An Effective Hybrid Segmentation Approach For Better Recognition Of Cursive Handwritten Characters.

Ashwini Kolekar, Pooja Kumbharkar, Shrineeta Kurud, Yashshree Shinde, Shalaka Deore

Computer Department, Modern Education Society College Of Engineering Pune,
Email:- pooja25kumbharkar@gmail.com shrineeta.kurud@gmail.com

ABSTRACT:

Segmentation of cursive handwriting is the challenging step of Optical Character Recognition (OCR).The recognition accuracy will highly depend on the good segmentation. Segmentation of cursive handwriting is very difficult. The segmentation can be done on the basis of zoning, a line segment of text, a word segment from line and character segment from word.This can be done by the use of horizontal, vertical methods. By reviewing many basic and advanced techniques of handwritten word segmentation, we try to find the most optimal segmentation bred out of them.The importance of segmentation is to find the accurate new segmented words in order to make a new data set. There are many segmentation algorithms available but each algorithm has accuracy. The Purpose of this project is to combine different structural and stochastic segmentation algorithms.

The proposed system will be segmenting cursive handwritten English text. For this free available dataset from mnist database is used. By comparing the accuracy of other segmentation techniques we have tried to implement edge and region segmentation from structural segmentation & CNN from stochastic segmentation as hybrid segmentation.

Keywords—: Machine Learning, Image Segmentation, Structural Segmentation, Stochastic Segmentation, CNN.

I. INTRODUCTION:-

Optical character recognition could be a program that translates a scanned image of a document into a text document which will be embedded. Segmentation of cursive handwriting is incredibly troublesome. Character segmentation is an operation of extracting an individual character from the text or as to decompose a picture into the sub-image of individual symbols. There are primarily 3 phases of a personality recognition system, particularly preprocessing, segmentation and recognition.

Preprocessing aims to produce information that is simple for the OCR system to figure accurately. Preprocessing consists of different activities, where it reduces noise and distortion, removes skewness and performs skeleton sing of the image, thereby simplifying the process for the next stages. Subsequent stage is segmenting the document into its sub-components. It separates the various logical components, like text from graphics, line of a paragraph, and character of a word. Preprocessing helps the next stage to enable proper segmentation which would further be used for the recognition.

Segmentation is the process which focuses on dividing an object into different parts according to the object of our interest or as per the features and their properties. Segmentation is the important step in cursive script recognition where it detects lines from the image, then detects words from each line and then extracting individual characters from each word.Segmentation has a wide range of applications such as image based searching, number plate identification , face recognition and many more. So, Segmentation is a very important step for further analysis. The higher recognition rate is achieved when the characters are correctly isolated from the input document or image. In this we need to improve the accuracy and precision of segmentation technique and

compare the results with some standard benchmark dataset.

There are different types of segmentation techniques such as Edge Segmentation,Threshold based Segmentation ,CNN based segmentation, which are under Structural Segmentation and Stochastic Segmentation algorithm.In Edge based Segmentation, different edge filters are applied on the images where the pixels are classified into edges and non-edges of an image depending on the output of the filter.Threshold based Segmentation is segmentation technique where it recursively finds the neighboring pixel based on the similar seed pixel based on the similarity measures.CNN based segmentation is an Artificial Neural Network based Segmentation algorithm where it uses discrete pixel values of an image for Segmentation.

Every single Segmentation algorithm has some drawbacks which result in low accuracy ,there is no such algorithm which would work perfectly for all types of images considering all factors such as noise,speed,accuracy,complexity. So the best solution to reduce the drawbacks and increase the accuracy is to make it a hybrid . Hybrid Segmentation combines the different segmentation algorithm from the Structural Segmentation and different Stochastic Segmentation algorithm.

Using just one Segmentation technique does not give us 100 % of accuracy so our approach is to combine the segmentation technique that is edge segmentation and region segmentation from structural Segmentation and CNN and machine learning classification from stochastic segmentation so as to create hybrid segmentation technique on handwritten cursive dataset which would reduce the drawbacks of the individual segmentation and will result in the better accuracy than the individual segmentation.

II. LITERATURE REVIEW

There is a lot of work done on handwritten documents whereas a very less amount of work is done on cursive handwritten documents.
Segmentation is the crucial task for classification and recognition of cursive handwritten documents.

Xiao and Leedham used the knowledge of English characters for segmentation. Bretto et al. presented a segmentation algorithm using Hypergraph theory. They have used Hypergraph theory for developing image-processing applications such as segmentation. However, the segmentation they have dealt with is for picture images and not for cursive handwritten words. Cheng et al. improved the segmentation process to detect ligatures by the analysis of direction features along with neural validation but made the whole

process computationally expensive by intruding lot of features.Hence cannot be used in real time scenario.

Most common challenges faced during segmentation of cursive scripts:-
There can be variation in shapes and writing styles of different writers . Cursive nature of handwriting i.e. two or more characters in a word can be written connected to each other. Characters can have more than one shape according to their position inside the word image. Words may be written by a pen having ink of different colors.
a.Over Segmentation
b.Bad Segmentation
c.Missed Segmentation

Through the study of previous work related to segmentation it is know that:-
1.Some have used the character shape analysis,ligatures analysis using the one from these four segmentation approaches or a combination of these two;-
i) projection-based, ii) smearing-based, iii) Hough transform-based, and iv) grouping methods
1.Some have used Connected Components Chains along with any other projection or segmentation technique.
2.Implicit segmentation technique achieved segmentation and recognition at the same time.
3.Implicit segmentation based recognition removed the class overlapping problem. The explicit segmentation based approach was computationally complex than the implicit segmentation approach, but the explicit segmentation technique achieved better results than implicit segmentation technique.
4..Some paper have used structural techniques like edge/region or any other segmentation technique along with the neural networks for validation so that whichever points have not been covered by the previous algorithm gets detected and then the proper action can be taken.

In literature, most of the researchers integrated segmentation approaches with some intelligent techniques such as neural networks, support vector machines and so on to enhance accuracy. For the sake of space, a brief comparison of achievement for segmentation rate is presented. Verma and Gader obtained 76.52% segmentation rate using neuro-feature based approach on words taken from CEDAR; however the number of word was not mentioned. Likewise, Blumenstein and Verma claimed 78.85% segmentation accuracy without mentioning number of words taken from CEDAR. In the same way, Verma claimed 84.87% segmentation accuracy for 300 CEDAR words. Similarly, Cheng et al.acquired 95.27% segmentation rate from 317 CEDAR words. Finally, Cheng and Blumenstein employed three trained neural networks for fine character segmentation. Accordingly, confidence obtained from each network fused to decide final

segmentation points. Beside training issues and computational complexity that was raised significantly, trained network could not perform desired objective . Recently, Lee and Verma [21] reported 83.46% segmentation accuracy on 200 words of CEDAR.

It was observed that neural networks are mainly used for validation not for actual segmentation purpose.

It is found that there is no perfect method for image segmentation because image segmentation depends on many factors, i.e., pixel color, texture, intensity, similarity of images, image content, and problem domain. Therefore, it is not possible to consider a single method for all type of images nor all methods can perform well for a particular type of image. Hence, it is good to use hybrid solution consists of multiple methods for image segmentation problem. Table1 we have rated the various segmentation algorithms against the parameters such as speed, Complexity, Noise resistance, Accuracy & Automation.
The below table shows our observations from previous study.

Parameter	Edge based	Threshold based based	CNN based
Speed	1	4	6
Computational Complexicity	1	3	6
Noise Resistance	2	3	1
Accuracy	3	2	1
Automation	5	3	1

Table1:-Comparison table(1 stands for best and 6 for worst)

.

Taking Table1 into consideration that is the advantages and disadvantages of every segmentation technique.
So that we can improve overall segmentation accuracy/efficiency.

Table1 shows the comparison that the speed and computation is faster of edge and region based but noise and accuracy is the issue with both the techniques whereas CNN fills this gap i.e the drawback of ege and region based segmentation approach.As the main aim of our is to increase segmentation accuracy for better recognition CNN is helping us with it.

By fusion or by taking different techniques (making hybrid)together can definitely increase our performance of accuracy.

As Edge detection cannot find exact boundaries of the image further CNN can be fed with the segmented images which would help in detecting and correcting over segmentation and also will help is marking exact boundaries of the characters within the image.
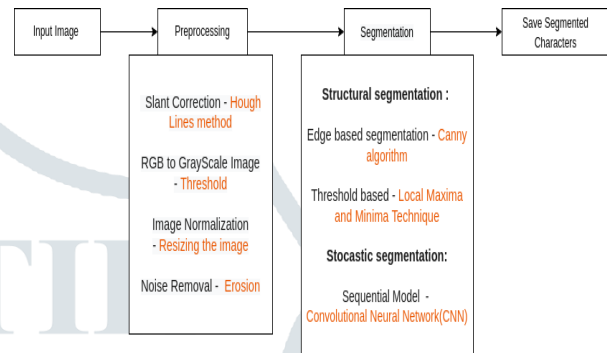
III. PROPOSED SYSTEM

Fig1. Proposed Model Architecture

Fig 1 shows the architecture of proposed system. As segmentation plays very important role in classification and recognition, its accuracy contributes to the great extend in classification. Proposed system will be segmenting the cursive text into individual characters with high accuracy with the help of Hybrid Segmentation. Datasets of cursive handwritten text from different writers has been collected from mnist database which is freely available.

- We are using Hybrid segmentation approach.
- Proposed system is that to combine different algorithms to efficiently segment the handwritten characters which in turn leads to better recognition.
- The main focus is upon segmenting the characters as segmentation is the building block of the accuracy of the model.
- System would use combination of structured and stochastic segmentation techniques.
- As the character being joint in nature segmenting them is the most difficult task to accomplish which we are trying to solve through this system.

IV. PREPROCESSING

Image Preprocessing

The first step to any document analysis is pre-processing. First the sample image is converted to gray scale image. A binary image is generated out of this grey scale image and one bit binary noise is removed by using erosion. The subsequent steps are all for normalizing the data to common axis. In normalization the various sized images are resized to fixed size image keeping aspect ratio intact.

Image Preprocessing plays very important role in segmentation or even in classification and recognition. Clean and noise free image makes it easier to compute any operation on it. Therefore before Segmenting the characters from the image it will undergo through preprocessing techniques. Firstly, we use median filter for reducing impulsive and salt or pepper noise and also used for reducing random noise. After median filter we get clean and noise free image. Secondly, we are going to apply binarization technique using OTSU's algorithm. Otsu method helps in finding optimal threshold value of an input image by going through all possible values(0-255). This helps in converting image pixels into bi-level collection of pixels ie. into black and white pixel. In this way we get an enhanced image.

- **Slant Correction**
  To make the letter recognition more efficient we corrected the slant angle of image. For slant correction we first calculate the angle for maximum lines with respect to vertical line using 'Hough Lines method'. After finding the angle, image is rotated with same angle in opposite direction around the baseline of text.

- **Line Segmentation**
  Considering the text lines are horizontal in position, the text lines are separated out by extracting the row wise consecutive black pixel existence.

- **Word Segmentation**
  After extracting the text lines, the image is dilated to join the distant neighboring pixel or to remove the small space between two letters. After that we tried to use contour tracing to accurately segment the intersecting components. Then bounding box is applied on contours to extract words. These word images are then further sent for preprocessing.

- **Top Contour Point Extraction**
  The top contour points are extracted by taking first black pixel of each column from top.

Segmentation

Character segmentation

- For the purpose of increasing the accuracy we have used a combination of pure segmentation based technique with a recognition based segmentation technique.
- Hybrid = Pure segmentation + Recognition based segmentation

- Pure segmentation:-
  1. Edge detection based segmentation
  2. Threshold based segmentation

- Recognition based segmentation
  1. CNN based sequential model

HYBRID SEGMENTATION

Edge based segmentation and Threshold based segmentation

Local maxima & Minima Technique In this we traverse through each top contour points from left to right. Then the vertical distance of each contour point with its next contour point is measured. If this distance is more than certain threshold value, then we consider this region as possible segment line.

Over Segmentation removal In the segmentation process, due to bad quality images, words are getting over segmented. We tried to minimize the over segmentation problem by setting some threshold. If the distance between two segment lines are less than the threshold value, then that segment line will be rolled out and consider its next segment line. The threshold was set by trial and error method initially, can be changed to a much concrete formula later.

Image Standardization

As we are going to evaluate the segmented letter images on standardize trained image model which is of fixed size 28*28 pixel, we resize the images with same size including padding of 4 pixels on each side. We cropped the various sized input images to cut the extra padding on each side and then resized them to 20*20 pixel.

CNN based sequential model

Using CNN sequential model for segmentation purpose.
In this application we propose the use of Convolution Neural Networks for recognition of characters which utilizes the unsupervised learning technique for training. We have created CNN multilayered model using KERAS backed by Tensorflow. For training purpose, we have used MNIST character dataset. The dataset consists of over 1L labeled images of handwritten letters. The images are in 28*28-pixel format. We trained the model on 70% of data and tested on 30% of data. The measured accuracy got from the CNN model is around 90%.

CNN Layers:-
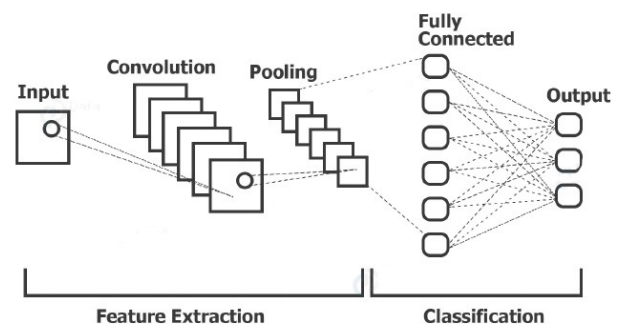
Firstly pure segmentation is applied and then recognitionbased segmentation is applied.
After applying pure edge based and threshold based segmentation we store the segment points in an array then the

neural network window is slided the characters are being recognized ,so after recognition it will choose that as a segment point.Now both the segment points(point in the pure segmented array and recognized array is being compared,if they different then the average of both points is to be considered as the correct segmentation point.

Pseudo code

Pseudo code:-
1.Find the feature points.
2.The text lines are separated out by extracting the row wise consecutive black pixel existence.
3.Use contour tracing to accurately segment the
intersecting components. Then bounding box is applied on contours to extract words.
4.The top contour points are extracted by taking first black pixel of each column from top.
5.Check if the pixel count is less than min-pixel threshold,if yes add segmentation point.
6.Check if 2 consecutive segmentation points are greater than min-seperation threshold in distance,if yes add to the segmented array.
7.In the segmentation process, due to bad quality images, words are getting over segmented. We tried to
minimize the over segmentation problem by setting some threshold. If the distance between two
segment lines are less than the threshold value, then that segment line will be rolled out and consider its
next segment line. The threshold was set by trial and error method initially, can be changed to a much
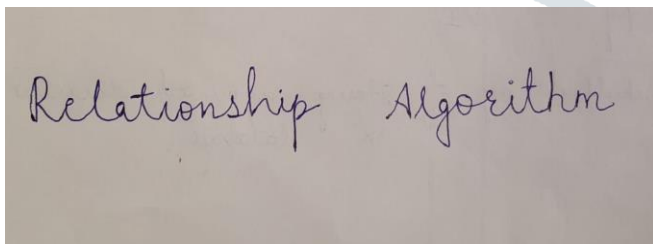concrete formula later.

Testing

Neural network is being tested upon 30% of MINIST dataset as well as with handwritings of people around us.

We have tested upon 100 samples of handwritings of people around us.

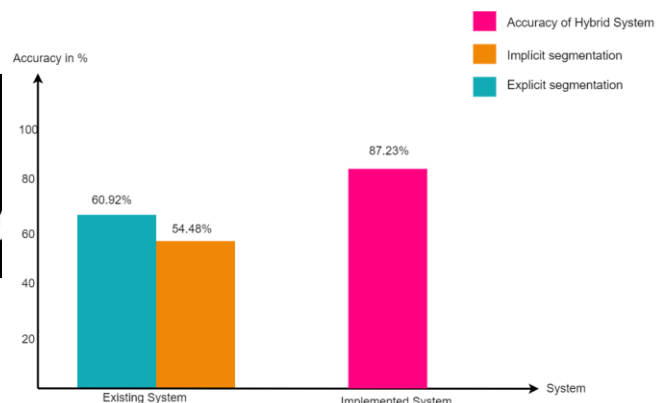## V. SEGMENTED RESULTS
Input Image:-

Segmented Output:-

VI. COMPARISON WITH PREVIOUS WORK

Comparision Table

Segmentation of Connected Handwritten Characters	Implicit segmentation and Explicit segmentation.	Implicite - 60.92% Explicit - 54.48%
Hybrid Segmentation of Connected Handwritten Characters	Used edge and region based alog with CNN based segmentation	accuracy achieved is 87.23%

Comparision Graph

techniques for offline cursive handwritten text", vol-3 Punjab University.

VIII.CONCLUSION

The Proposed Sysytem has been implemented and yield the expected results. The Accuracy of the model so far calculated is 87.23%.

The idea and technique of making a hybrid model for accuracy is successfully implemented. Once segmented the further work of recognize will hopefully be much easier.

VII.   REFERENCES

1. Amit Choudharya,*, Rahul Rishib , Savita Ahlawat, A New Character Segmentation 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.

2. A. Choudhary, "A review of various character segmentation techniques for cursive handwritten words recognition," International journal o Information and Computation Technology, vol. 4, pp. 559-564, 2014.

3. 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.

4. 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.

5. A. Rehman, D. Mohamad, and G. Sulong "Implicit Vs Explicit based script segmentation and recognition: A performance comparison on benchmark database," International Journal Open Problems Compt. Math., vol. 2, pp. 352-364, 2009.

6. H. Lee, and B. Verma, "Binary segmentation with neural validation for cursive handwriting recognition," Proceedings of International Joint Conference on Neural Networks, pp. 1730-1735, 2000

7. Alaa Hamid and Ramzi Haraty, "A Neuro Heuristic approach for segmenting handwritten Arabic Text and verification using CNN", Lebanese American University, P.O. Box 13-5053 Chouran Beirut, Lebanon 1102 2801

8. CATALIN AMZA, "A review on Neural Network-Based Image Segmentation Techniques", De Montfort University, Leicester, LE1 9BH

9. Varun Wahi1 and Prabukumar Manoharan2 , " Survey and Comparison of various Image Segmentation Techniques", International Journal of Pure and Applied Mathematics Volume 120 No. 6 2018, 11799-11812

10. Amandeep Kaur, Seema Baghla and Sunil Kumar, "Study of various character segmentation