, feed-forward, feed-back and recurrent networks. The most common neural networks used in the character recognition systems are the Multi Layer Perceptron (MLP) of the feed forward networks and the Kohonen's Self Organizing Map of the feedback networks.

**Indian Character Recognition** Not many attempts have been made on the character recognition of Indian character sets. However, some major works are reported on Devanagari. Some attempts are also reported on Tamil, Kannada, Gujarathi, Bengali, Malayalam and Telugu.

Character recognition of handwritten and printed text is of great importance for electronic conversion of historical information including letters, diaries, wills and other manuscripts. The problem is challenging because of human handwriting variability, uneven skew and orientation as well as noise and distortion such as smudges, smears, faded print, etc. identification of handwritten Indian scripts especially of Bangla, as well as English, Hindi, Malayalam, etc. Most of the Indian scripts have 500 or more characters or symbols used in running text, through the number of basic vowels and consonants is not more than 50. The number is multiplied by three types of vowel modifiers that may be glued below the consonants, thus generating threefold consonant-vowel combinations. Further increase in number is possible where consonant creates a complex orthographic shape called compound characters. For some scripts like Bangla, Gujarthi, Telugu and Devanagari languages consists of large number of compound characters. These compound characters can also take vowel modifiers to generate threefold more shapes. Thus orthographic shapes may run of the order of thousand. Only Tamil and Punjabi scripts are relatively simpler, where the number of characters/ symbol is about 150 and 70 respectively. Most Indian script lines can be partitioned into three sub-zones. The upper and lower zones may consist of parts of the basic characters as well as vowel modifiers. These parts of two consecutive text lines normally do not overlap or touch in case of printed script, but for handwriting, people have the tendency to write them bigger, leading to overlapping and touching characters. Overall these characteristics make handwritten and printed Indian text recognition more challenging. The earlier research work on character recognition related to Indian languages is discussed below.

### Telugu Character Recognition

A two stage recognition method for printed Telugu characters is proposed by Rajasekaran and Deekshatulu [11] . In the first stage, the primitive shapes are removed from the given Telugu character using a syntax-aided recognition scheme. After removal of the primitives only basic shapes were left. In the second stage, the basic letters can be recognized using either the classificatory or syntactic approach. The basic letters are recognized by a process called On the Curve Coding. The classification is achieved by means of a decision tree by using the knowledge of the primitives and the basic characters that are present in the input pattern. The extraction of primitive technique, being a very complex one, takes more time for recognition processing.

In [10] Sukhaswamy et.al., proposed an approach to recognize Telugu script using Neural Networks. The author developed a network architecture called Multiple Neural Network Associative Memory (MNNAM) for recognition of Printed Telugu characters. In this method, the exemplars to be trained are divided into groups, each of which is having a capacity less than the practical optimal storage capacity (POSC) of a network. Each group is trained into a separate network of same topology. The test pattern to be recognized is presented to each of these networks. The patterns to which each of the networks converge are then made as exemplars to train further levels of networks, called as combination networks. The main limitation of their work is that the character recognition is not invariant of size, translation and rotation. In [34], the authors P V S Rao and TM Ajitha have suggested Telugu script recognition using a feature based approach. Recognition is based on segmenting the characters into the component elements and identifying them. Feature vector parameters for individual basic characters are extracted from single specimen written in isolation. These are suitably combined to construct feature vectors for compound characters. These are compared with similar feature vectors extracted from the test samples to be recognized.

An OCR system for Telugu is developed by Atul Negi et.al.,[8]. They identified a total set of 370 connected components that exhaust each of the five categories, vowel, Consonants, vowel signs, conjunct consonants, vowel or consonant, combination of a consonant and vowel sign. Template matching is used to recognize the components. Frinze distance is used as a distance measure for comparison of images. In [9] the authors presented a system to locate, extract and recognize Telugu text. First, the Hough Transform for circles is performed on the Sobel gradient magnitude of the image to locate text. The located circles are filled to yield text regions, followed by Recursive XY Cuts to segment the regions into paragraphs, lines and word regions. A region merging process with bottom-up approach envelopes individual words. Local binarization of the word MBRs yields connected components containing glyphs for recognition. The recognition process first identified candidate characters by a zoning technique and then constructs structural feature vectors by cavity analysis. Finally, if required, crossing count based non-linear normalization and scaling is performed before template matching. [13], the author proposed a method which uses wavelet multi resolution analysis for the purpose of extracting features and associative memory model to accomplish the recognition tasks. Wavelet Basis function is used to extract the invariant features of the characters and Hopfield-based Dynamic Neural Network model used for the purpose of learning and recognition. Modular Neural Networks approach and minimum number of primitives in the training process for the recognition of complete set of printed Telugu characters is presented by Sandhya Rani[12]. Extraction of primitives, identifying the class of primitives and recognizing the characters using Neural Networks techniques are the significant works. Srivani [11], proposed a novel technique for processing Printed Telugu document using Modular Neural Network approach. The developed system is robust with rotation; size and noise efficient feature extraction techniques are adopted and achieved high recognition accuracy.

### Tamil Character Recognition

The Tamil alphabet consists of 12 vowels and 18 consonants. These combine to form 216 compound characters. There is one special character (*aaytha ezutthu*), giving a total of 247 haracters. Unlike other Indian languages, Tamil has single glyphs for ka, cha, ta, tha, pa, ra. But their sounds vary depending on the context where they occur. However, with the advantage of having a separate symbol for each vowel in composite character formations, there is a possibility to reduce the number of symbols used by the alphabet. In character recognition point of view, only 67 symbols have to be identified to recognize all 247 characters. Some of the research works regarding Tamil character recognition are as follows. Anbumani et.al., in [14] proposed an Optical Character Recognition of Printed Tamil Characters. Author used statistical parameters during recognition stage.

- **Kannada Character Recognition**

The Kannada alphabet is classified into two main categories: vowels and consonants. There are 16 vowels and 35 consonants and words in Kannada are composed of *aksharas* which are analogous to characters in an English word. While vowels and consonants are *aksharas*, the vast majority of *aksharas* are composed of combinations of these in a manner similar to most other Indian scripts. A few research works on character recognition of kannada is as follows.

In [4], Kannada characters written on the digitizer pad are recognized in an interactive way is proposed by R. Srinivasa Rao and Sudhakar Samuel. A document file of recognized text in Kannada computer fonts is produced as one keeps writing the text on the digitizer pad. Wavelet transforms have been used to extract the features of the Kannada characters, which have turned to be more efficient features. The multilayer feed forward neural networks are used to recognize and classify the Kannada characters. A font and size-independent OCR system for printed Kannada documents using support vector machines is presented by Ashwin and Sastry[15]. In this work recognition is achieved by employing a number of 2-class classifiers based on the Support Vector Machine (SVM) method.

- **Bengali Character Recognition :**

Bengali alpha numeric character recognition is proposed by Dutta et.al.,[5]. In this work, Curvature related characteristics were used as features and back propagation based learning scheme was used in the recognition strategy enables the system to learn from examples. Other works related to Bengali are reported in [6]. Few other research works are presented in other Indian languages like Gujarati, Gurumukhi [7].

- **Hindi Character Recognition**

Devanagari script is alphabetic in nature and the words are two dimensional compositions of characters and symbols which makes it different from Roman and ideographic scripts. I K Sethi [1] described Devanagari numeral recognition based on the structural approach. The primitives used are horizontal line segment, vertical line segment, right slant and left slant. A decision tree is employed to perform the analysis based on the presence/absence of these primitives and their interconnection. A similar strategy was applied to constrained hand printed Devanagari characters. An OCR system for printed Devanagari script [2] is presented by Pal and Chaudhuri claims an accuracy of 95% at the character level. In this system some standard and some new techniques have been used for preprocessing. However, thinning has not been carried out on the images. From zonal information and shape characteristics, the basic, modified and compound characters are separated for convenience of classification. Modified and basic characters are recognized by a structural feature based binary tree classifier while the compound characters are aimed to be recognized by a hybrid approach.

## III. CONCLUSION

Character Recognition is one of the vital tasks in Pattern Recognition. The popularity and use of Character Recognition is increasing day by day with the advent of new, fast and efficient hardware and software. But automatic character recognition of Indian languages is still in preliminary stage and hence there is a need of lot of research to address the various issues and their complexities. There are many factors such as noise, various font sizes, broken lines or characters, quality of the image, problems in segmentation that influence recognition process. India is a multi lingual country; so many more efficient and real-time text recognizers are required. A good text recognizer has many commercial and practical applications. Hence there is a need to develop a very good character recognition system which must achieve highest accuracy.

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.

## REFERNCES

[1] G. Nagy, S. Seth, and M. Viswanathan, "A Prototype Document Image Analysis System for Technical Journals," Computer, vol. 25, no. 7, pp. 10-22, July 1992.

[2] Fujisawa, Yasuaki Nankano, and Kiyomichi Kurino "Segmentation Methods for Character Recognition: From Segmentation to Document Structure Analysis" Proceedings of The IEEE. Vol. 80. No. 7. July 1992.

[3] L. O'Gorman, "The Document Spectrum for Page Layout Analysis," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 15, no. 11, pp. 1162-1173, Nov. 1993.

[4] H.S. Baird, H. Bunke, P. Wang and H.S. Baird "Background Structure in Document Images," Document Image Analysis, eds., pp. 17-34, World Scientific, 1994.

[5] Sylwester and S. Seth, "A Trainable, Single-Pass Algorithm for Column Segmentation," Proc. Int'l Conf. Document Analysis and Recognition, pp. 615- 618, Aug. 1995.

[6] I. Guyon, R.M. Haralick, J.J. Hull and I.T. Phillips, "Data Sets for OCR and Document Image Understanding Research," Handbook of Character Recognition and Document Image Analysis, H. Bunke and P. Wang, eds., pp. 779-799, World Scientific, 1997.

[7] O. Okun, M. Pietikainen, and J. Sauvola, "Robust Skew Estimation on Low-Resolution Document Images," Proc. Fifth Int'l Conf. Document Analysis and Recognition, pp. 621-624, Sept. 1999.

[8] G. Nagy, "Twenty Years of Document Image Analysis in PAMI," IEEE Trans. Pattern Analysis and Machine Intelligence,Vol. 22, no. 1, pp. 38-62, Jan. 2000.

[9] Jianbo Shi and Jitendra Malik "Normalized Cuts and Image Segmentation" IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No. 8, Aug 2000.

[10] S. Mao and T. Kanungo, "Empirical Performance Evaluation Methodology and Its Application to Page Segmentation Algorithms," IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 23, no. 3, pp. 242-256, Mar. 2001

[11] S. Mao and T. Kanungo, "Software Architecture of PSET: A Page Segmentation Evaluation Toolkit," Int'l J. Document Analysis and Recognition, Vol. 4, no. 3, pp. 205-217, 2002.

[12] L. Cinque, S. Levialdi, L. Lombardi and S. Tanimoto, "Segmentation of Page Images Having Artifacts of Photocopying and Scanning," Pattern Recognition, Vol. 35, pp. 1167-1177, 2002.

[13] T.M. Breuel, "High Performance Document Layout Analysis," Proc. Symp. Document Image Understanding Technology, Apr. 2003.

[14] S. Marinai, E. Marino and G. Soda, "Layout Based Document Image Retrieval by Means of XY Tree Reduction," Proc. Eighth Int'l Conf. Document Analysis and Recognition, pp. 432-436, Aug. 2005.

[15] Jean-Luc Meunier " Optimized XY-Cut for Determining a Page Reading Order", Proceedings Eighth International Conference on Document Analysis and Recognition, 2005.

[16] Faisal Shafait, Daniel Keysers and Thomas M. Breuel "Performance Comparison of Six Algorithms for Page Segmentation" 7th IAPR Workshop on Document Analysis Systems, DAS'06. , Feb. 2006.

[17] S. Mandal, S. Chowdhury, A. Das and B. Chanda, "A Simple and Effective Table Detection system from Document Images," Int'l J. Document Analysis and Recognition, vol. 8, nos. 2-3, pp. 172-182, June 2006.

[18] F. Shafait, D. Keyser and T.M. Breuel, "Pixel-Accurate Representation and Evaluation of Page Segmentation in Document Images," Proc. 18th Int'lConf. Pattern Recognition, pp. 872-875, Aug. 2006.

## CITE AN ARTICLE

Jasrotia, D. K., & Malik, A. (2018). A DETAILED STUDY AND ANALYSIS OF OCR USING MATLAB. *INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY*, *7*(8), 223-228.