*IOSR Journal of Computer Engineering (IOSR-JCE)*
*e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 19, Issue 3, Ver. II (May.-June. 2017), PP 91-96*
*www.iosrjournals.org*

# A Survey on Character Recognition using Geometry Based Feature Extraction

Bishakha sharma[1], Arun Agarwal[2], Aruna Bajpai[3]
[1]*Research scholar, Deptt. Of Computer Science & Engg. ITM Group Of Institutions,Gwalior(M.P.)*
[2, 3]*Assistant Professor, Deptt. Of Computer Science & Engg. ITM Group Of Institutions,Gwalior(M.P.)*

***Abstract:*** *The framework through which PC can perceive content went into the framework either by means of an information gadget, for example, an attractive peruser or through sweep records is known as character recognition system. On the premise of this whole wide character recognition system can be part into two sections. The classification is described in the later section of this paper. This paper gives an idea as to what is a character recognition system, its application areas, the various stages explained in the process. Character Recognition is a wide thought and this paper focuses particularly on the Online Character Recognition System or the Optical Character Recognition (OCR) and clarifies the diverse periods of the same. The paper is organized in the following way. We start by giving an introduction to the character recognition system, followed by its classification. We have also explained the various stages in a character recognition system along with a few algorithms that are explained in the particular stages. The last section of the paper focuses on certain achievements and advancements in this particular field and future scope to the study of character recognition system.*

***Keywords:*** *character Recognition, OCR, Image Processing, Segmentation, Feature Extraction*

## I. Introduction

Character Recognition is a procedure that can convert text, present in digital image, to editable text. Using character recognition machine can distinguish the characters throughout optical mechanisms. The result of the CR should preferably be same as entered data in format. In character recognition we can speak to the document and after that achieve an imperative data about printed content. That knowledge or data can be used to recognize characters. Character Recognition is becoming an important part of modern research based computer applications. Particularly with the beginning of Unicode and sustain of difficult scripts on personal computers, the importance of this application has increased [1].

Character recognition system helps people effectiveness and diminishes their occupations of physically dealing with and preparing of archives. To perceive the individual character and afterward change over checked record into machine encoded shape programmed preparing is required field. Character recognition is basically of two sorts on the web and offline [2].

## II. Types Of Character Recognition

### A. Offline Character Recognition

In offline character recognition all printed or type-written characters are classified in offline mode. Offline character recognition can perceive the characters in a content that have been examined from a surface, for example, a sheet of report and are put away carefully in dim scale sort out. The storage of scanned documents have to be huge in size and many processing applications as searching for a content, editing, protection are either hard or impossible. These types of documents need human beings to process them manually. Character recognition system unravel such examined pictures of printed archives into machine encoded content.These interpreted encoded content can be effortlessly altered, looked and these content can be handled in numerous different courses as indicated by necessities. It also requires tinny size for storage in contrast to scanned documents.

### B. Online Character Recognition

The characters can be recognized by using online mode of recognition. In this the character is captured and stored in digital form through different means. Usually, a special pen is used in conjunction with an electronic plane. In this method as the pen moves over the surface, the progressive purposes of two-dimensional directions can be spoken to as a component of time and are put away all together. Now a days, due to improved use of handheld devices online handwritten recognition concerned knowledge of worldwide researchers. The change of client and PC correspondence has turned into an awesome potential for online recognition. With a specific end goal to distinguish and revise misrecognized characters on the detect the client can checking the acknowledgment comes about as they show up. The user is confident to modify his writing style so as to

improve recognition accuracy. Also, a machine can be accomplished to a particular user's style. Tests of his misrecognized characters are put away to help character recognition.Accordingly both essayist adjustment and machine adjustment is conceivable [2].

## III. Application

The last years have seen a extensive appearance of commercial optical character recognition products meeting the requirements of different users. In this section we can disscuss a portion of the distinctive territories of utilization for CR. Three fundamental application ranges are ordinarily portrayed; data entry, text entry and process automation.

### A. Data Entry:-

This application area describes technologies for entering huge amounts of limited data. Initially such document reading machines were used for banking applications. The frameworks are portrayed by perusing just a to a great degree constrained arrangement of printed characters, normally numerals and a couple of extraordinary images.. They are designed to read data like account numbers, customers identification, article numbers, amounts of money etc. The paper formats are constrained with a limited number of fixed lines to read per document. Because of these restrictions, readers of this kind may have a very high throughput of up to 150.000 documents per hour. Single character error and reject rates are 0.0001% and 0.01% respectively. Also, as a result of the limited character set, these readers are usually extremely wide to bad printing quality. These systems are specially designed for their applications and prices are therefore high.

### B. Text Entry:-

The second branch of reading machines is that of page readers for text entry, mainly used in office automation. As indicated by limitations concerning text style and printing quality the confinements on paper arrangement and character set are traded. The perusing machines are utilized to enter a lot of content, regularly in a word handling condition. These page readers are in strong contest with electronic exchange of data and direct key-input. This area of application is as the outcome of diminishing importance. As the character set read by these machines is relatively huge, the performance is particularly dependent on the quality of the printing. However, under controlled conditions the single character error and reject rates are about 0.01% and 0.1% respectively. The perusing speed is commonly in the request of a couple of hundred characters for each second.

### C. Process Automation:-

Within this area of application the main concern is not to read what is printed, but rather to control some particular process. This is actually the technology of automatic address reading for mail sorting. Hence, the goal is to direct each letter into the appropriate bin regardless of whether each character was correctly recognized or not. The general approach is to read all the information available and use the postcode as a redundancy check. On the premise of properties of mail the acknowledgment rate of these frameworks is unmistakably extremely needy. This rate then changes with the rate of written by hand mail. [3].

## IV. Character Recognition Stages

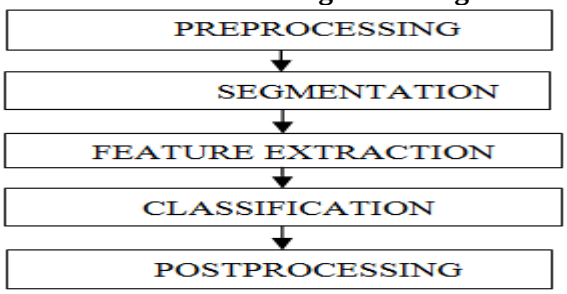


**Fig 1**: Stages of character recognition

### a. Preprocessing

The sequence of operations which performed on the scanned input image is known as preprocessing. It essentially enhances the image making it suitable for further processing. The various operations which are performed in pre-processing stage are noise removal, binarization, skew correction etc.

**A. Noise Removal**

It is a process of removing noise from scanned image by using appropriate filter for example smoothing linear filter, order statistic filter etc. Smoothing is the process of extracting large objects from the image by reducing noise and blurring , and deduction of small details.

**B. Binarization**

It change a dark scale picture into a double picture utilizing worldwide thresholding procedure like otsu's technique for thresholding. Otsu's provide optimum value of threshold.

**C. Skew Correction**

It is removal of skew in scanned document for its proper further segmentation. It is not necessary that handwritten documents are perfectly horizontally aligned, so skew correction methods are required to be performed. Projection profile examination, Hough changes, closest neighbor grouping, cross-connection, piece-wise covering by parallelogram and so on are the cases of skew correction [4].

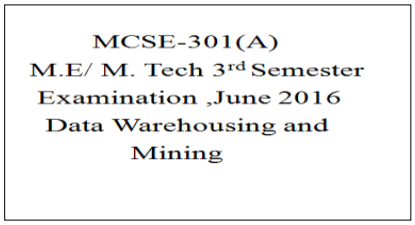


**Fig 2:** Text Template

**b. Segmentation**

In the segmentation stage, an image is decomposed into small parts of individual character. Segmentation includes: line segmentation which is separation of line from paragraph
Word segmentation which is separation of word from line.
Character segmentation which is separation of character from words[4].

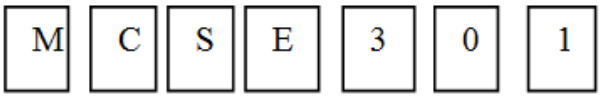


**Fig 3:** Segmented Image

**c. Feature Extraction**

In this stage, the processes that are necessary for classifying the character at recognition stage are extracted according to the features of the characters. Highlight extraction is an imperative stage and its effective methodology enhances the acknowledgment rate and diminishes the misclassification. Features like binary features, directional features etc are extracted and feature vector is created. Feature extraction methods falls among these categories.

**A. Statistical Features**

Statistical features is the part of feature extraction which is based on the probability theory and hypothesis. Statistical distribution of pixels of an image takes care of variations in writing styles. The statistical distribution of points is the process through which the statistical features are derived. Projections histogram, crossings, distances, zoning etc are comes under the category of statistical features.

**B. Structural Features**

Structural features give information about structure of the image. The geometrical and topological properties of character are clarified from auxiliary elements, for instance crossing focuses, Branches, circles, stroke length, stroke width, up, down, left and right projection profiles etc [4].

**d. Classification**

Classification is the final phase of character recognition system. It is the process to find the class labels of objects whose class label is unidentified. The component vector that we acquired from the element extraction stage can be utilized as a part of order stage. The features can be categorize according to its properties in

classification stage. Training and testing is done at the classification phase. The classifiers which can be used commonly are Artificial Neural Network, SVM, KNearest Neighbor and Nearest Neighbor classifier [5].

**e. Post Processing**

In this stage exactness of recognition is additionally expanded by associating lexicon to the framework keeping in mind the end goal to perform Syntax analysis, semantic analysis kind of higher level concepts, which is applied to check the recognized character[4].

## V. Character Recognition Approaches

There are different approaches used for the design of OCR systems are discussed below:

**A. Matrix Matching :-**

The technique through which each character can changes over into a case inside a structure, and after that differentiations the case and a document of recognized characters is known as matrix matching. The recognition of this stage is strongest on monotype and consistent single column pages.

**B. Fuzzy Logic :-**

The traditional assessments like yes/no, genuine/false, dark/white and so forth into which the middle of the road qualities are characterized is called fuzzy logic. In this approach an endeavor is made to viewpoint a more human-like method for sensible thinking in the programming of PCs. At the point when answers don't have a particular zero or one qualities and there are equivocalness included then fuzzy rationale have been utilized.

**C. Feature Extraction:-**

For the meaning of each character by the nearness or nonappearance of key components, including tallness, width, thickness, circles, lines, stems and other character qualities highlight extraction is utilized. Feature extraction is an established approach for OCR of magazines, laser print and fantastic pictures.

**D. Structural Analysis:-**

Auxiliary investigation approach gives an approach to dissect the character by looking at their sub highlight states of the picture, sub-vertical also, level histograms. Character repair limit is amazing for low quality substance and newsprints.

**E. Neural Networks :-**

This technique mirror the way the human neural framework works; it tests the pixels in each picture and matches them to a known record of character pixel designs. The ability to recognize the characters throughout abstraction is great for fixed documents and damaged text. For these sorts of issues, such as preparing securities exchange information or discovering patterns in graphical examples neural system end up plainly perfect. In all these approaches Neural Networks are efficient than others [6].

## VI. Literature Survey

Cui Xiaoxiao,et.al [7] Another strategy for computerized number acknowledgment for mechanical advanced meters in substation is clarified in this paper, which acknowledge straight SVM endless supply of Oriented Gradients (HOG) highlights.The grids of Histograms of Oriented Gradient descriptors considerably exceed for feature detection of the gray image which has more information than binary image. A original approach with segmentation of region of character image is proposed in this paper, which is important to the further HOG feature detection. SVM classifier is utilized as a part of the recognition parade and result demonstrates that HOG has better execution on digit arrangement in the substation examination robot instrument recognition.

Monika Lusa,et.al [8] Automatic traffic sign recognition by computers is becoming widely desirable in reality. Methods of automatic traffic sign detection are used in the automotive industry, not only in prototypes of automotive cars, but also in mass-produced models and mobile devices. In this paper, a two-phase algorithm based on key points feature detectors to detect and recognize road signs will be presented. The first stage of the algorithm locates objects present in the scene and determines their shape based on geometric properties. In order to reduce the number of found objects first phase includes two additional steps to remove too large and too small objects, and to merge objects of the same shape found in a similar area of the scene into one object. The second phase involves proper comparison of identified object with road signs from the knowledge database based on detected keypoints.

Hojin Cho [9] This paper gives a novel scene content location calculation, Canny Text Detector, which takes advantage of the contrast between picture edge and content for viable content limitation with enhanced

review rate. As closely associated edge pixels construct the structural information of an object, we observe that consistent characters compose a meaningful word/sentence which can shared a parallel properties such as spatial location, size, color, and stroke width in spite of language. However, common scene text detection approaches have not fully utilized such similarity, but mostly rely on the characters classified with high confidence, can lead to a low review rate. With a specific end goal to rapidly and heartily confine an assortment of writings we can misuse a correlation. By the utilization of unique Canny edge indicator, our calculation makes utilization of twofold limit and hysteresis following to recognize writings of low certainty. As indicated by exploratory outcomes on open datasets we can show that our calculation beats the state-of the-art scene content identification techniques in wording of detection rate.

Monica patel,et.al [4] This paper describes a comprehensive review of Character Recognition in English language. The manually written character acknowledgment has been helpful in assortment of uses like Banking divisions, Health mind enterprises and numerous such associations where translated reports are overseen. Handwritten Character Recognition is the procedure of change of written by hand message into machine reasonable frame. For written by hand characters there are challenges arrive like it contrasts starting with one author then onto the next, notwithstanding when same individual composes same character there is distinction fit as a fiddle, size and position of character. For the lessening of the multifaceted nature of perceiving written by hand message unmistakable sorts of technique has been utilized as a part of the most recent research of this zone.

Sukhpreet singh [1] This paper gives different thoughts regarding on English OCR strategies. For the transformation of various distributed books of English into editable PC content documents english OCR framework is fundamental. Some new techniques has been created to defeat the unpredictability of English composition style in the most recent research of this territory. Still these algorithms have not been experienced for complete characters of Engish Alphabet. Hence, a system is required which can handle all classes of English text and identify characters among these classes.

Karishma Tyagi**,** et.al [10] The application of OCR has become important in day-to-day life. OCR has been widely used in banking, legal, health care, finance etc. The most intriguing and testing research regions in field of picture handling and example acknowledgment in the current years is penmanship acknowledgment. This paper gives different thoughts for changing over literary substance from a paper record into machine lucid frame. The computer actually recognizes the characters in the document during a transforming technique called Optical Character Recognition. A few procedures like OCR utilizing connection strategy and OCR utilizing neural systems has been talked about in this paper.

Shalin A. Chopra1,et.al [11] Presently a days, keyboarding remains the most well-known method for contributing information into PCs. This is presumably the most tedious and work serious operation. Optical Character Recognition is the machine delineation of human perusing and has turned into a serious research for over three decades. The procedure through which filtered pictures where pictures can be written by hand, typewritten or printed content can be depicted as mechanical or electronic trade is known as OCR. This is a strategy for digitizing printed messages with the goal that they can be consequently looked and can be utilized as a part of machine procedures. It is a procedure of changing over the pictures into machine-encoded content that can be utilized as a part of machine interpretation, content to-discourse and content mining. This paper gives a simple, efficient, and less exorbitant definition to build OCR for perusing any record that has settle text dimension and style or manually written style. In this paper OCR utilizes database for the accomplishment of conviction and less computational cost to perceive English characters which makes this OCR extremely easy to oversee.

## VII. Conclusion

A number of techniques that are used for character recognition have been discussed. The main research is currently going on in extending Character Recognition to English alphabet, numbers, etc. Template matching method which is easy to implement Due to algorithmic simplicity and higher degree of flexibility the template matching method is easy to implement. In this paper we have discussed as to what is character recognition. Also this paper describes various stages that are given in detail. There are various algorithms for each stage and also new algorithms are being formulated in recent times. This paper also gives a set of references which we have used and also so that the readers can get through and have a better understanding of the different algorithms portrays in the diverse stages.

## Refrences

[1]. Sukhpreet Singh, Optical Character Recognition Techniques: "A survey International Journal of Advanced Research in Computer Engineering & Technology" (IJARCET) Volume 2, Issue 6, June 2013 ISSN: 2278 – 1323.

[2]. Priya Sharma,Randhir Singh, "India Survey and Classification of Character Recognition System" International Journal of Engineering Trends and Technology- Volume4Issue3- 2013.

[3]. Line Eikvil, "Optical character recognition" Unpublished

[4]. Monica Patel1 , Shital P. Thakkar 2, " Handwritten Character Recognition in English": A Survey International Journal of Advanced Research in Computer and and Communication Engineering Vol. 4, Issue 2, February 2015.
[5]. Arya A1 , Anil Kumar A2, "A Review on Geometrical Analysis in Character Recognition" IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 17, Issue 2, Ver. V (Mar – Apr. 2015), PP 61-65.
[6]. N. VENKATA RAO, 2DR. A.S.C.S.SASTRY, 3A.S.N.CHAKRAVARTHY, 4 KALYANCHAKRAVARTHI P "OPTICAL CHARACTER RECOGNITION TECHNIQUE ALGORITHMS" © 2005 - 2015 JATIT & LLS. ISSN: 1992-8645 E-ISSN: 1817-3195.
[7]. Cui Xiaoxiao, Fang Hua, Yang Guoqing, Zhou Hao, "A New Method of Digital Number Recognition forSubstation Inspection Robot" 978-1-5090-3228-0/16/$31.00 ©2016 IEEE.
[8]. Monika Lusa, "Recognition of Multiple Traffic Signs using Keypoints Feature Detectors" 2016 international Conference and Exposition on Electrical and Power Engineering (EPE 2016), 20-22 October, Iasi, Romania.
[9]. Hojin Cho Myungchul Sung Bongjin Jun Stradvision, Inc. Canny Text Detector: "Fast and Robust Scene Text Localization Algorithm" 2016 IEEE Conference on Computer Vision and Pattern Recognition.
[10]. Karishma Tyagi, Vedant Rastogi Department of Computer Science & Engineering, IET Alwar, Rajsthan, INDIA Survey on Character Recognition using OCR Techniques International Journal of Engineering Research & Technology (IJERT) Vol. 3 Issue 2, February – 2014.
[11]. Shalin A. Chopra1, Amit A. Ghadge2, Onkar A. Padwal3, Karan S. Punjabi4, Prof. Gandhali S. Gurjar5, "Optical Character Recognition" International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 1, January 2014,ISSN (Online) : 2278-1021, ISSN (Print) : 2319-5940