OCR for Hindi Handwritten Text

Abstract

Optical Character Recognition (OCR) is one of the most important applications of Computer Science. The accuracy of Character Recognition in the case of printed characters has improved , But same is not in the case for Indian languages which are complicated in terms of structure and computations ,especially in case of Hindi Handwritten documents researchers are still working on accuracy for character Recognition .The task to implement the same is very Complexas Handwriting of any individual varies from person to personand even in the same document a single character or word can be represented differently by the same writer, which leads to less accurateresults.

The building block of the complete task is Segmentation and it plays very important role. To achieve this, one has to efficiently handle problems like unequal spacing between text lines, over- lapped text lines, connected components between the lines, unequal height of characters, separation of upper and lower modifiers.

As we want to make our OCR more accruable and effective by adding scan copy into searchable, editable and giving the command/feature in form of handwritten textand then it will get converted into typed text and after that it is in dictation way that is an assistive technology (AT),It helps those people who struggle in writing . You may hear it referred to as “speech-to-text,” “voice-to-text,” “voice recognition” or “speech recognition” technology as the text will get converted into speech .we want to make this bot more compatible ,more efficient and more optimized.

A few classification techniques have been talked about in this paper, such as Convocational neural network.

List OF ALGORITHMS

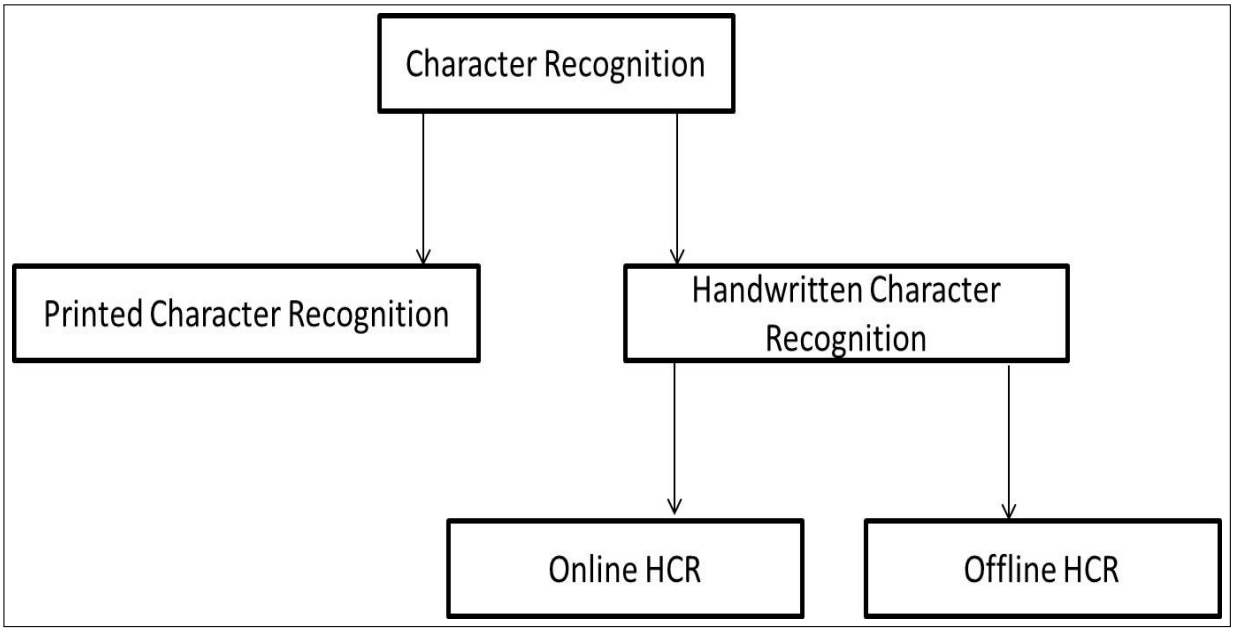
Introduction

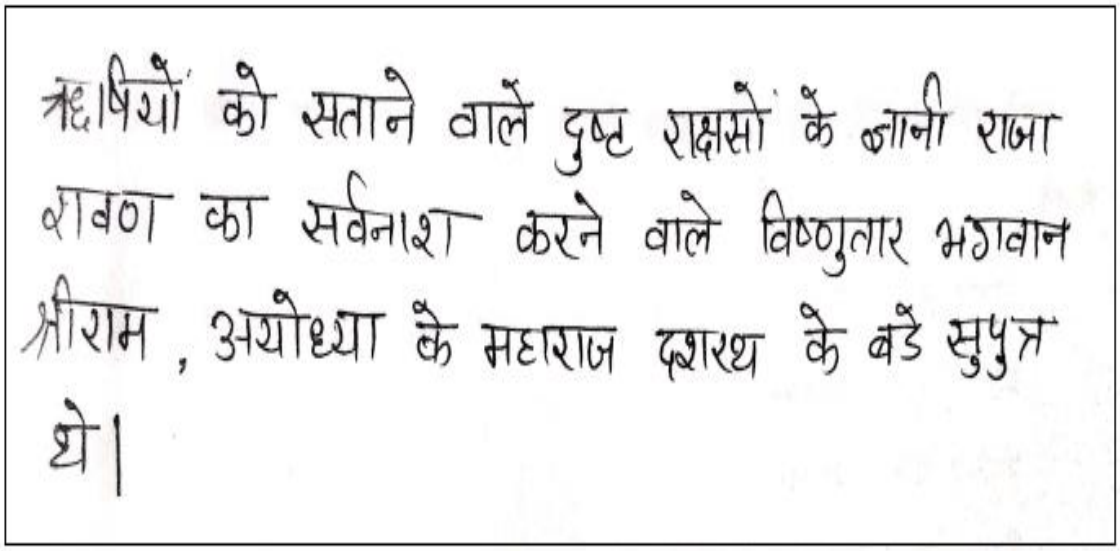
OpticalCharacterRecognition(OCR) or Optical Character Reader is the electronic or mechanical conversion of images of typed, handwritten or printed text(letters,numerals,andsymbols) into machine-encoded text(such as ASCII), whether from a scanned document, a photo of a document a scene photo or from subtitle text superimposed on an image.

OCR plays an important role for digitization of paper documents. Nowadays, the process of digitization from handwritten documents into computer editable form is going on in sectors like government offices, libraries, banks, colleges, *etc.* So the OCR system is developed to convert the scanned documents into computer editable format .Major application of OCR includes

(i) Signature verification and identification
(ii) Reading bank deposit slips
(iii) Automatic number-plate reader
(iv) Make electronic images of printed documents search-able.

Optical Character Recognition is broadly categories into two sub-fields as shown in (i) Printed Character Recognition (ii)Handwritten Character Recognition.





Sample of Hindi Handwritten Text

Application OCR

---

1. **Data identification**
2. **Readingaddresses**
3. **Readingpassportdata**
4. **Automatic number plate Recognition**
5. **BillProcessingReaders**
6. **Legal**
7. **Health care**
8. **Page Readers**
9. **QuickDigitalSearch**

Introduction of Devanagari Script

After English and Chinese, Hindi is the third most widely used language and all over the world, there are approximately 500 billion people who write and speak Hindi. In India, Devanagari is the basic script of many languages like Sanskrit and Hindi. Hindi language have 11 swar (vowels) , 33 Vyanjan (consonants) and 12 modifiers (special symbols) .Almost all the words are combination of these two along with the modifiers placed on left, right, bottom and above. It is written from left to right.

combination of these two along with the modifiers placed on left, right, bottom and above. It is written from left to right.

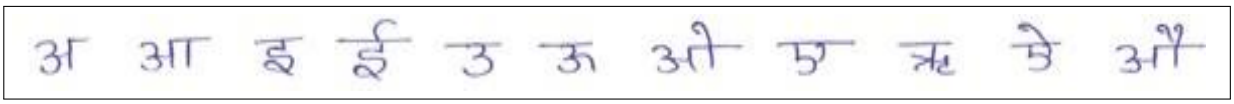


Figure: Swar in Hindi language

Figure: Vyanjan in Hindi language

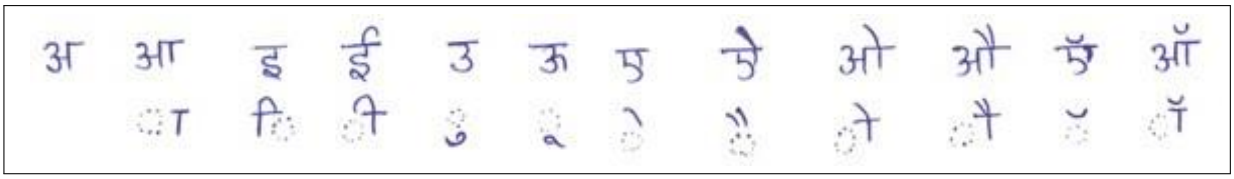
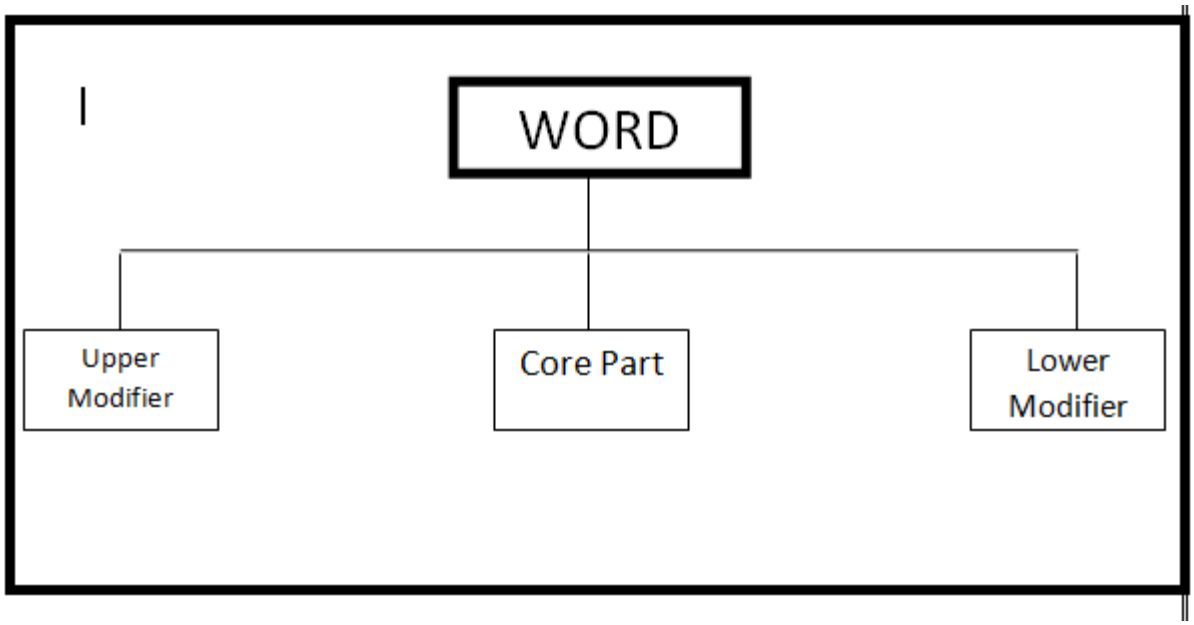


Figure: Modifiers in Hindi language



Stages in OCRSystem

The OCR system accepts scanned images as input. The quality of the input image isdetermined by measuring the number of DPI (dots per inch). **Image quality** is **directly proportional** to **DPI**. We scanned the images at **300 DPI** to digitize documents for ourexperiment because 300 DPI produces sufficient image quality for OCR. If the resolution is too high, thenitwilltakemoretimeforprocessingandrequiresmorememoryspace.Recognitionof devanagari script includes the following steps as shown in.

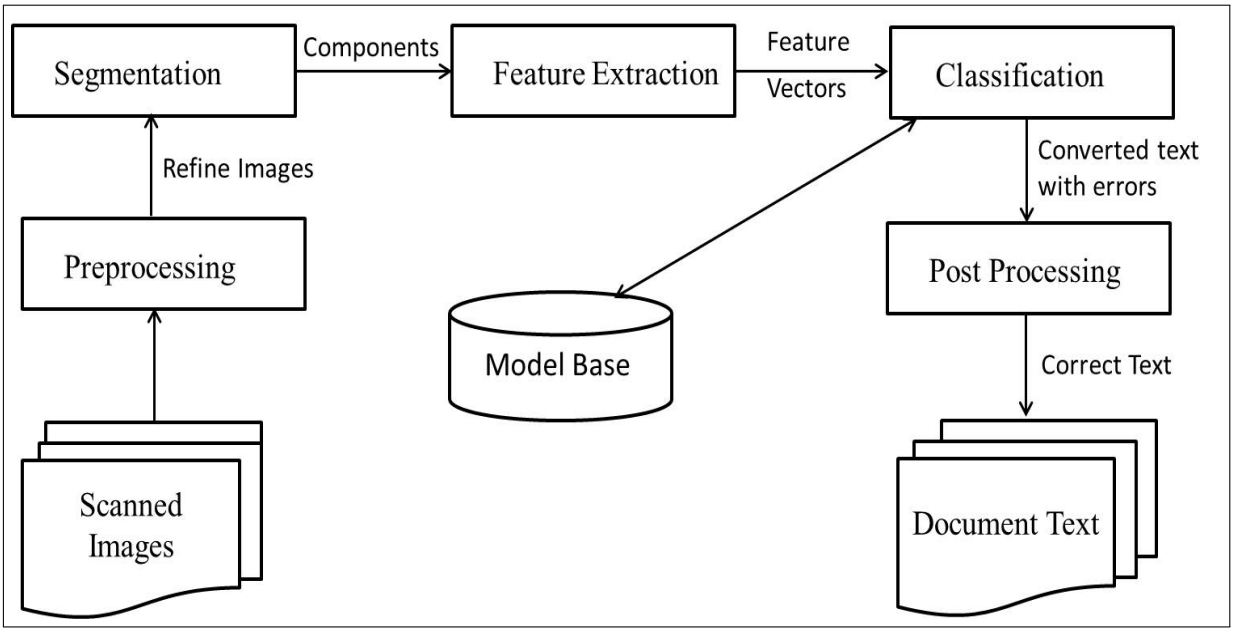


Figure: Steps in recognition of Hindi Documents

Libraries

1. Pandas to read and write Tables
2. Numpy for effective computations
3. Matplotlib for data visualization.
4. Sk-learn
5. Pytsxx3(text to speech)
6. Gingerit(For auto-correction)
7. Tensorflow (For build model)

Reference video link:-

https://youtu.be/ZgeFFBfaJ7c