OCR of Offline Hindi Handwritten Text

B. Tech Major Project

By:-

Raghav Jajoo — 2019KUCP1082

Rishabh Chauhan — 2019KUCP1071

Ayush Kumar Mohanta — 2019KUCP1033

Mentor: Dr. Basant Agarwal

Computer science and Engineering

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY KOTA

MNIT Campus, Jaipur, INDIA.

2021-2022

DECLARATION

I hereby *declare* that the Major Project entitled **"OCR of offline Hindi handwritten text"**, which is being submitted to the ***Indian Institute of Information Technology Kota*** in partial fulfilment of the requirements for the award of the Degree of ***Bachelor of Technology-CSE*** in ***Raghav Jajoo, Ayush Kumar Mohanta and Rishabh Chouhan*** is a *bonafide report of the Project work carried out by me*. The material contained in this thesis has not been submitted to any University or Institution for the award of any degree.

Raghav Jajoo — 2019KUCP1082
Rishabh Chauhan — 2019KUCP1071
Ayush Kumar Mohanta — 2019KUCP1033
Computer Science and Education
Indian Institute of Information Technology
Kota

Place: IIIT Kota, Jaipur

Date: May 12, 2022

CERTIFICATE

This is to *certify* that the Research Thesis entitled **“OCR of offline Hindi handwritten text ”**, submitted by **Raghav Jajoo, Rishabh chouhan and Ayush kumar mohanta** as the record of the research work carried out by him, is *accepted* as the *Research Thesis submission* in partial fulfillment of the requirements for the degree of ***B. Tech***.

Dr Basant Agarwal
Research Guide
Professor
Computer Science and Engineering
IIIT Kota

Abstract

Optical Character Recognition is a system which can perform the translation of images from handwritten or printed form to machine-editable form. OCR converts normal scanned documents text-searchable so to allow content search on the same. Hindi being the national language of India, with such huge population makes document managing and preservation difficult in government sector. Hence, this paper presents an efficient algorithm CRNN and CTC for recognition of Hindi script characters from printed documents. One of the major reasons for the poor recognition rate is error in character segmentation. The presence of touching characters in the scanned documents further complicates the segmentation process, creating a major problem when designing an effective character segmentation technique. Preprocessing, character segmentation, feature extraction, and finally, recognition are the major steps which are followed by a general OCR.

Keywords - OCR, CRNN, CTC, Preprocessing, character segmentation, Feture Extraction, Recognition.

Contents

Chapter 1

Introduction

Optical Character Recognition abbreviated as OCR is the electronic translation of images of handwritten, typewritten or printed text into a machine editable text. An OCR system enables you to take a book or magazine article, feed it directly into an electronic computer file, and then edit the generated text file using a word processor. Thus it can convert the printed characters on the scanned page into editable text. OCR is a field of research which comes under the area of pattern recognition and artificial intelligence. OCR converts printed or handwritten scanned documents into ASCII characters that a computer can recognize.

1.1 Problems in OCR:-

- Non-Latin language advancements have been slow.
- Lack of data
- Different styles and inconsistent way of writing

1.2 Uses of OCR:-

- Authentication of signatures in banks
- Processing of archived institutional records
- Recognizing ZIP code addresses on letters

Chapter 2

Working in Hindi

Devanagari script is the script for writing Hindi language. Hindi is the official language of India. Hindi is spoken in almost all of India. It includes 13 vowels and 36 consonants. Apart from this, it has basic 11 modifiers which are combined with different consonants and vowels.

Each vowel except the first one has a corresponding modifier using which we can modify a consonant. This line which is available on the upper side of a character is called "Shirorekha". Based on this shirorekha each character is divided into three distinct parts. The portion in the upper side of shirorekha is called upper modifiers, in the middle portion the character is available and in the last portion lower modifiers are available.

The major problems with this script which require special attention are:

1. "Shirorekha" or the header line above each and every character.
2. Attachment of modifiers before, after, above, below and within the base vowels and consonants.
3. Large number of symbols
4. Joint, touching and broken characters

Chapter 3

Processes of OCR

The algorithm that is used to develop the OCR software for printed Hindi characters is based on the different geometrical features/shapes of Hindi characters. Input image is parsed into many sub parts/images based on these features. Then other properties such as distribution of points/pixels and edges within each sub image are features used to recognize parsed symbols.

There are mainly four steps performed in any OCR system:-

1. Pre Processing
2. Segmentation
3. Character Recognition
4. Post Processing.

3.1 Pre Processing

The pre-processing phase includes the steps that are necessary to bring the input data into an acceptable form for the further phases. The steps are:

1. RGB to GRAY
2. Binarization
3. Noise removal and smoothing
4. Skew detection and correction

3.1.1 GrayScale Conversion

In the first process, The input is converted in a gray scale image. A gray scale image is an image having BW shades. The image is converted into a grayscale image to further convert it into an BW image. Gray image has shades from black to white.

3.1.2 Binarization

Since the developed system is only able to perform its task only on binarized images, we have to perform the binarization operation before the actual task starts. Basically in binarization, all the pixels above the threshold value are assigned a particular value for instance 1 and all the values below the threshold value are given the value = 0.

There are different types of thresholding for binarization:-

- Simple thresholding.
- Otsu's Binarization.

3.1.3 Smoothing and noise removal:-

Images do have some stray pixels and some unwanted marks. By using filter noise can be filtered from the image. Smoothing operation in gray image is used for noise reduction and filtering is used for noise removal. Basically there are two types of filters, linear filter and order statistics filter. Order statistics filters are non-linear filters whose response is based on the ranking of the pixel and then replacing the value of the center pixel with the value known by ranking result. The best example of a non-linear filter is a median filter.

Median Filtering:-

The median filter is a non-linear digital filtering technique, often used to remove noise from an image. Such noise reduction is a typical pre-processing step to improve the results of the later processing. This is highly effective in removing salt-and-pepper noise.

3.1.4 Skew detection and correction:-

The deviation of the baseline of the text is called skew. During the scanning process, the whole document or a portion of it is fed through the scanner. Our main goal during skewing will be splitting the rotated image into text blocks, and determining the angle from them. We here create two functions, one of determining the skew angle and the other of rotating the image.

What we are trying to do here is to create bounding boxes around the text and then choosing the largest bounding box and then determining the skew angle from the box. The process at first involves finding the areas of text. To make text block detection easier we will invert and maximize the colors of our image, that will be achieved via thresholding.

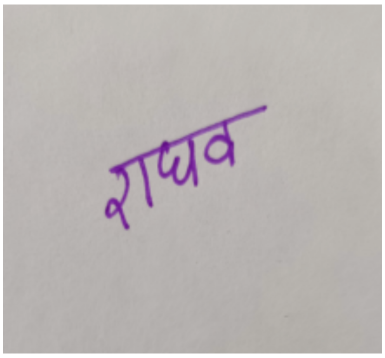


Figure 3.1: Before Pre processing

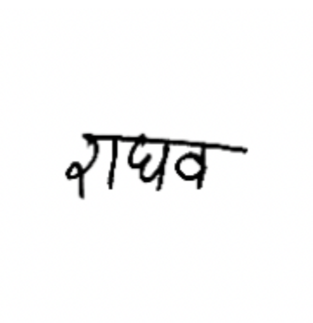


Figure 3.2: After Pre processing

3.2 Segmentation

Segmentation is the way toward parceling a picture/record into disjoint and homogeneous areas. Segmentation is one of the most significant and fundamental procedures that improve the precision pace of character recognition framework. Devanagari report is apportioned into grouping of lines and words by vertical and even projection separately. Process of segmentation involves:-

1. Segmentation of lines
2. Segmentation of words
3. Segmentation of characters

3.2.1 Segmentation of lines

- A horizontal scanning method is applied for segmenting the text paragraphs into lines.
- While performing the segmentation to extract the lines from the text blocks, it performs horizontal scanning starting from the top of the scanned document till it locates the last row containing all white pixels, before a black pixel row is encountered.
- It continues the scanning further, till it locates the first row containing all white pixels, just after the end of the last row of black pixels.
- This determines a line, and is eventually extracted. This whole process is repeated on the entire text page to segment all the text lines present in that particular page/paragraph.

3.2.2 Segmentation of Words

- After segmenting the lines it segments the individual words embedded in each line.
- To perform this operation a vertical scanning method is applied. The vertical scanning is applied to the width of the line only.
- Analysis of this projection will give us a clear idea about the starting and ending column of each character lying within that text line and amount of space between two adjacent characters.

3.2.3 Segmentation of Characters

- A further segmentation process is applied to achieve the individual characters out of the segmented words.
- Before segmenting words at character level, the header line or shirorekha is identified and removed.
- Once the shirorekha is properly removed, the word is divided into three horizontal zones known as upper, middle and lower zones. Individual characters are separated from each zone by applying vertical scanning

3.3 Character Recognition

After the extraction of individual characters occurs, a recognition engine is used to identify the corresponding computer character. Several different recognition techniques are currently available.

3.3.1 K-Nearest Neighbor(KNN) Algorithm for Machine Learning:-

- K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
- The K-NN algorithm assumes the similarity between the new case/data and available cases and puts the new case into the category that is most similar to the available categories.
- KNN algorithm is used for classification (most commonly) and regression. It is a versatile algorithm also used for imputing missing values and resampling datasets.
- There are three limitations of KNN:-
  - When K is greater than one and if numbers of train samples of different classes are the same then there will be a tie for assignment of a specific class.
  - When any input vector (test sample) is assigned to a class, it does not indicate intensity of the vector to that class.

  - All classes are considered with equal strength in assignment of the class label to the test sample.

- Fuzzy KNN:-
  - To avoid the above mentioned disadvantages of KNN algorithm, fuzzy set concept is introduced into it.
  - The Fuzzy K-Nearest Neighbor algorithm assigns class membership to a test sample rather than defining specific class.

3.3.2 Convolutional Recurrent Neural Network:-

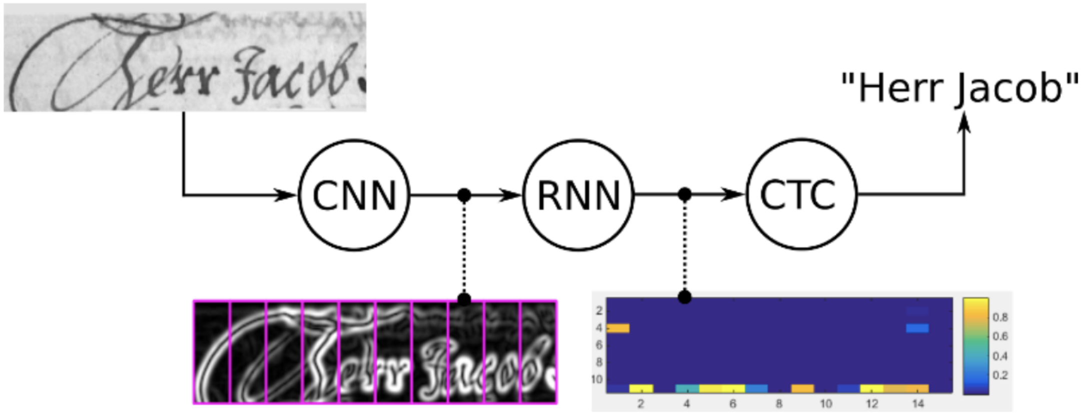
- The Convolutional Recurrent Neural Networks is the combination of two of the most prominent neural networks. The CRNN (convolutional recurrent neural network) involves CNN(convolutional neural network) followed by the RNN(Recurrent neural networks).
- We use CRNN to extract the important features from the handwritten line text Image.
- Most of the time, the Convolutional Neural Network analyzes the image, sending it to the recurrent part of the important features detected.
- The recurrent part analyzes these features in order, taking into consideration previous information in order to realize what are some important links between these features that influence the output.

3.3.3 Connectionist Temporal Classification (CTC):-

The NN outputs character-scores for each sequence-element, which simply is represented by a matrix. Now, there are two things we want to do with this matrix:

- Train: calculate the loss value to train the NN
- Infer: decode the matrix to get the text contained in the input image

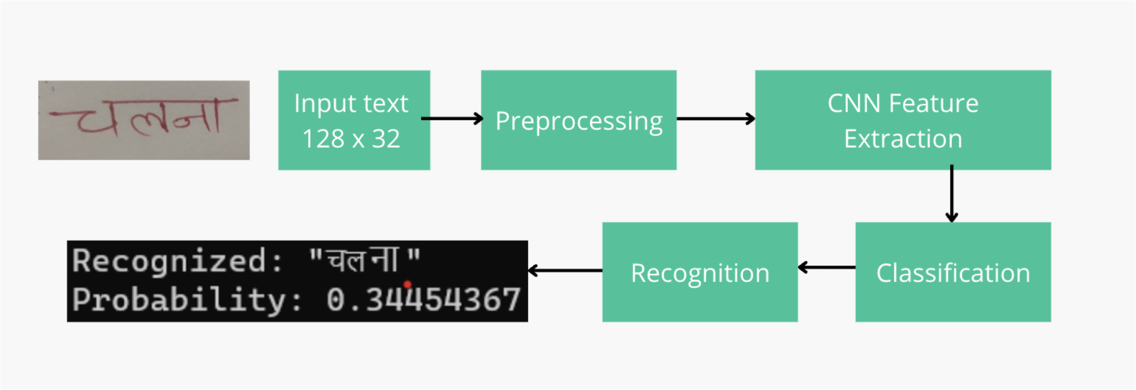
Both tasks are achieved by the CTC operation.



The NN-training will be guided by the CTC loss function. We only feed the output matrix of the NN and the corresponding ground-truth (GT) text to the CTC loss function.

3.4 Post Processing

Post-handling stage is the last phase of the proposed recognition framework. The post processing phase includes the conversion of the UNICODE in to standard output into any standard text encoding scheme.



```
Recognized: "चलना"
Probability: 0.34454367
```

Figure 3.3: The Complete process of an OCR

3.5 Bibliography

1. D. Yadav, S. Sanchez-Cuadrado, and J. Morato, “Optical Character Recognition for Hindi Language Using a Neural-network Approach,” Journal of Information Processing Systems, vol. 9, no. 1, pp. 117–140, Mar. 2013.

2. K. Dutta, P. Krishnan, M. Mathew and C. V. Jawahar, "Towards Spotting and Recognition of Handwritten Words in Indic Scripts," 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR), 2018, pp. 32-37, doi: 10.1109/ICFHR-2018.2018.00015.

3. Nafiz Arica , Student Member , Fatos T. Yarman-vural , Senior Member, "Optical Character Recognition for Cursive Handwriting", IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 24, NO. 6, JUNE 2002

4. Gunna, S., Saluja, R., Jawahar, C.V. (2021). Transfer Learning for Scene Text Recognition in Indian Languages. In: Barney Smith, E.H., Pal, U. (eds) Document Analysis and Recognition – ICDAR 2021 Workshops. ICDAR 2021. Lecture Notes in Computer Science(), vol 12916. Springer, Cham.

5. Abhishek Mehta, Dr. Subhashchdra Desai, Dr. Ashish Chaturvedi, 2021, Hindi Handwritten Character Recognition from Digital Image using Deep Learning Neural Network, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH TECHNOLOGY (IJERT) ICRADL – 2021 (Volume 09 – Issue 05)

6. Nisha Goyal1 , Er. Shilpa Jain, "Optimized Hindi Script Recognition using OCR Feature Extraction Technique", International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 8, August 2015.

7. Bairagi, Prasanta Pratim. “Optical Character Recognition for Hindi.” (2018).