methodology

Segmentation of devanagri script

After binarization of image we have to isolate individual letter from an image so that we can recognize it with the help of classifier. Segmentation includes three four processes,

1)Separation of word from the sentence

2) Isolation of individual character(with lower or upper connected modifier if any) from the word

3)Removal of connected lower or upper modifiers

4)If the letter is conjugate letter(Jodakshara),we have to separate letter into its participating letter (e.g. separation of mya into m,ya)

In devanagri script all words are connected by the shirolekaha, so if the document if noise free, number of connected components will be equal to number of words

For the separation of words, we use vertical projection approach for separation of letters. We gate of row matrix consisting of sum of all columns ,Then we find minimum value of sum from the matrix, the columns whose sum is minimum are

columns between two letters e.g.

consider the following word:

C:\Users\Soman\Desktop\Character recognition\ty_mini_project\13.tiff

(white pixels represented by one and black pixels are represented by zero in binary image)

the sum of columns between n(na) and its kana(a) is minimum by using this approach we can segment above word like

C:\Users\Soman\Desktop\Character recognition\ty_mini_project\1.tiff

now connected components in the image will be the letter with connected modifier(like ri)

Now by removing shirolekha(upper line) we can get individual isolated letters and modifiers

Using this approach we can track about presence of modifier for the letter which can be useful in recognition process

Separation of lower modifier is difficult because there is no feature like shirolekha to separate lower modifier from letter.

For separation of lower modifiers we calculate minimum height of letter in the word, that letter will be the letter having no lower modifiers, If height of any letter in word is considerably greater than minimum height ,we separate part below that height and treat it as a lower modifier

normal

height

C:\Users\Soman\Desktop\Character recognition\ty_mini_project\segmented_images\1.2.1.tiffC:\Users\Soman\Desktop\Character recognition\ty_mini_project\26.tiff

letter with lower modifier

Segmentation of jodakshara

In devanagri script conjugate charactors or jodakshara formed when two or more consonant combine togather.

Examples of jodakshara are(mya,@t,$)

For segmentation of jodakshara we use vertical projection approach

We form row matrix by taking sum of all columns in a image, When two consonant combine to form jodakshara the sum of column where first constituent consonant ends and second starts is minimum

D:\laptop-data\Desktop\Character recognition\ty_mini_project _upadte\jodakshara\1.2.1.tiff

Sum of this column is minimum

II)Pre processing of individual characters

Preprocessing includes thinning of image so that stroke width of image is one pixel for that we using standard skelatanization of image and morphological operations by converting image into its thin form we will able to calculate features related to number of pixels in image(statistical features) Independent of font size

II)Feature extraction

In optical character recognition it is important to find feature vector associated with input symbol so that we can recognize character in input symbol with the help of classifier.

We use three features

1)Shadow with zoning

We divide input image into 16 parts, then find length of horizontal and vertical projection of each part

2)Vertical line

We detect presence and position of vertical line in an image for example, vertical line is present at middle for k, f ,at end for pa,sa and absent for d,D

3) Eular number:

Eular number is total number of objects minus total number of holes in the image

Future Scope:

1)ukar segmentation can be improved by considering minimum height of character from all document

2) Non segmentable conjugate characters like $,W should be treated like individual characters and added in training set

3)Segmentation algorithm should be designed for jodakhara involving three constituent characters

4)Algorithm should be designed for segmentation and recognition of punctuation marks and numerals