Optical Character Recognition for different types of English Documents

A Project Report

submitted by

GUBBALA ROSHAN

in partial fulfilment of the requirements for the

award of the degree of

BACHELOR OF TECHNOLOGY &

MASTER OF TECHNOLOGY

DEPARTMENT OF ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY, MADRAS

June 2021

THESIS CERTIFICATE

This is to certify that the thesis titled **Optical Character Recognition for different types of English documents**, submitted by **Gubbala Roshan (EE16B138)**, to the Indian Institute of Technology, Madras, for the award of the degree of **Bachelor of Technology & Master of Technology,** is a bonafide record of the research work done by him under our supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Prof. Srinivasa Chakravarthy
Research Guide
Professor
Dept. of Biotechnology
IIT Madras, 600036

Place: Chennai
Date: 27th June 2021

ACKNOWLEDGEMENTS

Foremost, I am very grateful to my research guide **Professor Srinivasa Chakravarthy** for his constant invaluable guidance and support throughout the project. His expertise in the area and discussions with him has motivated me to learn more. It has been a great learning curve for me under his guidance.

I would also like to acknowledge **Professor K.Sridharan** for supporting me as a co-guide for this project. Also, special thanks to **Vigneswaran**, for constantly resolving my doubts and helping me in my thought process. I would also like to thank my other colleagues from the lab for always being ready to help and accepting me as a part of the lab.

Any acknowledgement will be incomplete without a mention of my parents and brother. I am extremely grateful for their unceasing support and encouragement. They have always stood by me through my years of study.

ABSTRACT

There is a strong need for storing information to a computer storage drive from data available in printed or handwritten documents or photographs so that it can be re-used by computers in a variety of industries. Scanning the documents and then storing them as image files could be a straightforward approach to save information to a computer system from these printed materials. However, reading or querying text or other information from these image files would be extremely difficult to re-use. As a result, a method for retrieving and storing information, particularly text, from image files is required. Optical character recognition (OCR) is a study topic that aims to create a computer system that can automatically extract and analyse text from photographs. The goal of OCR is to convert any type of text or text-containing document, such as handwritten text, printed text, or scanned text images, into an editable digital format for deeper and more advanced processing. As a result, OCR allows a machine to recognise text in such documents automatically. In order to accomplish successful automation, some main problems must be identified and addressed. Character font qualities in paper documents and image quality are just two of the most recent difficulties. Characters may not be recognised accurately by computer systems as a result of these issues. This paper investigates OCR in three different types of English Documents namely Normal Printed text , Old Newspapers and Handwritten text Documents.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

It is normal and expected for humans to require that machines identify patterns be built and designed. From automated optical character recognition to face recognition, fingerprint identification, speech recognition, DNA sequence identification, and many other applications, machine pattern recognition is clearly beneficial. Optical character recognition(OCR) is a study topic that aims to create a computer system that can automatically extract and analyse text from photographs. There is a large demand these days for storing information on a computer storage drive from data available in printed or handwritten documents so that it can be re-used by computers later. One simple way to store information to a computer system from these paper documents could be to first scan the documents and then store them as image files. However, reading or querying text or other information from these image files would be extremely difficult to re-use. Therefore a technique to automatically retrieve and store information, in particular text, from image files is needed. Of course, this is not a very trivial task. Some major challenges need to be laid out and handled in order to achieve successful automation. The font characteristics of the characters in paper documents and quality of images are only some of the recent challenges. Due to these challenges, characters sometimes may not be recognized correctly by the computer system. Thus there is a need for mechanisms of character recognition to perform Document Image Analysis (DIA) which overcomes these challenges and produces electronic format from the transformed documents in paper format.

Similarly, Optical Character Recognition (OCR) is the process of modification or conversion of any form of text or text-containing documents such as handwritten text, printed or scanned text images, into an editable digital format for deeper and further processing. Optical character recognition technology allows a machine to recognise text in such documents automatically. In the real world, it's similar to the brain and eye of a human being.Although the human eye can detect, view, and extract text from images, it is the human brain that interprets the detected or extracted text read by the eye. Of course OCR technology has not advanced enough to compete with human’s ability. The performance and accuracy of OCR is directly dependent upon the quality of input documents. Again, when we think of a human's ability to recognize text, the performance of the brain's process directly depends upon the quality of the input read by eye. Several issues and concerns can arise while creating and deploying a computerised OCR system. For example, there is just enough difference between some digits and letters for computers to recognise and differentiate them accurately. For example, it may not be very easy for computers to differentiate between digit “0” and letter “o”, especially when these characters are embedded in a very dark and noisy background. One of the main focuses of OCR research has been to recognize cursive scripts and handwritten text for its broad application area. Today, to solve the text recognition problem several different types of OCR software exist such as Desktop OCR, Server OCR, web OCR and so on.

RELEVANT WORKS

The rise in popularity of Deep Learning has resulted in a greater demand for annotated data for supervised learning tasks. The majority of early datasets were manually annotated, and it was only recently that attempts were made to automate the annotation process and reduce the human work involved. However, it is vital to remember that in some circumstances, user feedback is required in order to provide the greatest quality of annotated data that is practically error-free.

In disciplines as diverse as real-time video feeds, object detection, and even the semantic web, initiatives to automate data annotation have been made. It's not unexpected that there's been a lot of research in this field, given the obvious requirement for annotated data for offline handwritten text recognition, especially with the transition towards Deep Learning-based systems for offline handwriting text recognition. A number of papers have proposed a systematic arrangement of stages to develop a comprehensive annotation engine for handwritten text, with varied levels of automation. "*A Semi-automatic Annotation Scheme for Bangla Online Mixed Cursive Handwriting Samples,*" *by U. Bhattacharya, R. Banerjee, S. Baral, R. De and S. K. Parui* in 2012, and "*Automated Semantic Annotation of Species Names in Handwritten Texts " by Stork L., Weber A., van den Herik J., Plaat A., Verbeek F., Wolstencroft K* in 2019 are few such attempts. However, as mentioned in a section of a study by *Ung et al,* creating a comprehensive end-to-end pipeline to annotate handwritten text with very little human interaction remains a difficult issue .

The two main components of our annotation pipeline are a word detection system and a handwriting recognition system. There has been extensive research in these fields for the past many years. There have been non-Machine Learning approaches, Machine Learning approaches, and most recently, Deep Learning approaches. *Lavrenko et al.* presented a holistic word recognition strategy based on Hidden Markov Models, motivated by cognitive psychology findings, for handwritten historical texts. This approach, which did not use character segmentation, had a recognition accuracy of 65 percent, which was higher than the findings of other systems at the time. Optical character recognition (OCR) technologies have been available for a long time, and many attempts at handwritten OCR for many languages have been made. Over this time, OCR approaches have progressed significantly, and in recent years, with the advent of cloud computing, GPUs, and a stronger research community, have switched toward some extremely amazing Deep Learning-based models. However, as noted in the preceding section, OCR algorithms encounter a number of difficulties and have yet to produce compelling results for cursive handwritten text documents with no specified structure.

A work done by *Shiedl et al.* implements a handwriting recognition system, with an architecture based on Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN) and a Connectionist Temporal Classifier (CTC), which provides impressive results for handwriting recognition. This was one of the main inspirations for the recognition architecture we present in this work. The 2D LSTM implemented in the recognition system of our work, is inspired by a work by *Graves et al.* Text detection has also gotten a lot of attention recently, thanks to advances in image processing, object detection, and deep learning. *Youngmin Baek.* proposed a simple yet robust deep learning based scene text detector known as CRAFT text detector. In our work we build upon this CRAFT model as it outperforms a majority of the state-of-the-art text detectors in terms of accuracy and speed ,and is easy to build as well.

CHAPTER 2

Text Detection

CRAFT text detector is a novel text detector and a pre-trained model that effectively detects text area by exploring each character region and affinity between characters. The bounding box of texts are obtained by simply finding minimum bounding rectangles on a binary map after thresholding character region and affinity scores.

Figure 1. Visualization of character-level detection using CRAFT. (a) Heatmap predicted by the framework. (b) Detection result

The detector localizes the individual character regions and links the detected characters to a text instance. It is designed with a convolutional neural network producing the character region score and affinity score.The region score is used to localize individual characters in the image, and the affinity score is used to group each character into a single instance. Figure. 1 is a visualization of CRAFT's result on a sample image.

CRAFT adopts a fully convolutional network architecture based on VGG-16 as its backbone. In simple words, VGG16 is essentially the feature extracting architecture that is used to encode the network's input into a certain feature representation. The decoding segment of the CRAFT network is similar to UNet. It has skip connections that aggregate low-level features.

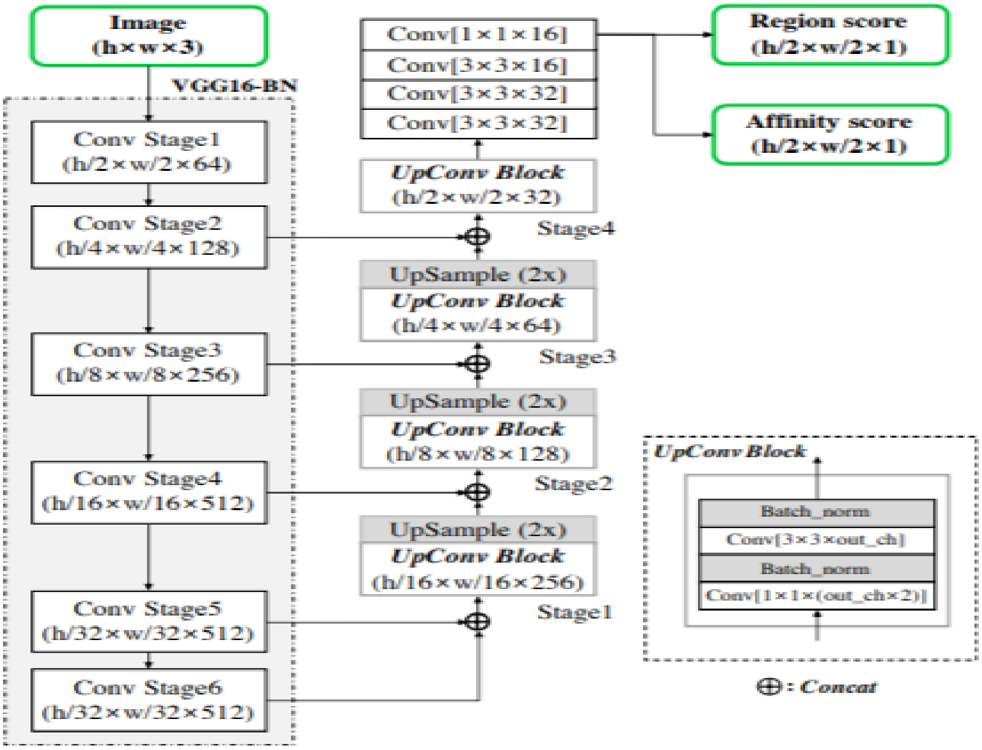


Figure 2. Schematic illustration of CRAFT network architecture

CRAFT predicts two scores for each character:

- **Region Score:** As the name suggests, it gives the region of the character. It localizes the character
- **Affinity Score:** 'Affinity' is the degree to which a substance tends to combine with another. So, an affinity score merges characters into a single instance (a word).

CRAFT generates two maps as output: Region Level Map and Affinity Map.Let's understand what they mean by looking at an example:

Figure 3. Sample image applied with Craft text detector

The areas where the characters are present are marked in the Region Map:

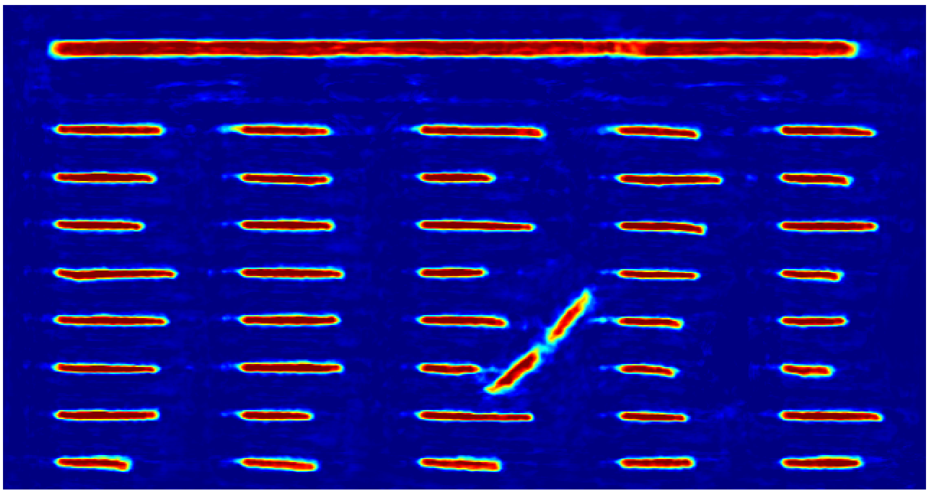


Figure 4. Region map of Sample image

The Affinity Map pictorially represents the related character. Red symbolizes the characters have a high affinity and must be merged into a word:

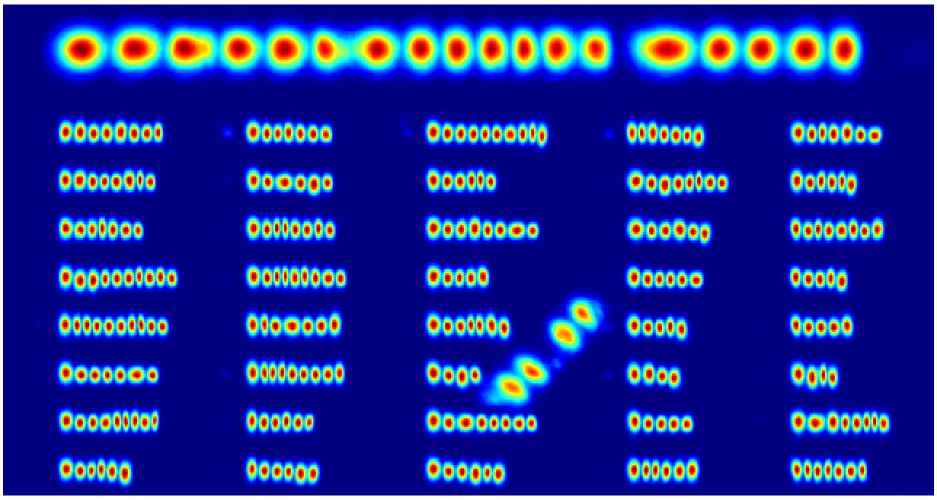


Figure 5. Affinity Map of Sample image

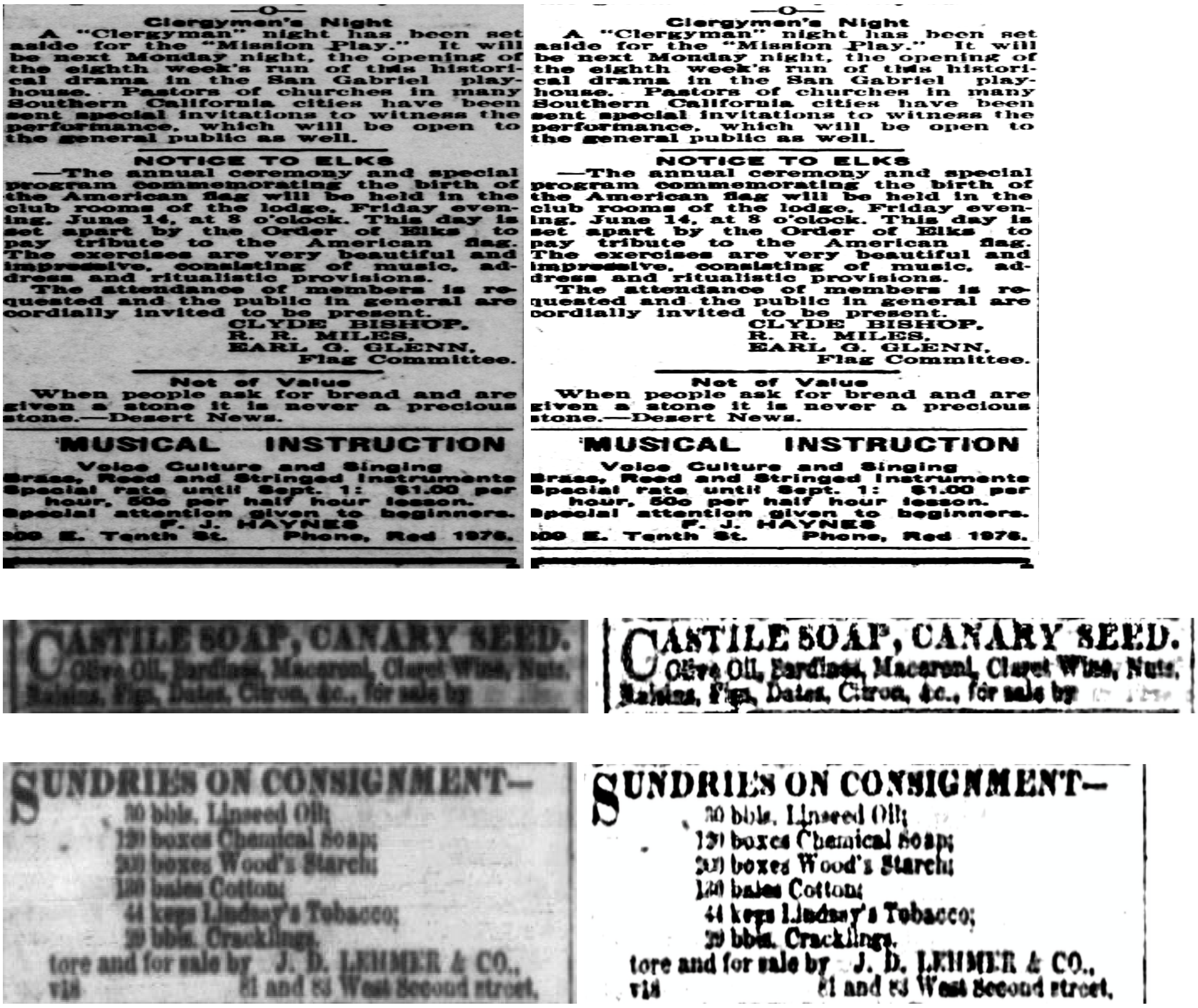
Finally, the affinity and region scores are combined to give the bounding box of each word. The coordinates are in the order:(left-top), (right-top) (right-bottom), (left-bottom), where each coordinate is an (x, y) pair. The result of text detection:

Figure 6. Detected text output of Craft detector

Image Binarization & Enhancement

For good quality and high accuracy character recognition, OCR techniques expect high quality or high resolution images with some basic structural properties such as high differentiating text and background. The way images are generated is an important and determining factor in the accuracy and success of OCR, since this often affects the quality of images dramatically. Preprocessing techniques are needed on colour, grey-level or binary document images containing text and/or graphics.

Binarization is the starting step of most document image analysis systems and refers to the conversion of the gray-scale image to a binary image. Since historical document collections are most of the times of very low quality, an image enhancement stage is also essential. The scheme used for image binarization and enhancement consists of five distinct steps: a preprocessing procedure using a lowpass Wiener filter, a rough estimation of foreground regions using Niblack's approach, a background surface calculation by interpolating neighboring background intensities, a thresholding by combining the calculated background surface with the original image and finally a post-processing step that improves the quality of text regions and preserves stroke connectivity.



Clergymen's Night

A "Clergyman" night has been set aside for the "Mission Play." It will be next Monday night, the opening of the eighth week's run of this historical drama in the San Gabriel playhouse. Pastors of churches in many Southern California cities have been sent special invitations to witness the performance, which will be open to the general public as well.

NOTICE TO ELKS

—The annual ceremony and special program commemorating the birth of the American flag will be held in the club rooms of the lodge, Friday evening, June 14, at 8 o'clock. This day is set apart by the Order of Elks to pay tribute to the American flag. The exercises are very beautiful and impressive, consisting of music, address and ritualistic provisions. The attendance of members is requested and the public in general are cordially invited to be present.

CLYDE BISHOP,
R. R. MILES,
EARL G. GLENN,
Flag Committee.

Not of Value

When people ask for bread and are given a stone it is never a precious stone.—Desert News.

MUSICAL INSTRUCTION

Voice Culture and Singing
Brass, Reed and Stringed Instruments
Special rate until Sept. 1: $1.00 per hour, 50c per half hour lesson.
Special attention given to beginners.
F. J. HAYNES
909 E. Tenth St. Phone, Red 1976.

CASTILE SOAP, CANARY SEED. Olive Oil, Sardines, Macaroni, Claret Wine, Nuts, Raisins, Figs, Dates, Citron, &c., for sale by

SUNDRIES ON CONSIGNMENT—
30 bbls. Linseed Oil;
120 boxes Chemical Soap;
200 boxes Wood's Starch;
180 bales Cotton;
44 kegs Lindsay's Tobacco;
20 bbls. Cracklings.
tore and for sale by J. D. LEHMER & CO.,
v18 81 and 83 West Second street.

Figure 7. shows Image enhancement used on Old Newspaper Articles.

Text Recognition

The images of the individual words , within each of the bounding boxes, are passed as inputs to the recognition system . Each input image that is passed into the recognition system goes through the three architectures of our recognition model : the Convolutional Neural Network (CNN), the multi-dimensional LSTM, and the Connectionist Temporal Classifier(CTC). The exact details and nature of these architectures are discussed in the next section.

For each input of an individual image to the recognition system, the output is a sequence of characters. The recognition system thus outputs a sequence of characters/ individual characters, for each of the detected, and serialized images passed into it from the previous phases of the pipeline. As the image passes through each of the layers of the CNN, the trained layers extract all the required features from that image. There are three main operations that are carried out on the image in the CNN, in each layer: the convolutional operation, a non-linear activation and a pooling function. Apart from these three operations, we add a Gaussian noise layer that adds standard Gaussian noise to the input, for reducing any chances of overfitting the dataset, and enable the recognition system to recognize the word to a good accuracy level irrespective of the gray values and contamination of the new image due to various natural causes. Then finally, after passing through two fully connected layers, a feature map is output.

The output feature map from the CNN is then passed as input to our 2D LSTM. An LSTM is used instead of a standard unidirectional or bidirectional RNN, as LSTMs prevent loss of information over long distances, and so is very helpful when dealing with long character sequences, which is very important for the task at hand. Our custom 2D LSTM was designed and implemented instead of using a standard one dimensional LSTM, because we felt that considering both the horizontal and vertical dimensions of handwritten text while recognizing it, would be much more effective than just working along one dimension. This is because the English cursive handwritten text has a myriad of variations along both the dimensions, and the system would be very robust if it could learn the features across both these dimensions. *Moysset et al. 's* paper shows that 2D LSTMs give great results for recognition of handwritten text, and provide higher performances as compared to single dimensional RNNs or LSTMs even when used on complex, challenging and real life data.

The output sequence of the LSTM is mapped to a matrix which becomes the input to the final CTC layer. Connectionist Temporal classification, a work by *Graves et al.* proposes a method that serves two purposes : it not only calculates the loss values required for training, but also decodes the matrix that is output from the LSTM, to obtain the final text that is present in the input image. During the training process, both the ground truth and the LSTM output matrix are fed to the CTC layer, and based on these, a loss value is calculated which is used to train the system to recognize the right sequence of characters. During the inference phase, only the LSTM output matrix is fed, and is decoded by the CTC layer, using a decoding algorithm, to get the text from the images. As the text from each input image is recognized by the recognition system, they are checked for any misspells, and are corrected using a state-of-the-art python spell checker known as Pyspellchecker. This spell checker, which was developed and released very recently, even offers a feature where the users can add words of their choice to the dictionary, thus allowing them to customize the dictionary to suit the task at hand. After this stage, all the recognized words and symbols are stored in the same serialized order.

RECOGNITION ARCHITECTURE USED:

The combination of a CNN, RNN and a CTC layer has been gaining popularity in the recent past, especially for text recognition tasks. We modify this architectural combination, and improve upon it by building our own 2 Dimensional LSTM to replace the standard RNN, which results in a significant increase in performance. Apart from this, the CNN in our system has also been designed in a way such that it is powerful and robust enough to deal with handwritten text images. The CNN model contains 10 layers, out of which each of the first 8 layers perform convolutional operations, non linear activations and pooling functions. The convolutional operations are carried out by filters of varying kernel sizes from 7x7 to 3x3, and a standard RELU non linear activation function is used. These 8 layers are followed by a Gaussian noise layer and 2 fully connected layers. The input to the CNN is the preprocessed image of dimensions 128x32, and the output is a feature map of size 32x512.The dimension of this 32x512 feature map that is passed as input to the LSTM, represents 512 features per time step, where each time step represents the position for the characters that may possibly be present in the word to be recognized. Each of these timesteps contain 512 relevant features extracted by the CNN layers. There are 32 timesteps because we set the maximum length of the character sequence that can be recognized, to 32. We found that for values greater than 32 the system performed worse and the loss values were considerably higher.

The 2D LSTM, which was built using 256 hidden cells, processes this feature map further, by only carrying forward the relevant information. The output of the 2D LSTM is finally mapped to a matrix of dimension 32x80, where 80 is the number of possible characters that can be recognized. This is because, apart from the 79 different characters present in our training dataset, another extra character is required for CTC operations, known as the CTC blank. Therefore, this 32x80 matrix contains the probability scores with respect to the 80 different possible entries for each timestep.

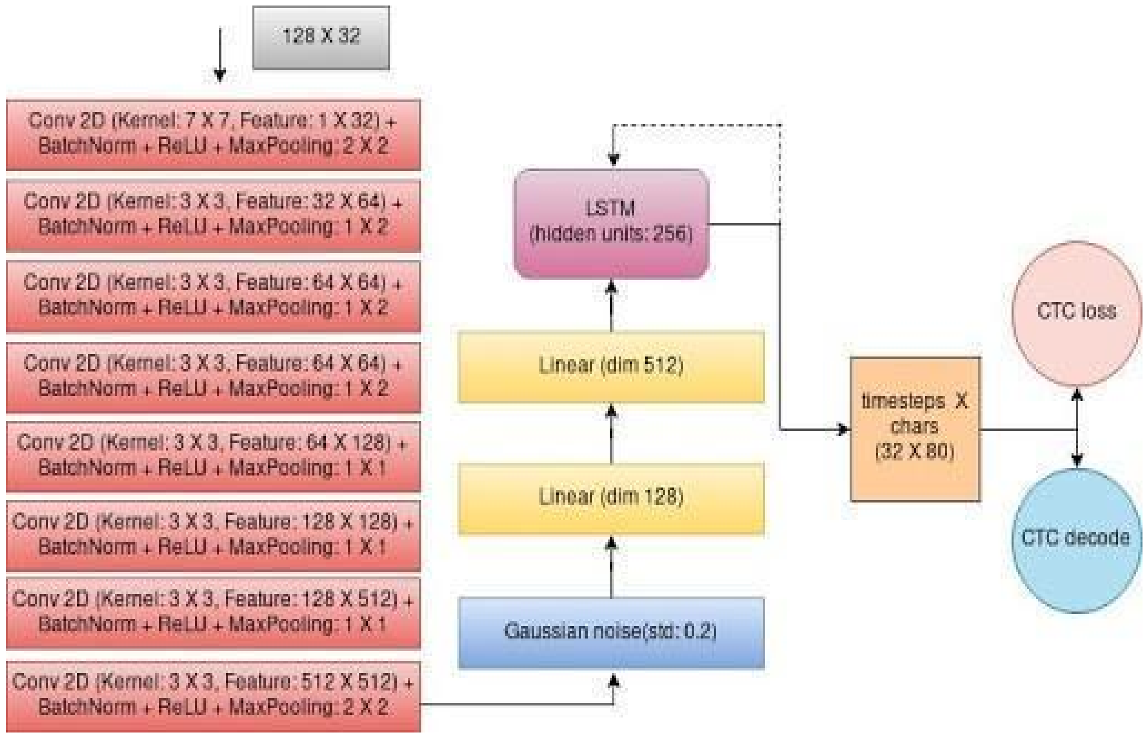


Figure 8. shows Summary of Recognition architecture

This matrix is provided as input to the CTC layer, which during training is compared with the ground truth tensor to generate a CTC loss value. This CTC loss value was the error metric that was considered for training. We used an RMSProp optimizer with a decaying learning rate that was initialized to 0.01, and a batch size of 50 for the training process. During the inference phase, the text in the image is decoded by the CTC layer using a CTC beam search decoding algorithm, which is offered as a feature in the Neural Net module of Tensorflow. This algorithm was used instead of the standard greedy best path decoding algorithm, because of which we were able to improve the accuracy of the recognized words even further.

Experimentation and Results

The OCR model is investigated in three different types of Documents namely Normal Printed text , Old newspaper articles and Handwritten text.

For Normal Printed text:

For training the data set is prepared by synthesising words which were randomly chosen from ASCII value range of 33-127.Further fonts ,noise level and padding white spaces are randomized to create different word images of uniform dimension 32x128. Then the OCR model is trained with 2,00,000 synthesized words for 146 epochs.

Character error rate (CER)	0.863
Word Accuracy	95.21

Table 1. Result parameters for Normal Printed text

--->y*\I7zJV`1E*H


--->(5R~FBr%O

Figure 9. Recognition results for Normal printed text

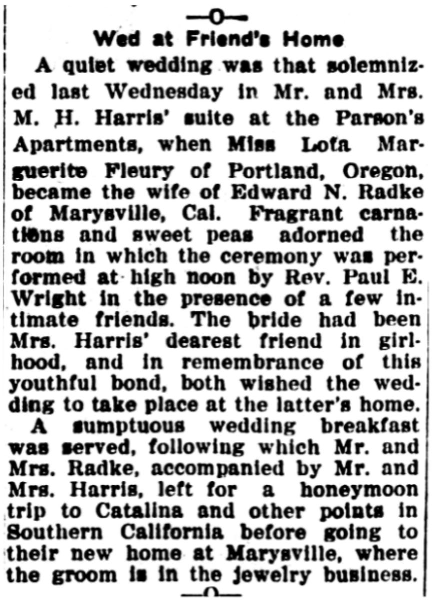
For Old Newspaper text:

For training the data set is prepared by generating different unique newspaper fonts words and then randomly dilated /eroded to make text resemble real historic newspaper words.For training, no image enhancement was used but for testing Histogram based Binarization was used as Image Enhancement technique.Then, the OCR model trained on printed text was fine tuned to Newspaper data of size of

2,00,000 words and ran for 65 epochs.

Character error rate (CER)	6.96
Word Accuracy	61.58

Table 2. Result parameters for Old Newspapers text



Wed at Friend's Home

A quiet wedding was that solemniz-ed last Wednesday in Mr. and Mrs. M. H. Harris' suite at the Parson's Apartments, when Miss Lota Mar-guerite Fleury of Portland, Oregon, became the wife of Edward N. Radke of Marysville, Cal. Fragrant carna-tions and sweet peas adorned the room in which the ceremony was per-formed at high noon by Rev. Paul E. Wright in the presence of a few in-timate friends. The bride had been Mrs. Harris' dearest friend in girl-hood, and in remembrance of this youthful bond, both wished the wed-ding to take place at the latter's home.

A sumptuous wedding breakfast was served, following which Mr. and Mrs. Radke, accompanied by Mr. and Mrs. Harris, left for a honeymoon trip to Catalina and other points in Southern California before going to their new home at Marysville, where the groom is in the jewelry business.

Wed at Friend's Home

A quiet wedding was that solemniz ed last Wednesday in Mr. and Mrs. M. H. Harris' suite at the Parson's Apartments, when Miss Lota Mar guerite Fleury of Portland, Oregon, became the wife of Edward N. Radke of Marysville, Cal. Fragrant carna tions and peas adorned the room in which the ceremony was per formed at high noon by Rev. Paul E. Wright in the presence of a few in timate friends. The bride had been Mrs. Harris' dearest friend in girl. hood, and in remembrance of this youthful bond, both wished the wed ding to take place at the latter's home.

A sumptuous wedding breakfast was served, following which Mr. and Mrs. Radke, accompanied by Mr. and Mrs. Harris, left for honeymoon a trip to Catalina and other points in Southern California before going to their new home at Marysville, where the groom is in the jewelry business.

Figure 10. Recognition result for an Old Newspaper Article

For Handwriting text:

The IAM offline dataset was used to train the recognition system, with a train-valid split ratio of 95:5. Therefore, a total of 115320 words were used to train the model. The model was trained for 127 epochs.


--->Kings

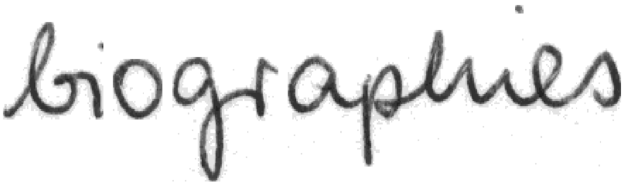

---->biographies

Figure 11. Recognition Result for Handwritten text

Character error rate (CER)	8.62
Word Accuracy	57.41

Table 3. Recognition parameters for Handwritten text

Conclusion and Future Works

This work has potential to further be improved, by using deeper networks and larger datasets in the case of availability of powerful computational systems and hardware, which we did not have access to. Better preprocessing techniques and more powerful CTC decoding algorithms are also aspects of the work that can be focussed on for significant improvements. Future work can focus on extending our current work to regional languages, where there is a clear lack of significant amounts of annotated data. Designing annotation pipelines for regional languages may require more sophisticated segmentation techniques and better recognition systems, and is definitely something that requires much more meticulous research. Even though there are a large number of problems that can be solved by developing systems that aim to digitize or restore documents in regional languages, one of the main limiting factors for extensive research in this field is the dearth of high quality annotated data.

REFERENCES

1.Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk LeeClova AI Research, NAVER Corp.: Character Region Awareness for Text Detection(CRAFT), 2019

2. U. Marti and H. Bunke. The IAM-database: An English Sentence Database for Off-line Handwriting Recognition. Int. Journal on Document Analysis and Recognition, Volume 5, pages 39 - 46, 2002.

3.. Zhou, Xinyu, et al. "EAST: An Efficient and Accurate Scene Text Detector." 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, doi:10.1109/cvpr.2017.283.

4.Vinciarelli, Alessandro, and Juergen Luettin. "A New Normalization Technique for Cursive Handwritten Words." Pattern Recognition Letters, vol. 22, no. 9, 2001, pp. 1043–1050

5.T. Bluche, "Deep neural networks for large vocabulary handwritten text recognition," Ph.D. dissertation, Universite Paris Sud Paris XI, 2015.

Online references:

1.https://towardsdatascience.com/pytorch-scene-text-detection-and-recognition-by-craft-and-a-four-stage-network-ec814d39db05

2.https://stackoverflow.com/questions/59879611/remove-noise-and-staining-in-historical-documents-for-ocr-recognition

3.https://link.springer.com/article/10.1007/s10032-006-0023-z

4.https://towardsdatascience.com/build-a-handwritten-text-recognition-system-using-tensorflow-2326a3487cd5