Optical Character Recognition for English Handwritten Text Using Recurrent Neural Network

R.Parthiban[1,]R.Ezhilarasi[2],D.Saravanan[3]
Associate Professor[1,3,] Department of CSE, IFET College of Engineering, Villupuram[1,3]
Final Year Students[2],Department of CSE, IFET College of Engineering, Villupuram[2]
parthineyveli@gmail.com[1],ezhilarasir98@gmail.com[2], saranmds@gmail.com[3]

***Abstract*—Manually written Text Recognition is an innovation that is genuinely necessary right now of today.Before appropriate execution of this innovation we have depended on composing writings with own may leads to some mistakes. It is hard to maintain safely and gathering that information with effectiveness.Difficult work is required so as to keep up appropriate association of the data.Recurrent neural networkis utilized to discover the arrangement of character.Today we have OCRs effectively accessible for the English language.We can discover OCRs for formal texted English also yet OCRs for written by hand content are uncommon.Furthermore,those which are accessible don't have a goodaccuracy.We expect to make such an OCR which gives us an impressiverecognition exactness for manually written Text using recurrent neural network.The proposed model is implemented using Conda, used with TensorflowFramework. The purpose of Recurrent neural network is to improve accuracy.**

Keywords—English handwritten character; optical character recognition; Recurrent neural network; accuracy.

I.INTRODUCTION

English is most commonly used language in the world and also official language among 53countries.The quantity of scientific papers written in English has begun to exceed the quantity of papers written in the local language of the analyst. In the Netherlands, for instance, the proportion is an amazing 40 to 1. Thus, having an information on English is unbelievably essential to those working in the logical field. Nowadays there is an enormous interest in putting away the data accessible in these paper records into a PC storage disk and afterward reusing this data is hard via looking through procedure. Optical Character Recognition (OCR) passes on the which means of perceived English characters just as digits that perhaps pictures of manually written content, or might be simply PC content text styles of different kinds. It is an application programming which examinations and forms a picture record to perceive productively the characters present inside it. The picture record can be a written by hand and examined photograph.It makes an interpretation of pictures into unmistakable machine encoded editable content. It perceives just those characters for which the framework has been prepared for utilizing explicit arrangement calculation.There are many algorithms are utilized and tested for English handwritten text. Though many technologies were present we will not get 100 percent accuracy. The principle issue emerges because of the way that we are doing it for manually written content. So, our example set is endless and alsodifferent samples have differentcharacteristics.The handwriting samples are collected from different persons; hence it is very unlikely that they will follow a similar pattern.

The section of this paper is organized as follows, and part 2 contains related works on handwritten recognition using optical character recognition with algorithms. In section 3, the processing flow of modules with consensus rules and roles assigned to the process, section 4 reviews tool analysis and requirements for proposed system implementation. Section 5 concludes the paper with future research.

II. LITERATURE SURVEY

This area audits related works of optical character recognition for manually handwritten texts.*Mujadded Al RabbaniAlif*[1] proposed an altered adaptation of ResNet-18 engineering which is especially strong in characterizing Bangla separated transcribed characters. *Tapan Kumar Hazra*[5] proposed OCR for separate particular highlights from the info picture for arranging its substance as characters explicitly letters and digits.*Dhara S. Joshi*[6] proposed a methodology that utilizes KNN to perceive manually written or printed content. They utilized the Gujarati Characters for the acknowledgment into the machine editable format and also include deep learning.*BalaMallikarjunaraoGarlapati*[3] proposed a methodology for machine print and transcribed content order at word level utilizing force and shape basic highlights of checked content. The proposed technique accomplished great characterization productivity on IAM dataset. *ThulasiKishna N.P*[8] proposed a methodology for the most part centersaround the acknowledgment of written by hand Malayalam characters. Consequently, cursive Malayalam characters can be perceived by Hidden Markov Model.The grouping is finished with Artificial Neural System.*SujalaK,Ajay James*[14] proposed a work that executes a transcribed Malayalam character acknowledgment framework for 15 vowels and 36 consonants. The proposed technique for OCR is a mixture approach for highlight extraction joining basic and measurable highlights. *Mustafa Ali Abuzaraida*[2] proposed framework is intended to bargain with a book. Not with

standing, the examination is constrained to manage procured digits which will give a comprehension of utilizing preprocessing steps right now.*RyosukeOdate*[13] proposed a tree-based bunching strategy consolidated with Linear Discriminant Analysis and it worked fine with ETL9B dataset comprising of Japanese written by hand characters. *BinnyThakral*[15] proposed another system for the division of conjuncts, and covering characters in Devanagari content on Hindi language. The proposed calculation is centeredaround Cluster Detection system.

III. PROPOSED METHODOLOGY

This proposed frameworkintroduces a Recurrent neural network for recognize English handwritten text. Presently we are making an OCR for written by hand English content. The primary issue emerges because of the way that we are doing it for written by handwritten text.So, our example set is infinite.Likewise, various examples have various attributes. The penmanship tests are gathered from various people, henceforth it is impossible that they will follow a comparative example. We have followed a bottom up strategy in our methodology, for example we start with a particular example and afterward approach towards the general arrangement. We take a specific example apply our philosophy to it and discover the outcomes. At that point we re-play out the calculation on a subsequent example set and relying on the presentation of our strategy on this set we continue improving our procedure until it insinuates towards a general arrangement. Let discuss deeply about the proposed system with its flow chart that consist of four main stages with output image.It RNN layers and a last Connectionist Temporal Classification (CTC) layer.

Fig. 1.Flow chart of proposed approach.

A-Dataset

Gathering transcribed information from various writers.Each personhas different style of handwriting.All those different documents are collected and Scannedthose datasets. From these Crop each character physically and Label each character separately as a image.Then it is used for optical character recognition.

B-Pre-processing

Resize picture to 30*30 pixels and Convert to grayscale structure.Alter pixels force. Include cushioning of 2 pixels all sides.it is a grayscale picture of size 128×32.

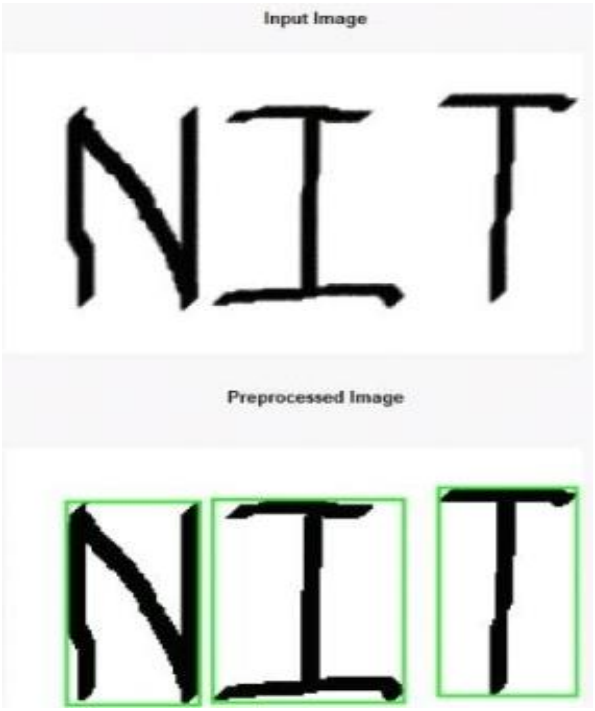

Fig. 2.Pre-processed image.

Ordinarily, the pictures from the dataset don't have precisely this size, along these lines we resize it until it either has a width of 128 or a tallness of 32. At that point, we duplicate the picture into an objective picture of size 128×32. This procedure is indicated in.Finally, we standardize the dark estimations of the picture which streamlines the errand for the NN. Information increase can without much of a stretch be coordinated by duplicating the picture to irregular situations as opposed to adjusting it to one side or by arbitrarily resizing the picture.

C-Training and Testing Set

Randomly split data into training and testing set.the characters are anticipated precisely at the position theyshow up in the picture (for example look at the situation of the "I" in the picture and in the diagram). Just the last character "e" isn't adjusted. However, this is OK, as the CTC activity is without division and couldn't care less about outright positions. From the base most diagram demonstrating the scores for the characters "l", "I", "t","e" and the CTC clear name, the content can undoubtedly be decoded: we simply take the most plausible character from each time-step, this structures the alleged best way, at that point we discard rehashed characters lastly all spaces: "l - ii- - t-t- - l-… - e" → “l-- t-t- - l-… - e" → "little”.

In the Fig.3, we see that the OCR has effectively distinguished the characters (digits and letters in order) from the picture of a manually written paper for example address of individual and composed digits.we see that the OCR has effectively distinguished the digits from the picture all together from option to left.

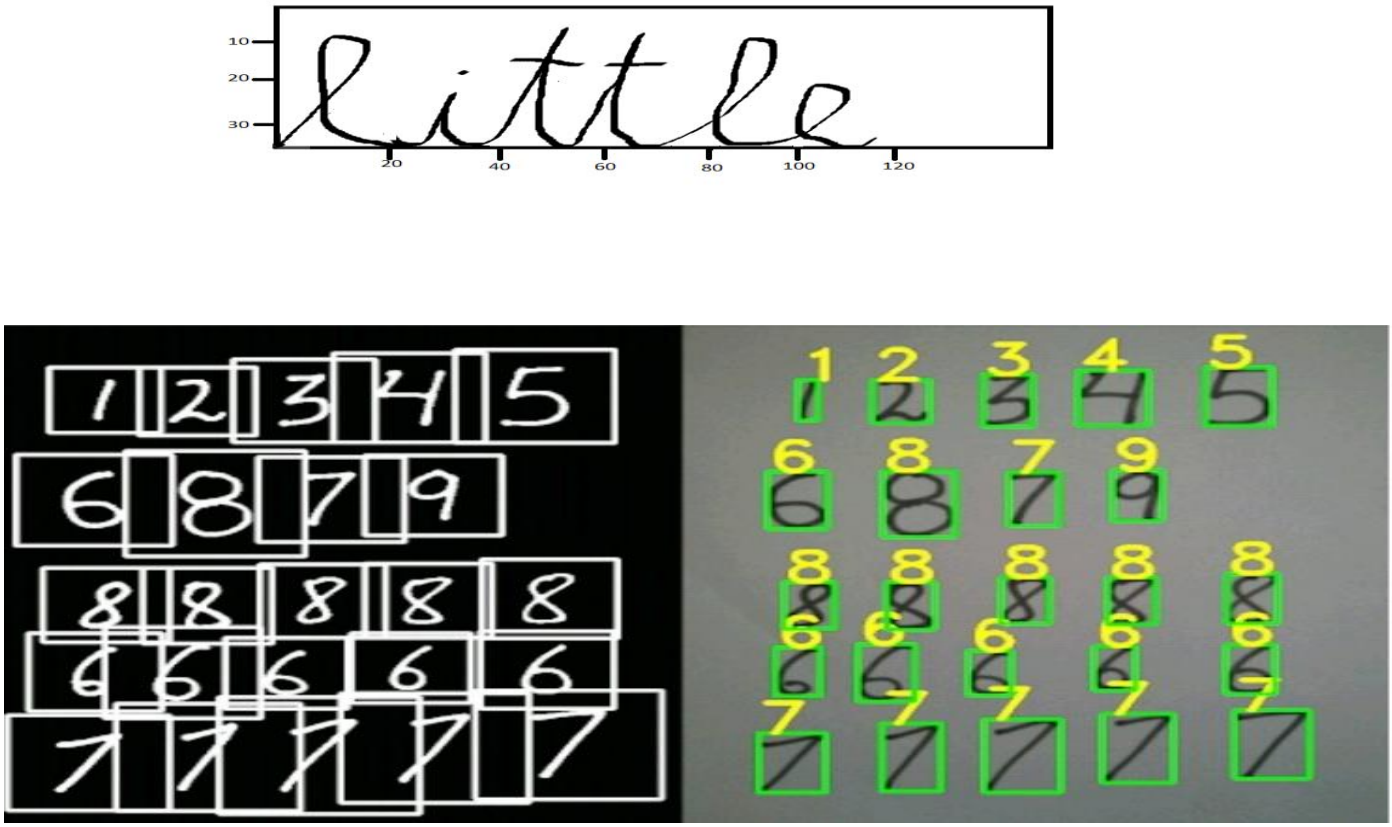

Fig. 3.Trained and Testing of English character and digits.

D-Recurrent Neural Network

In customary neural systems, all the data sources and yields are autonomous of one another, however in cases like when it is required to foresee the following expression of a sentence, the past words are required and consequently there is a need to recollect the past words. Accordingly, RNN appeared, which comprehended this issue with the assistance of a Hidden Layer.The Input layer will hold the pixel estimations of the picture.

The Recurrent layer will decide the yield of neurons of which are associated with nearby locales of the contribution through the count of the scalar item between their loads and the district associated with the information volume. The amended direct unit (generally abbreviated to ReLu) plans to apply an 'elementwise' actuation capacity, for example, sigmoid to the yield of the initiation created by the past.TheRNN spreads pertinent data through this arrangement. The well known Long Short-Term Memory (LSTM) execution of RNNs is utilized, as it can spread data through longer separations and gives more vigorous preparing qualities than vanilla RNN. The RNN yield succession is mapped to a network of size 32×80. The IAM dataset comprises of 79 distinct characters, further one extra character is required for the CTC activity (CTC clear name), along these lines there are 80 passages for every one of the 32 time-steps.

E-Output Image

Finally, the output image is given as formal printed text.Word exactness rate (WAR) level of words with all characters effectively recognized.Character precision rate(CAR) normal Levenshtein separation standardized by the length of the longest word and subtracted from 1.Then the recognized output are shown in following figure.

When compared to existing system our proposed system executed with high accuracy.

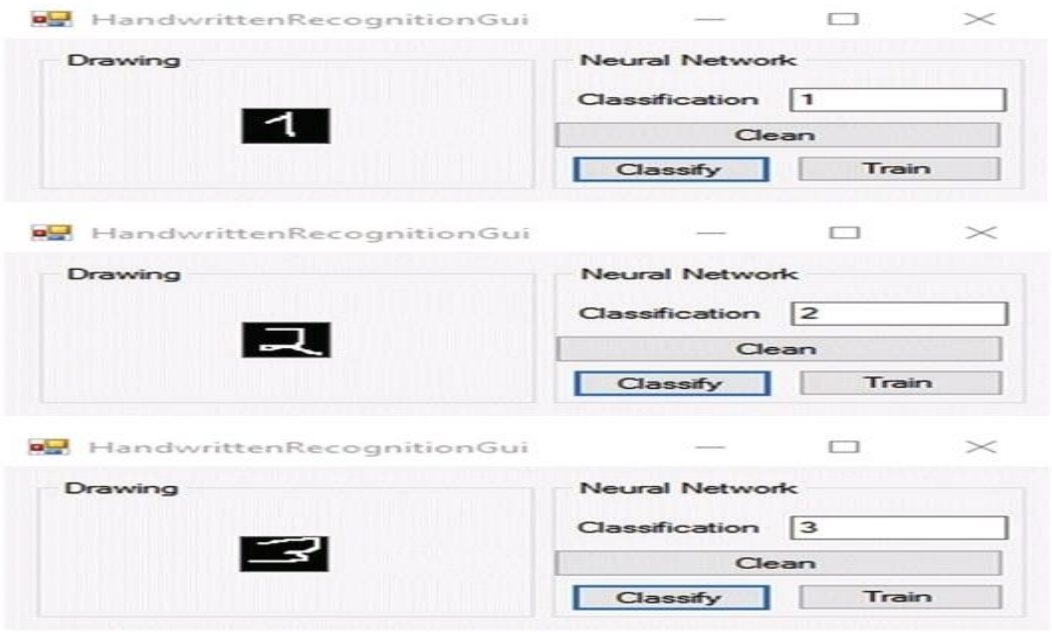

Fig. 5.Recognized output for digits.

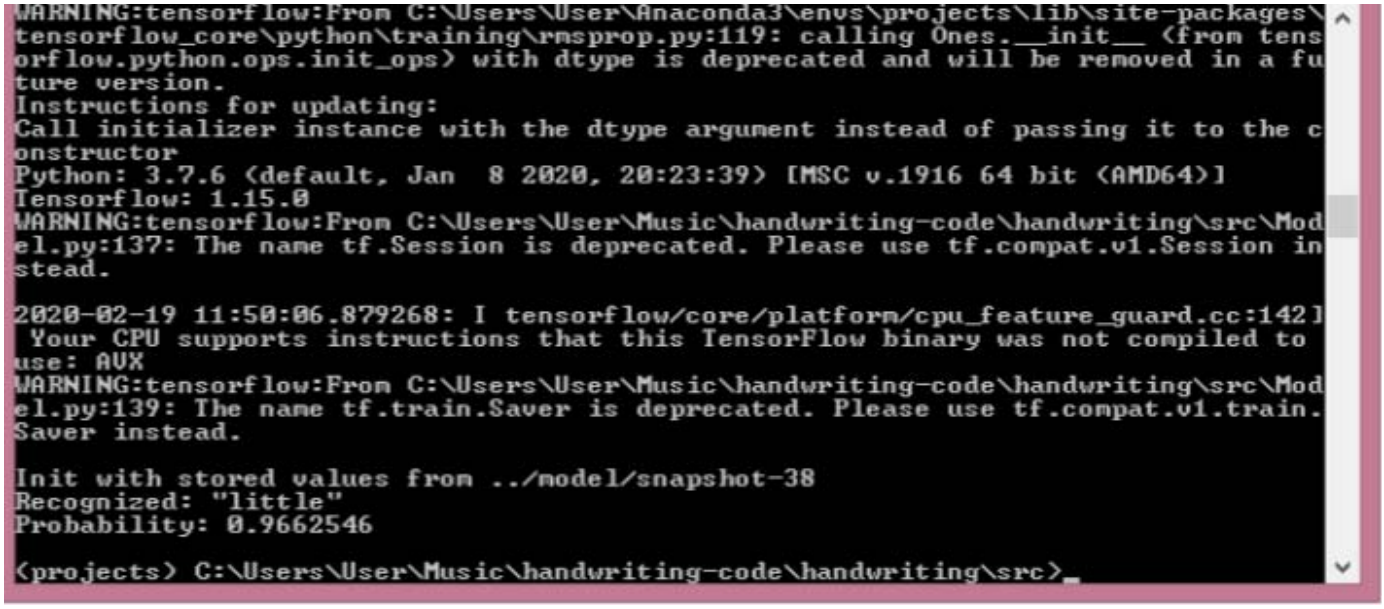

Fig. 6.Recognized output for text.

Requirements needed for our proposed system:

Hardware Requirements:System: Pentium IV 2.4 GHz.HardDisk: 40 GB.FloppyDrive : 44 Mb.Monitor :15 VGA Colour.Ram : 512 Mb.Software Requirements:- Operating system : Windows XP/7.Coding Language: Python.Tool: Python IDE.

3 Python language is created under an OSI-endorsed open source permit, which makes it allowed to utilize and circulate, including for business reason.

The accompanying propels to choose python as advancement device:

1 Very easy to comprehend and utilize

2 Extensive Support Libraries - Python gives a huge standard library with absences of helpful capacities most appropriate for creating application programming

4 The size of the code is decreased.

5Python has worked in rundown and word reference information structureswhich can be utilized to develop quick runtime information structures. Further, Python likewise gives the alternative of dynamic elevated levelinformation composing which decreases the length of help code that is required.

IV.CONCLUSION

In this Proposed framework the handwritten English document are scanned and recognized optically.Byusing RNN algorithm output are got as printed text with 90% accuracy.So as to additionally improve the accomplished exhibition, we needto examine the issue further for discovering better arrangement by planning a totally new engineering for English content. We left that part as our future work.

REFERENCES

[1] Mujadded Al RabbaniAlif, Sabbir Ahmed, Muhammad AbulHasan, "Isolated Bangla Handwritten Character Recognition with Convolutional Neural Network", 978-1-5386-1150-0/17/$31.00 c 2017 IEEE

[2] Mustafa Ali Abuzaraida, Salem MeftahJebriel," The Detection of the Suitable Reduction Value of DouglasPeucker Algorithm in Online Handwritten Recognition Systems", 978-1-4673-8480-3/15/$31.00©2015 IEEE

[3] BalaMallikarjunaraoGarlapati, SrinivasaRaoChalamala, "A System for Handwritten and Printed Text Classification" 978-1-5386-2735-8/17 $31.00 © 2017 IEEE

[4] Soumik Bhattacharya, DurjoySenMaitra, Ujjwal Bhattacharya, Swapan K. Parui, "An End-to-End Systefor Bangla Online Handwriting Recognition", 2167-6445/16 $31.0© 2016 IEEE

[5] Tapan Kumar Hazra, Dhirendrapratapsingh, NikunjDaga," Optical Character Recognition using KNN onCustom Image Dataset", 978-1-5386-2215-5/17/$31.00 ©2017IEEE

[6] 1Dhara S. Joshi, 2Yogesh R. Risodkar," Deep Learning Based Gujarati Handwritten CharacterRecognition", 978-1-5386-0926-2/18/$31.00 ©2018 IEEE

[7] Kha Cong Nguyen, Nakagawa Masaki, "Enhanced Character Segmentation for Format-Free Japanese TextRecognition", 2167-6445/16 $31.00 © 2016 IEEE

[8] ThulasiKishna N.P, Seenia Francis, "Intelligent Tool For Malayalam Cursive Handwritten CharacterRecognition Using Artificial Neural Network And HiddenMarkov Model", 978-1-5386-4031-9/17/$31.00©2017 IEEE

[9] Qi Li, Weihua An, Anmi Zhou, Lehui Ma, "Recognition of Offline Handwritten Chinese Characters Usingthe Tesseract Open Source OCR Engine", 978-1-5090-0768-4/16$31.00 © 2016 IEEE

[10] Jianjuan Liang, Bilan Zhu and Masaki Nakagawa, "ACandidate Lattice Refinement Method for OnlineHandwritten Japanese Text Recognition", 2167-6445/16 $31.00© 2016 IEEE

[11] Michael Murdock, Jack Reese, Shawn Reid, "ICFHR2016 Competition on Local Attribute Detection for Handwriting Recogntion", 2167-6445/16 $31.00 © 2016 IEEE

[12] Pranav P Nair, Ajay James, C Saravanan, "Malayalam Handwritten Character Recognition Using Convolutional Neural Network", 978-1-5090-5297-4/17/$31.00©2017 IEEE

[13] RyosukeOdate, Hideaki Goto, "FAST AND ACCURATE CANDIDATE REDUCTION USING THE MULTICLASS LDA FOR JAPANESE/CHINESE CHARACTER RECOGNITION", 978-1-4799-8339-1/15/$31.00 ©2015 IEEE

[14] SujalaK , Ajay James, C. Saravanan, " A Hybrid Approach for Feature Extraction in Malayalam Handwritten Character Recognition", 978-1-5090-3239-6/17/$31.00©2017IEEE

[15] BinnyThakral*, Manoj Kumar**,"Devanagari Handwritten Text Segmentation for Overlapping and Conjunct Characters- A Proficient Technique", 978-1-4799-681/14/$31.00 ©2014 IEEE

[16] Matthias Zimmermann, Jean-Ce´dricChappelier, and Horst Bunke, "Offline Grammar-Based Recognition of Handwritten Sentences", 0162-8828/06/$20.00 2006 IEEE