# AN ITERATIVE GRAPH BASED HANDWRITTEN TEXT SEGMENTATION USING OPTICAL CHARACTER RECOGNITION

**Mrs.Y.Neeraja,**
Associate Professor, Department of ECE, Narayana Engineering College, Gudur, AP, 524101

**B.Hemalatha, K.Bhargavi, CH.Hema, B.Keerthi**
UG Student, Department of ECE, Narayana Engineering College, Gudur, AP, 524101

**Abstract:** The main objective of this paper presents a simple approach for the character segmentation of the handwritten words. Segmentation of handwritten text recognition is better to employ the dependency on context of strokes written before and after it. The novel part of this paper is to avoid over segmentation by thresholding automatically. The characters which are separated are then subjected to recognition using OCR. The Optical Character Recognition (OCR) offers several important applications. This paper presents a technique for recognition of Printed text with noise using Optical Character Recognition (OCR). The main steps of this system are pre-processing of the text and remove the noise from the text. Segmentation of the text image to each character. The feature extraction using coming-based technique and classification. The performance of the recognition of handwritten character is greatly depending on the segmentation rate. This paper achieves the segmentation rate 76.9%. The quality of the input documents is very important to achieve high accuracy and increase the recognition of text depends on the segmentation.

**Keywords** – Optical character recognition, Handwritten text, Features Extraction, Segmentation.

## 1.INTRODUCTION

Segmentation of handwritten pages into text lines and words. It is still a difficult problem as the writing is not constrained and the writing varies a lot depends on the writer. And yet, it is important step for the recognition of the content: the segmentation into words can significantly help the process of the recognition. In this domain, which are able to deal with various kinds of unconstrained writing styles. In order to segment text from a given input document image, it is necessary to detect all the possible text regions. In the case of overlapped scripts, segmentation is a relatively simple task. In the case of overlapped scripts, broken characters, connected characters, loosely configured characters, and mixed scripts, segmentation is difficult. Overlapped, broken, connected and loosely configured characters are major causes of segmentation errors. Segmentation of Text Image is used to locate each individual character and its boundaries. It involves process of labelling, which assigns the some label to spatially align units. Handwritten character recognition (HCR) is the process of conversion of scanned handwritten documents into the text document so that it becomes editable and researchable. It contains conversion of text image into letter codes which are useful in computer and text processing applications.

All collecting databases from all different handwritten style samples is considered as static images of handwriting. So that the recognition is slightly problematic as different individuals have a dissimilar style of writing. Sometimes a document written in the past used for recognition, then scanned image will need to be extracted for getting individual characters. Various tools exist for extraction, but may occur common imperfection in it. Commonly, when characters are connected, then both characters considered as single sub-image. Therefore there is a recognition problem but Yet there are many algorithms are available to minimize the problem related to connected characters. After the extraction of individual characters occurs, a recognition system is used to find equivalent computer character. The handwritten character recognition is globally divided into two types: 1. Offline, 2. Online character recognition.

Off-Line: The writing is typically taken optically by a scanner and the completed writng is accessible as an image.

On- Line: It has the two dimensional organizes of sequential points are denoted as a function of time and the order of knocks prepare by the writers are also accessible.

In the research area, handwritten character recognition system is exploring with new techniques and improving performance accuracy.

## 2.LITERATURE SURVEY

In this section, detailed literature review is done that aims to review the critical points of current works. Here the information collected about researches and innovations carried out on the related technologies have been done. This section will highlight the recent trends and innovations in the concerned technology.

In 1993 C.Y. Suen [1] obtained Pattern Recognition Letters, followed by R. Legault and C. Nadal. Based on a review of the recent achievements in off-line computer recognition of totally unconstrained handwritten characters, and extensive research, the authors attempt to identify new frontiers for research which may lead to further breakthroughs in this field. They will present some evidences and novel ideas on ways of stretching the limits of handwriting recognition systems aiming at outperforming human beings.

In this review focused on the shape analysis of binarized images, quite often assuming good quality document and isolated characters. Such assumptions are challenged by the conditions met in practice : binarization is difficult for low contrast documents, characters often touch each other, not only on the sides but also between lines, etc.

In 1995 H. I. Avi-Itzhak[2] obtained Pattern Analysis and Machine Intelligence, followed by T. A. Diep and H. Garland. Optical Character Recognition (OCR) refers to a process whereby printed documents are transformed into ASCII files for the purpose of compact storage, editing, fast retrieval, and other file manipulations through the use of a computer. The recognition stage of an OCR process is made difficult by added noise, image distortion, and the various character typefaces, sizes, and fonts that a document may have. In this study a neural network approach is introduced to perform high accuracy recognition on multi-size and multi-font characters; a novel centroid-dithering training process with a low noise-sensitivity normalization procedure is used to achieve high accuracy results.

The study consists of two parts. The first part focuses on single size and single font characters, and a two-layered neural network is trained to recognize the full set of 94 ASCII character images in 12-pt Courier font. The second part trades accuracy for additional font and size capability, and a larger two-layered neural network is trained to recognize the full set of 94 ASCII character images for all point sizes from 8 to 32 and for 12 commonly used fonts. The performance of these two networks is evaluated based on a database of more than one million character images from the testing data set.

In 1993 S.N.[7] Srihari obtained Pattern Recognition Letters, postal address interpretation is the task of assigning to letter mail pieces a delivery point encoding. The encoding is determined from images of destination address on mail piece faces; addresses that are handwritten, are of poor-quality machine printing, are incomplete or incorrect. This paper describes several recognition algorithms used in the interpretation of handwritten and machine-printed address text.

Based on a review of the recent achievements in off-line computer recognition of totally unconstrained handwritten characters, and extensive research, the authors attempt to identify new frontiers for research which may lead to further breakthroughs in this field. They will present some evidences and novel ideas on ways of stretching the limits of handwriting recognition systems aiming at outperforming human beings.

In 1996 R.G.Casey[11] obtained Pattern Analysis and Machine Intelligence, Character segmentation has long been a critical area of the OCR process. The higher recognition rates for isolated characters vs those obtained for words and connected character strings well illustrate this fact. A good part of recent progress in reading unconstrained printed and written text may be ascribed to more insightful handling of segmentation. This paper provides a review of these advances. The aim is to provide an appreciation for the range of techniques that have been developed, rather than to simply list sources. Segmentation methods are listed under four main headings. What may be termed the "classical" approach consists of methods that partition the input image into sub images, which are then

classified. The operation of attempting to decompose the image into classifiable units is called "dissection". The second class of methods avoids dissection, and segments the image either explicitly, by classification of prespecified windows, or implicitly by classification of subsets of spatial futures collected from the image as a whole. The third strategy is a hybrid of the first two, employing dissection together with recombination rules to define potential segment, but using classification to select from the range of admissible segmentation possibilities offered by these sub images. Finally, holistic approaches that avoid segmentation by recognizing entire character strings as units are described.

In 2005 C. K. Cheng[21] obtained Documentation Analysis and Recognition, followed by M. Blumenstein. In this view enhanced heuristic segmenter (EHS) and improved neural-based segmentation technique for segmenting cursive words and validating prospective segmentation points respectively. The EHS employs two new features, ligature detection and a neural assistant, to locate prospective segmentation points. The improved neural-based segmentation technique can then be used to examine the prospective segmentation points by fusion of confidence values obtained from left and centre character recognition outputs in addition to the segmentation point validation (SPV) output.

The improved neural-based segmentation technique uses a recently proposed feature extraction technique for representing the segmentation points and characters to enhance the overall segmentation process. The EHS and the neural-based segmentation technique have been implemented and tested on a benchmark database providing encouraging results. Areas with low pixel density are then identified as prospective segmentation points. For example of vertical histogram for the word image before thinning; the vertical histogram is formed basaed on the middle region of the word after removal the dots. Disadvantage using the original image before thinning, excessive number of "low" density regions appears like a noise, which means excessive number of anomalous points still in the word image.

One weakness of the modified vertical histogram is that it is not suitable for characters with overlapped strokes. But in this research, since the overlapped strokes are removed in most cases, the advantage of the modified vertical histogram can then be maximized. After filtering and thinning the word image, only one or two vertical pixels are defined as prospective segmentation point.

In 1993 S.N. Srihari obtained Pattern Recognition Letters, postal address interpretation is the task of assigning to letter mail pieces a delivery point encoding. The encoding is determined from images of destination address on mail piece faces; addresses that are handwritten, are of poor-quality machine printing, are incomplete or incorrect. This paper describes several recognition algorithms used in the interpretation of handwritten and machine-printed address text.

Based on a review of the recent achievements in off-line computer recognition of totally unconstrained handwritten characters, and extensive research, the authors attempt to identify new frontiers for research which may lead to further breakthroughs in this field. They will present some evidences and novel ideas on ways of stretching the limits of handwriting recognition systems aiming at outperforming human beings.

In 1997 S.B. Cho obtained Neural Networks, Artificial neural networks have been recognized as a powerful for pattern classification problems, but a number of researchers have also suggested that straight forward neural-network approaches to pattern recognition are largely inadequate for difficult problems such as handwritten numeral recognition. In this paper, we present three sophisticated neural-network classifiers to solve complex pattern recognition problems: multiple multilayer perceptron(MLP) classifier, hidden Markov model(HMM)/MLP hybrid classifier, and structure-adaptive self-organizing map (SOM) classifier.

In order to verify the superiority of the proposed classifiers, experiments were performed with the unconstrained handwritten numeral database of Concordia University, Montreal, Canada. The three methods have produced 97.35%, 96.55% and 96.05% of the recognition rates, respectively, which are better than those of several previous methods reported in the literature the same database.

In this paper, we have presented three sophisticated neural-network classifiers to recognize the totally unconstrained hand-written numerals: multiple MLP classifier, HMM/MLP hybrid classifier and structure-adaptive SOM classifier. All of them have produced better results than several previous methods reported in the literature on the same database. Actually, the proposed methods have a small, but statistically significant.

The multiple MLP classifier leads to a reliable recognizer without great effort to fine-tune the individual MLP classifiers. Also, the HMM/MLP hybrid classifier complements each method for improving the overall performance, and the structure-adaptive SOM classifier automatically finds a network structure and size suitable for the classification of complex patterns through the ability of structure adaptation.

Even though our work to date is concentrated on handwritten numeral recognition, we believe that the methods presented can be easily generalized to more difficult problems, such as handwritten Roman character recognition and Hangul recognition. The further works are under going with the more difficult task of recognizing handwritten Hangul.

## 3.PROPOSED METHOD

Optical character recognition (OCR) is the use of technology to distinguish printed or handwritten text characters inside digital images of physical documents, such as a scanned paper document. The basic process of OCR involves examining the text of a document and translating the characters into code that can be used for data processing. OCR is sometimes also referred to as text recognition.

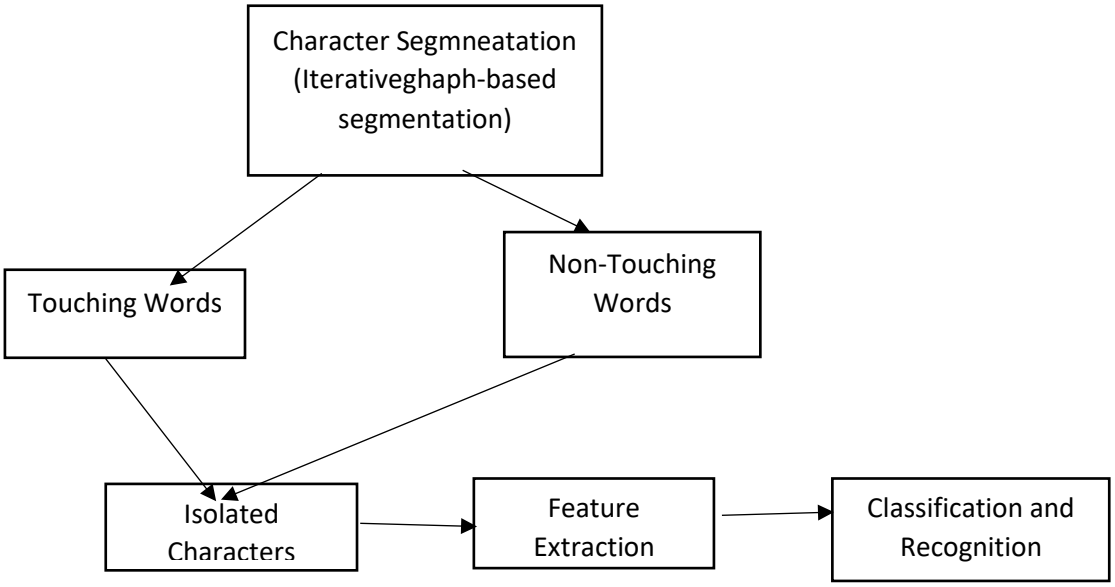


**Fig 1: Block diagram**

A. Image acquisition: This involves scanning a document and storing it as an image. Their solution (number of dots per inch, dpi) determines the rate of process.

B. Pre-processing: Process of representing the scanned image of further processing. Pre -processing aims to produce data that are easy for the OCR system to operate accurately

It reduces noise and distortion, removes skewness and performs skeletonising of the image, thereby simplifying the processing the rest of the stages.

C. Segmentation: After the pre- processing stage, a 'clean' document is obtained. The next stage is segmentation. In this stage, segmenting the document into its sub-components. It separates the different logical parts, like text from graphics, line of a paragraph, and characters of a word. Segmentation is an important phase of OCR, because it can reach in separation of words, lines or characters directly affect the recognition rate of the script [4]. In fact correct recognition based on correct segmentation.

A character is the smallest unit of any language script and the segmentation of characters is the most crycial step for any OCR (Optical Character Recognition) System. This system translates scanned or printed image of the document into a text document that can be edited. The selection of segmentation algorithm being used is the key factor in deciding the accuracy of the OCR system. If there is a good segmentation of characters, the recognition accuracy will also be high. Segmentation of

words into characters becomes very difficult due to the cursive and unconstrained nature of the handwritten script.

**3.1Character Segmentation Technique**

The proposed technique segments the words in an iterative manner by focusing on presence of headline,aspect ratio of characters and vertical and horizontal projection profiles. The proposed approach of segmentation can be used for handwritten text for English.Graph based methods are an effective tools for segmentation since they model the impact of pixel neighborhoods on a given cluster of pixels or pixel. In this methods, the image is modelled as weighted, undirected graph. Usually a pixel on a group of pixels are associated with nodes and edge weights define the similarity between the neighbourhood pixels. The graph is then partitioned according to a criterion designed to model. Each partition of the nodes output from these algorithms are considered an object segment in the image.It is used for segment the foreground and text in the given image.Our unique approach for learning English graph segmentation rules using the iterated Version space Algorithm is presented. After defining the problem and our representation foer the instances.

Graph based methods are an effective tools for segmentation since they model the impact of pixel neighborhoods on a given cluster of pixels or pixel. In this methods, the image is modelled as weighted, undirected graph. Usually a pixel on a group of pixels are associated with nodes and edge weights define the similarity between the neighbourhood pixels. The graph is then partitioned according to a criterion designed to model. Each partition of the nodes output from these algorithms are considered an object segment in the image.It is used for segment the foreground and text in the given image.

D. Touching Words and Non-Touching Words: Touching words are recognized by using pixel wise approach and non touching words are recognitized by using bounding box approach.

E. Feature Extraction: A set of rules stored on OCR engine comparing against character's shape and its features that distinguishes each character identify a character. The main part of the recognition system design is the selection of a stable representative set of features. It is the most consequential issue in the designing issues involved in building an OCR system.

F. Classification and Recognition: The main decision making stage of an OCR system is classification. Classification uses the features extracted in the feature extraction stage to identify the text segment.

## 6. RESULTS AND DISCUSSION

The performance of the recognition of handwritten character is greatly depending on the segmentation rate. Experiments have been performed to test the proposed system. It can be seen that the K-Nearest Neighbour algorithm can be used to classify images into alphabets in an OCR.

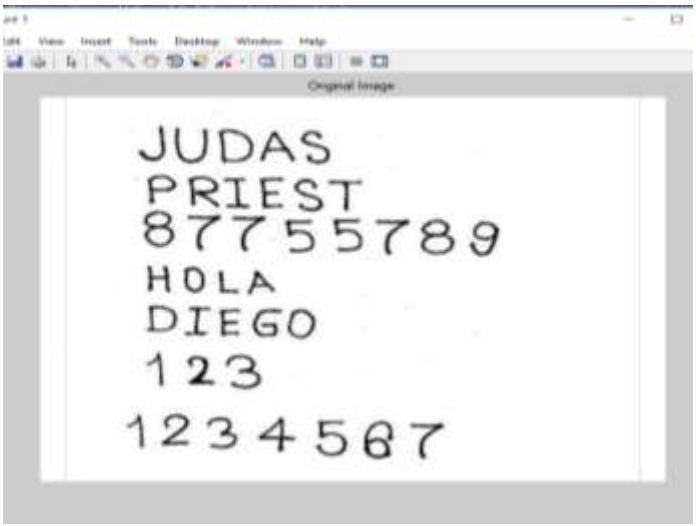


**Fig 2: Original Image**

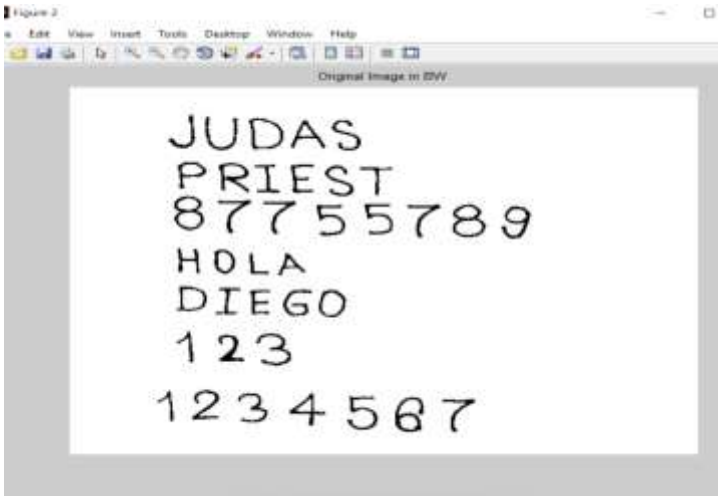


**Fig3 : Binarized Image**

**Fig 4: Character segmentation**

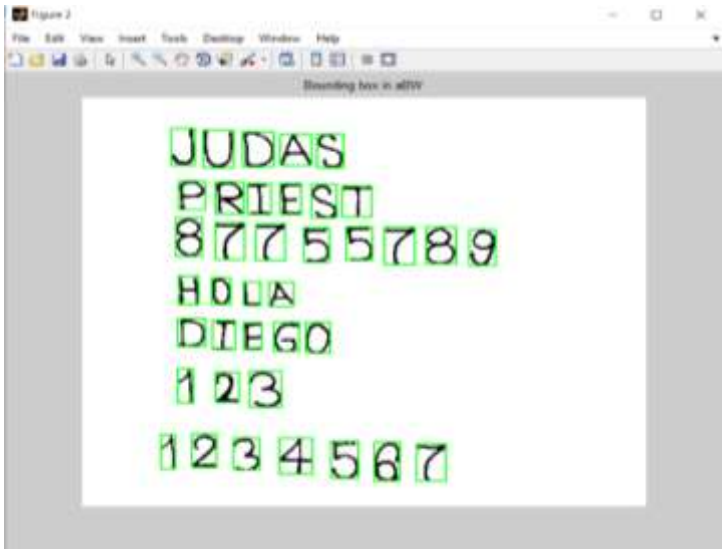


**Fig 5: Bounding Box**

## 7. CONCLUSION

In this paper, an effective character recognition system using OCR has been proposed. The proposed system is based on image processing, characters segmentation, feature extraction and classification process. English numbers and characters database were used. K-nearest Neighbour

Algorithm is used for classification process. The proposed character recognition system performance was evaluated and high recognition rate was achieved. i.e.76.9%.

## 8.REFERENCES


1. Frucci, Maria, Sanniti di Baja, Gabriella(2008) "From Segmentation to Binarization of Gray-level Images".
2. R. Rasika, Mr.D.D.Dighe, "Handwritten English Character Recognition Using KNN".
3. J.Leena Hepzi, I. Muthumai, S. Selvabhatathi,"English Cursive Hand Written Character Recogniton".
4. Amandeep Kaur, Seema Baghla, Sunil Kumar," Study of Various Character Segmentation Techniques For Handwritten Off-Line Cursive Words: A Review".
5. Veronica Ong, Derwin,Suhartono, "Using K-Nearest Neighbour In Optical Character Recogntion".
6. Vassilis Katsouros, Vassilis Papavassiliou, " Segmentation of Handwritten Document Images Into Text Lines".
7. M. Saravanan, Siddhartha Roy, " Handwritten Character Recognition Uisng K-NN Classification Algorithm".
8. Nawaf Hazim Barnouti, Mohammed Abomaali, Mohannad Hazim Nsaif AI-Mayyahi," An Efficient Character Recognition Technique Using K-Nearest Neighbour Classifier".
9. Rakesh Rathi, Ravi Krishan Pandey," Offline Handwritten Devanagari Vowels Recognition Using KNN Classifier".
10. A. Sinchez, P.D. Suarez, C.A.B.Mello,A.L.I.Oliveira,V.M.O.Alves,"Text Line Segmentation In Images of Handwritten Historical Documents".
11. A. Phukan, M. Borah, "A survey paper on character recognition focusing on offline character recognition," International Journal of Computer Engineering and Applications, vol. 6, pp. 51-60, 2014.
12. A. Choudhary, "A review of various character segmentation techniques for cursive handwritten words recognition," International journal of Information and Computation Technology, vol. 4, pp. 559-564, 2014.
13. A. Choudhary, R. Rishi, and S. Ahlawat, "A new approach to detect and extract characters from off-line printed images and text," Information Technology and Quantitative Management, pp. 434-440, 2013.
14. A. Choudhary, R. Rishi, and S. Ahlawat, "A new character segmentation approach for off-line cursive handwritten words," Information Technology and Quantitative Management, pp. 88-95, 2013.
15. Cover, T., Hart, P., Nearest-neighbour pattern classification, Information Theory, IEEE Transactions on, Jan. 1967, pp. 21-27.
16. R. M. Bozinovic and S.N. Srihari, Off-Line Cursive Script Recognition, IEEE Trans on Pattern Analysis and Machine Intelligence, vol 11, no.1, page 68, 1989.
17. R.G. Casey, Text OCR by solving a cryptogram, Proc. 8th Int. Conf. on Pattern Recognition, Paris,pp. 349-351, Oct. 1986.
18. R.G. Casey, Segmentation of touching characters in postal addresses, Proc. 5th US Postal ServiceTechnology Conference.
19. M. Cesar and R. Shinghal, Algorithm for segmenting handwritten postal codes, Int. J. Man MachStud., vol. 33, no. 1, pp. 63-80, Jul. 1990.
20. M.Y. Chen and A. Kundu, An Alternative to Variable Duration HMM in Handwritten Word Recognition, Pre-Proceedings IWFHR III, Buffalo, page 82, May 1993.
21. P. D. Gader, M. Mohamed and J. H. Chiang, "Handwritten word recognition with character and inter-character neural networks", IEEE Transactions on Systems, Man and Cybernet—Part B: Cybernetics, vol.27, pp.158-164, 1997.
22. Kang L, David D ,"Local segmentation of touching character using contour based shape decomposition"