Hindi Optical Character Recognition For Printed Documents Using Fuzzy K-Nearest Neighbor Algorithm: A Problem Approach In Character Segmentation

Prof. Amit Choksi[1], Kajal Kumari[2], Shivani Kanojiya[3], Pragya Sahu[4], Nishtha Rindani[5]

[1](EC Department, BVM Engineering College, V.V.Nagar, Gujarat,India)
[2](EC Department, BVM Engineering College, V.V.Nagar, Gujarat,India)
[3](EC Department, BVM Engineering College, V.V.Nagar, Gujarat,India)
[4](EC Department, BVM Engineering College, V.V.Nagar, Gujarat,India)
[5](EC Department, BVM Engineering College, V.V.Nagar, Gujarat,India)
Corresponding Author: Prof. Amit Choksi

***Abstract :** Optical Character Recognition (OCR) is a technology that extracts all the text from the images, .pdf documents or scanned files. So 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 Fuzzy KNN for recognition of Hindi script characters from printed documents. Optical Character Recognition (OCR) systems developed for the Hindi language carry a very poor recognition rate due to shirorekha as well as joint characters. This paper proposes an OCR for printed Hindi text in Devanagari script, using Fuzzy KNN which improves its efficiency. One of the major reasons for the poor recognition rate is error in character segmentation also. 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. Here, Fuzzy KNN classifier in pair with two different features Geometric and Wavelet features are used to handle this problem.*

***Keywords –** Optical Character Recognition, Fuzz-KNN, Wavelet Transform*

I. Introduction

TO OCR 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 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 in to editable text. OCR is a field of research which comes under the area of pattern recognition and artificial intelligence. The challenges in the task can be realized by knowing the fact that there are thousands of fonts available for most of the scripts and the text typed with any of these fonts may be in an of regular bold or italics styles and of various sizes. This results into large number of possible variants. As a result of this, earlier OCR systems were dependent on a number of factors including the font style, size and orientation. There are mainly four steps performed in any OCR system. The block diagram of OCR system is shown in figure below [2][11].1) Pre Processing 2) Segmentation 3) Recognition 4) Post 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 GRAY2) Binarization 3) Noise removal and smoothing 4) Skew detection and correction 5) Character normalization.

The segmentation phase includes two steps:
1) Line segmentation 2) Character segmentation

In recognition phase each character in the document is recognized. For example on of the recognition technique is called *template matching*, has been used wherein each character in the input image as seen in OCR is compared against a set of templates and the UNICODE of the template that matches the best output. Classification and feature extraction is done in this phase. The post processing phase includes the conversion of the UNICODE in to standard output into any standard text encoding scheme [1].

System Block Diagram
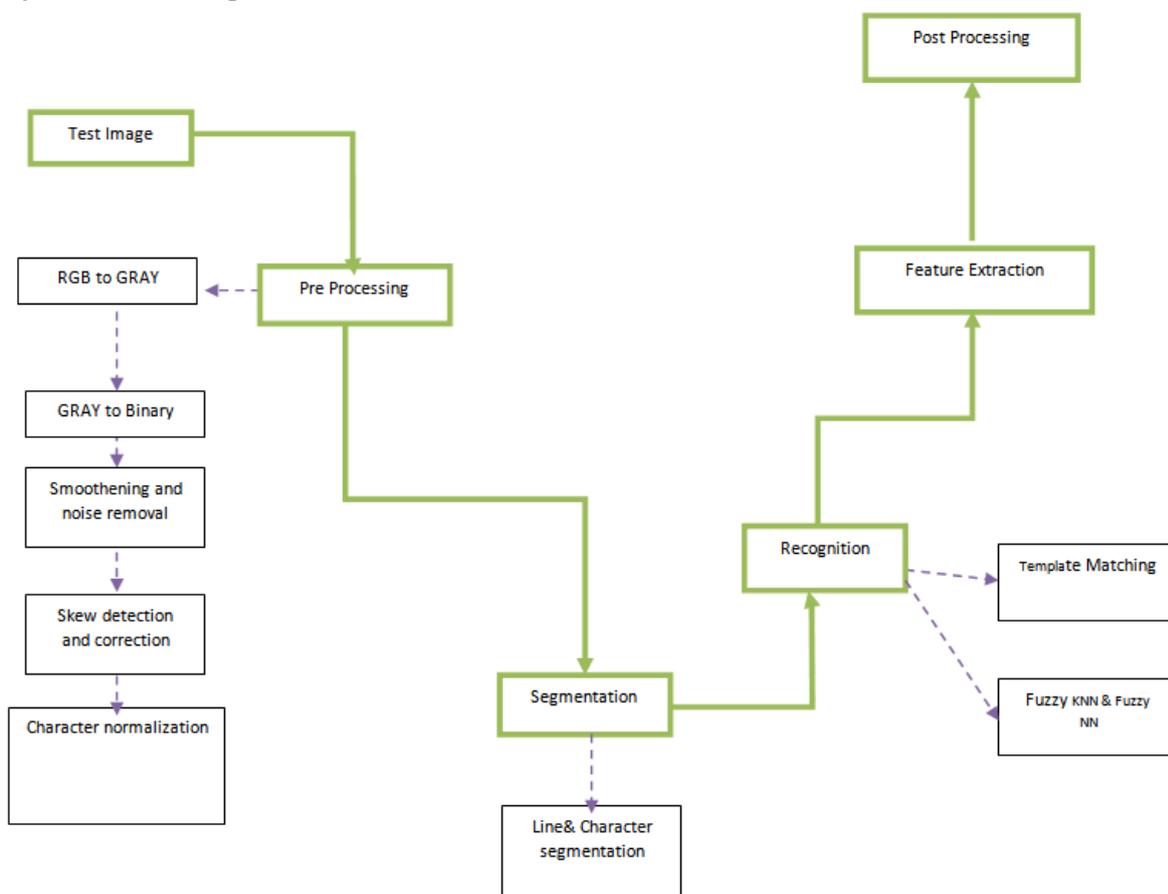
Introduction To Hindi Script

Hindi is spoken in almost all of India. It includes 12 vowels and 34 consonants. Apart from this, it has basic 11 modifiers which are combined with different consonants and vowels. There appear before, after and below the consonant or vowel. They are similar to those of Gujarati language.

क ख ग घ च छ ज ट ठ ड ढ ऑ ो ौ ि ी ी ु ू े ै फि भी बि मौ खू

Fig.1 Consonants, **Vowels and Modifiers of Hindi Script**

Same as in other languages, Hindi script characters also have their own unicodes. Figure below shows some Hindi characters along with their unicodes.

ब—092C क—0915

Fig.2 Unicode of Hindi character

1.1 Challenges in recognition of Hindi script

Unlike English and Gujarati, Hindi poses many challenges as far as development of OCR technology is concerned. Like most of the Indian scripts, it is difficult and more complex to recognize Hindi characters than any other Latin base scripts.

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

As this project also includes the recognition of handwritten documents, handwriting of different persons may vary in size, font, curves, header line and more

The figure above shows complete block diagram of Hindi OCR system. The system performs the following steps described in detail.

II. Creation Of Database

Initially a database is created for 260 Hindi characters including consonants, special characters, modifiers and digits from 0 to 9. The database fonts are random and not of some similar font. Greater the number of database characters, greater is the system efficiency.

2.1 Pre Processing Techniques

The Pre-processing step is very essential step in this technique of image processing field [6] [20]. In some cases, the original data or image is of poor quality due to blurred image. Pre-processing Phase is concern with reduction of noise in the input image. Pre-processing is concern with the reduction of noise and variability in the input [17]. Some of the common operation performed prior to recognition are; thresholding, binarization, Noise removal etc.

2.2.1 Gray Scale Conversion

Here, it is required to covert RGB image into Gray image for further converting into binary image. If image is not in a gray form then it is important to converts image into gray form. In gray scale conversion the image will be comprised as black at weakest intensity and white at strongest intensity and there will be many shades in between. It replaces every pixel of image after calculation of gray conversion into new required gray scale pixel value. If gray level is done at 8 bit then it will give 256 shades. Here, gray scale image is having value from 0 to 255 pixel value.

2.2.2 Binarization

Normally pixel intensity values of an image are in the range of 0 to 255. Binarization is process which converts gray image into binary image. Binary is often produced by thresholding a Gray scale image. Most of the time the goal of it is that to separate an object in image from its background. To perform binarization, to find threshold value for particular image is necessary.

The task of **thresholding** is to extract the fore ground from the background. A number of thresholding techniques have been previously proposed using global and local techniques.

The **histogram** of gray scale values of a document image typically consists of two picks: A high pick corresponding to the white background and a smaller peak corresponding to the foreground.

Hence, threshold gray scale value can be determined by an optimal value in the valley between the pick.

Here Otsu's method is used for binarization [5]

2.2.3 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.

2.2.3.1 Order Statistics Filter

Order statistics filter are non linear filter whose response is based on the ranking of the pixel and then replacing the value of centre pixel with the value known by ranking result. In Median Filter which is the best example of non linear filter, replaces the value of the median of the gray levels in the neighborhood of that pixel. Median filter are popular because they provide excellent noise removal capabilities with less blurring of the pixel. Median filters are particularly effective in the presence of impulse noise, also called salt and paper noise, because of its appearance as white and black dots superimposed on an image.

2.2.4 Skew detection and correction

The deviation of the base line of the text is called skew [12]. During the scanning process, the whole document or a portion of it is fed through scanner. The digital image of the document may be skewed arbitrarily because of how it was places on the platen when it was scanned or because of a document feeder malfunction. However, skew is unintentional in many real cases and it should be eliminated because it dramatically reduces the accuracy of the subsequent process such as page segmentation and OCR. Most of the OCR and document retrieval are very sensitive to skew in document images. Hence it is important to correct the skew.

There are several algorithms for skew detection mentioned as: 1) Projection profile 2) Hough transforms technique 3) Fourier method 4) Nearest neighbor clustering 5) Correlation

2.2.4.1 Skew detection using projection profile

A straight forward method to determine the skew angle of a document is the horizontal projection profile. This is a one-dimensional array with a number of locations equal to the number of rows in an image. Each location in the projection profile stores a count of the number of black pixels in the corresponding row of

the image. This histogram has the maximum amplitude and frequency when the text in the image is skewed to zero degrees since the number of co-linear black pixels in maximized in this condition. Histogram of any image represents the number of pixels in different shades.

2.3 Segmentation

There are only fifty two possible character symbols. Since there is always some space between characters of a word, a general strategy for handling such scripts would be to segment a word into individual characters and then recognize each character separately.

It is required to group the lines, words and characters in proper order [16], we have to go for segmentation part. Segmentation phase is an important phase and accuracy of any OCR heavily depends upon segmentation phase. Incorrect segmentation leads to incorrect recognition.

Here we have performed two types of segmentation:
1) Line segmentation
2) Word Segmentation and Character segmentation

2.3.1 Line Segmentation

The image is segmented into the lines based on the information provided by the procedures. Here horizontal projection profile technique is used for line segmentation. The digitize image is processed to line and words using Horizontal Projection Profile.

2.3.2 Word and Character Segmentation

The lines are then segmented into characters and given to the classifier to recognize that particular character. For Hindi characters, vertical projection profile approach alone will not give the desired output as the characters in Hindi are composed by the attaching glyph of a consonant, modifier, vowel and the header line.

So here the character segmentation is done using two methods
1) Vertical projection profile
2) Combination of connected component labeling and vertical projection profile.

2.3.2.1 Vertical projection profile

Like horizontal projection, even in vertical projection profile we shall be able to gather the information about black pixels. Unlike horizontal histogram, in case of the vertical histogram the projection will be taken vertically. Vertical histogram has to be taken for each line one by one, so line starting and ending data has to be given precisely which are available through horizontal histogram analysis. Analysis of this projection will give us a clear idea about starting and ending column of each character lying within that text line and amount of space between two adjacent characters.

2.3.2.2 Connected Component Algorithm

CCs are generally considered in binary images. Two pixels are said to be 8-connected if they are connected by a chain of 8-connected pixels. A CC is a set of pixels in which each pixel is connected to the rest. Touching characters have stroke pixels in a common CC should be then split at the points of touching. By CC labeling, the pixels of different components are stored in different sets of labeled with different pixel values. There have been many effective algorithms for labeling CCs which can be roughly divided into two categories: raster scan and contour tracing. By raster scan, all the CCs can be found in two passes or a forward scan with local backtracking. By contour tracing the pixels enclosed by different contours belong to different CCs.

2.4 Feature Extraction

OCR systems extensively use the methodologies of recognition which assigns an unknown sample into a pre-defined class. Numerous techniques for OCR can be investigated in four general approaches of recognition, as suggested:
1) Template Matching 2) Fuzzy KNN Technique 3) Neural Network

2.4.1 Template Matching

OCR techniques vary widely according to the feature set selected from the long list of features, described in the previous section for image representation. Features can be as simple as gray-level image frames with individual characters or words or as complicated as graph representation of character primitives. The simplest was of character recognition is based on matching the stored prototypes against the character or word to be recognized [2]. Generally speaking matching operation determines the degree of similarity between two vectors in feature space.

2.4.1.1 Direct matching

A gray-level or binary input character is directly compared to a standard set of stored prototypes. According to a similarity measure, a prototype matching is done for recognition. The matching techniques can be as simple as one to one comparison or as complex as decision tree analysis in which only selected pixels are tested. A template matcher can combine multiple information sources including match strength and k-nearest neighbor measurements from different metrics. Although direct matching method is intuitive and has a solid mathematical background, the recognition rate of this method is very sensitive to noise.

2.4.2 KNN Algorithm

In KNN algorithm [23] [24]Euclidean distance between Test samples and correctly identified samples set is measured. Label to test sample is given a class based on K-closest Neighbors. Euclidean distance between two samples a1= (a1, 1 ...a1, k) and a2= (a2, 1 ...a2, k) is,

$$d(a_1, a_2) = \sqrt{\sum_{j=1}^{k} \left(a_{1,j} - a_{2,j}\right)^2}$$

(1)
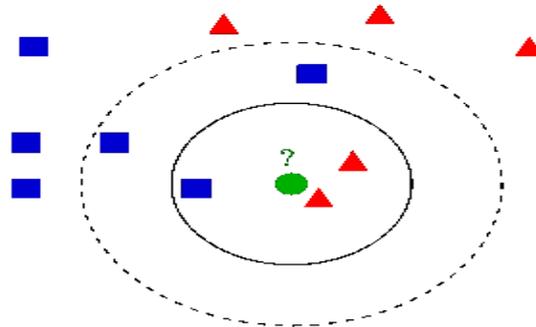
As shown in figure, the test sample is circle. First class is squares and second class is of triangles. If closest neighbors considered are three (*K=3*) then circle is classified to class of triangle. If *K = 5* then it is classified to square. KNN has advantages of being Nonparametric and works well on small sample size. There are three limitations of KNN: 1) when K is greater than one and if numbers of train samples of different classes are same than there will tie for assignment of a specific class. 2) When any input vector (test sample) is assigned to class, it does not indicate intensity of vector to that class. 3) All class is considered with equal strength in assignment of the class label to test sample.

2.4.3 Fuzzy KNN

To avoid above mentioned disadvantages of KNN algorithm fuzzy set concept is introduced into it. Fuzzy set was used by Zadeh in 1965. A "Fuzzy KNN" algorithm [26] utilizes strength of test sample into any class called fuzzy class membership and thus produces fuzzy classification rule. An example of Fuzzy set is the set of real numbers much larger than zero, which can be defined with a membership function as follows:

$$u(x) = \begin{cases} x^2/(x^2 + 1), & x \geq 0 \\ 0, & x < 0 \end{cases}$$

Numbers less than zero are not in the set because value of membership function for those is zero. While numbers larger than zero are in the set based on strength of numbers with respect to zero. This makes Fuzzy Set a useful tool for classification of samples having imprecise boundary. Fuzzy Set gives degree of presence of any sample into specific class. As mentioned in, use of Fuzzy Set in KNN improves its classification results for pattern recognition. Thus, Fuzzy KNN can be a useful algorithm to deal classification of similar appearing symbols mentioned in table 1. Suppose, input sample set is {*x1,…………..,xn*} is to partitioned into *C* classes. Then it is required to find fuzzy membership value of each input vector into *C* classes given by *uik = ui(xk) for i = 1,.....,C* and *k = 1,.......,n.* Here *uik* is the degree of membership of kth sample into classes *i*. It is denoted by the *C* by *n* matrix on U. The following properties must be true for *U* to be a fuzzy *C* partition.

$$\sum_{i=1}^{c} u_{ik} = 1$$

$$0 < \sum_{i=1}^{m} u_{ik} < n$$

$$u_{ik} \in [0,1]$$

The Fuzzy K-Nearest Neighbor algorithm assigns class membership to a test sample rather than defining specific class. For example, an input character sample is assigned a membership of 0.8 to one class and 0.2 to other class than that sample belongs to class one than the other. But if a vector is assigned membership value of 0.55 to one class and 0.45 to other class than input vector must be examined further for assignment of a class. $ui(x)$ can be found by

$$u_i(x) = \frac{\sum_{j=1}^{\kappa} u_{ij}\, (1/\|\,x - x_j\,\|^{2/(m-1)})}{\sum_{j=1}^{\kappa} (1/\|\,x - x_j\,\|^{2/(m-1)})}$$

As shown in equation membership value depends on the inverse of the distances from the nearest neighbors and their class membership. Because of inverse distance input sample's membership is more if it is near and less if it is farther from the sample vector into consideration. In equation the variable m determines how heavily the distance is weighted when calculating each neighbor's contribution to the membership value. If m is two, then contribution of each neighboring point is weighted by reciprocal of its distance from the point being classified. As m increases, the neighbors are more evenly weighted, and their relatives' distances from the point being classified have less effect. As m approaches one, the closer neighbors are weighted far more heavily than those farther away, which have the effect of reducing the number of points that contribute to the membership value of point being classified.
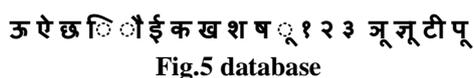
III. Results

3.1 Create database

ऊ ऐ छ ि ौ ई क ख श ष ू १ २ ३ ऋ ू टी पू
Fig.5 database
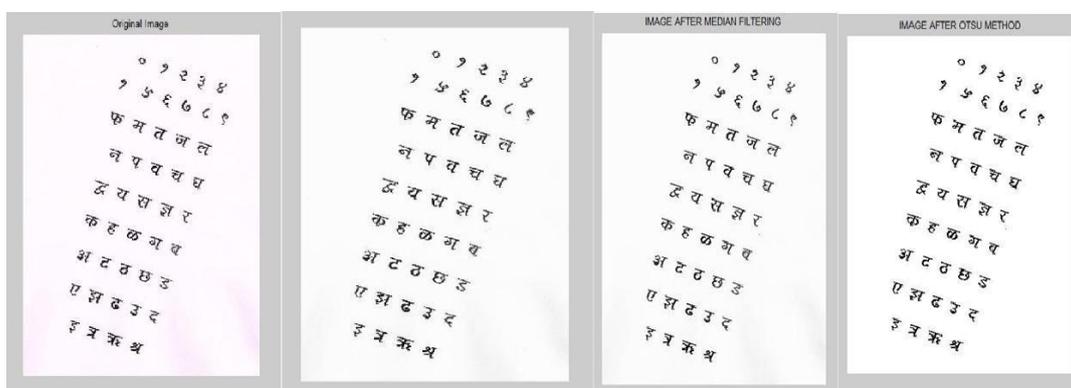
3.2 Pre-processing Techniques

Fig.3 Pre-processing techniques

3.3 Skew detection and correction using projection profile method
Here we have taken non skewed image and then we have manually skewed it and we have checked that this algorithm is also working properly when skew angle is unknown to us.
- Define the amount of range (-5 to +5)
- Define the step size (0.05, 0.10, 0.15)
- Add step size to amount, rotate the image by it and get projection profile image.
- Repeat the step 3 until all the amount is calculated for projection profile.
- Get the variances at particular amount in projection profile. Find the maximum variance at one of the angle, is called skew angle.
- Skew correction is done by rotating the image by above obtained skew angle.
- By applying above algorithm we obtain skew corrected image. The corrected image and its projection profile are show below.
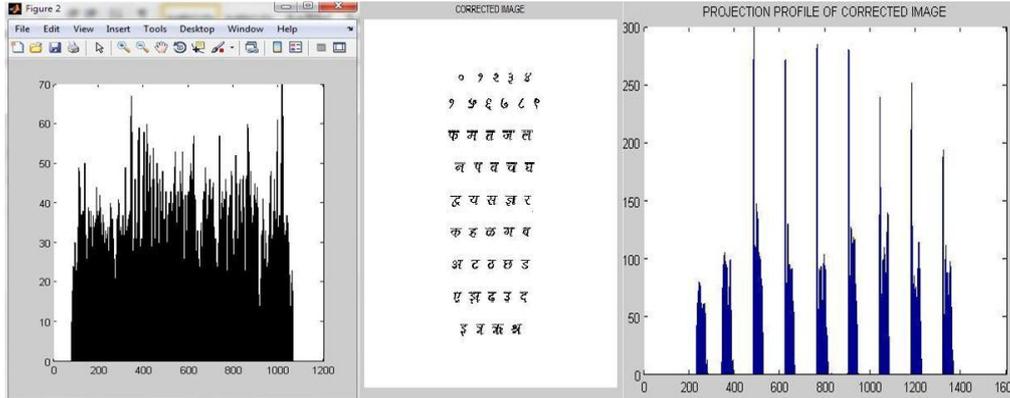
Fig.4 horizontal projection profile of skewed image and corrected image.

3.4 Segmentation
3.4.1 Line Segmentation
- Call the skew corrected image in the line segmentation function.
- Find the indices and the values of non-zero elements of image.
- Find minimum and maximum element of row.
- Find minimum and maximum element of column.
- Cropped image = image (minr:maxr, minc:maxc).
- Complement the cropped image
- Compute the horizontal projection profile of cropped image.
- Crop the portion of first line from the horizontal projection profile.
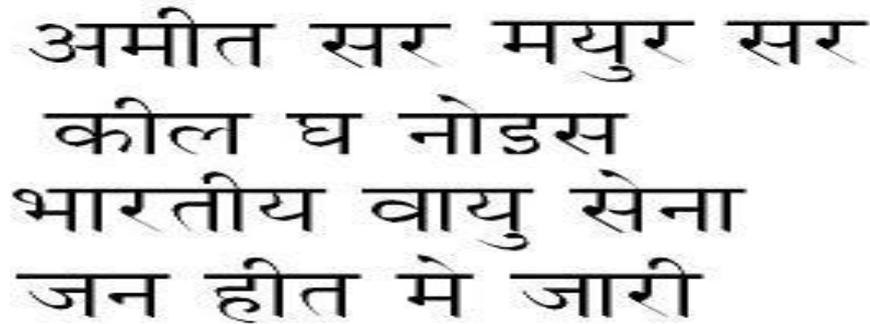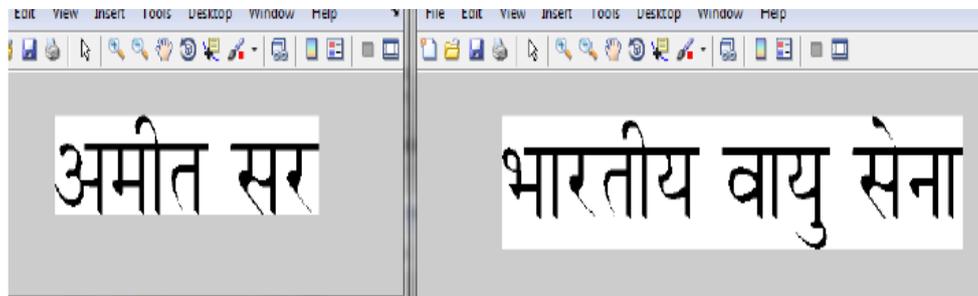- Repeat above step until portion of all lines are cropped.

Fig.6 Line Segmentation

Fig.7 Line Segmentation using Horizontal Projection Profile

3.4.2 Word and Character Segmentation
- Call the line image in the character segmentation function.
- Find the indices and values of non-zero elements of line image.
- Find minimum and maximum element of row.
- Find minimum and maximum element of column.
- Cropped line image = line image.

- Complement the cropped line image.
- Compute the vertical projection profile of cropped line image.
- Crop the portion of first character from the vertical projection profile.
- Repeat until portion of all characters are cropped.

3.4.2.1 Vertical projection profile
Pseudo code:
For (j = start col to end col)
{
Column blacks = 0;
For (I = start row to end row)
{
If (binarized [i][j]) then increment the column blacks counter by one.
Verticals[j] = column blacks}
Return an array vertical [] which contains no. of black pixels in each column }}

Fig.8 Line Segmentation

Fig.9 Character Segmentation using Vertical Profile Projection

Problem during character segmentation is Overlapping Characters. For example this kind of problem has been solved using CCA as mentioned above.

3.5.1 Feature extraction and selection
The heart of any optical character recognition system is the formation of feature vector to be used in the recognition stage. Feature extraction can be considered as finding a set of parameters (features) that define the shape of the underlying character as precisely and uniquely as possible. The term *feature selection* refers to algorithms that select the best subset of the input feature set. Methods that create new features based on transformations, or combination of original features are called *feature extraction* algorithms. The features are to be selected in such a way that they help in discriminating between characters. Features that capture topological and geometrical shape information are the most desired ones. Features that capture the spatial distribution of the black (text) pixels are also very important.

Wavelet Features
A wavelet is a wave-like oscillation with amplitude that starts out at zero, increases, and then decreases back to zero. Wavelets are purposefully crafted to have specific properties that make them useful for digital image processing. A wavelet transform is the representation of a function by wavelets. Wavelet transforms have

advantages over traditional Fourier Transforms for representing functions that have discontinuities and sharp peaks, and for accurately deconstructing and reconstructing finite, non-periodic and non-stationary signals. Wavelets are very good in time and frequency localization, they are used by many people as feature extractor [4], [5]. Families of wavelet base vectors by Daubechies are very well localized in space and frequency. In [6] authors have shown Daubechies D-4 wavelet which gives better results for handwritten Chinese character recognition than the other wavelets. It is been also observed that D-4 wavelets gives better results than other statistical methods used for feature extraction.

3.5.2 Template Matching

This is the simplest way of character recognition, based on matching the stored prototypes against the character or word to be recognized. The matching operation determines the degree of similarity between two vectors (group of pixels, shapes, curvature etc.) A gray-level or binary input character is compared to a standard set of stored prototypes. A template matcher can combine multiple information sources, including match strength and k-nearest neighbor measurements from different metrics. The recognition rate of this method is very sensitive to noise and image deformation. For improved classification Deformable Templates and Elastic Matching are uses.

Table 1:- Accuracy efficiency using Template Matching

Image name	Total Characters	Correct Recognized Characters	Incorrect Recognized Character	Efficiency
Image01	44	11	33	25%
Image02	34	24	10	70%
Image 03	34	26	8	76%
Image04	57	9	48	15%
Image 05	44	17	27	38%

Table 2:- Accuracy and efficiency using KNN

Image name	Total Characters	Correct Recognized Characters	Incorrect Recognized Character	Efficiency
Image01	44	42	2	95%
Image02	34	31	3	90%
Image 03	34	26	8	76%
Image04	57	53	4	92%
Image 05	44	39	5	90%

Table 3:- Fuzzy KNN is a Better Approach Than Template Matching And Nearest Neighbor

Image name	Total Characters	Correct Recognized Characters	Incorrect Recognized Character	Efficiency
Image01	44	44	0	100%
Image02	34	34	0	100%
Image 03	34	34	0	100%
Image04	57	55	2	98%
Image 05	44	42	2	98%

IV. Conclusion

India is a country in which more than 22 languages are used as a medium of communication among various populations. All individual language has its own printed as well as handwritten characters. So many researches have been made in the area of pattern recognition. But it is observed that very little work is carried out in Hindi language due to poor recognition rate in context of joint characters. As it is very hard to extract the features of joint characters स्पष्ट, प्रज्ञा, कर्तव्य. It is also observed that no work has been done for handwritten Hindi character as handwriting varies from people to people. Still there is huge scope of research in the field of pattern recognition.

V. Future Work

Currently we are working on joint character component and Handwritten Hindi character recognition.

Acknowledgement

References

[1].   Gonzalez R.C and Woods, R.E "Digital Image Processing", Second Edition, Pearson Education, Singapore.
[2].   "Character Recognition System" by Mohammed Cheriet, Nawwwaf Kharma, Chen-Lin Liu, Ching Y.Suen.
[3].   "Design Implementation of OCR to recognize Gujarati script using Template Matching" by Prof. K S Shah, A. Verma.
[4].   "Optical Character Recognition" by Jesse Hansen.
[5].   "Otsu.N, 1979" A threshold selection method from grey level histogram. IEEE. Trans. Sys. Man. And Cyber, 9:62-66.
[6].   "Pre-processing algorithms for the recognition of tamil hand written characters" by N Shanthi and K Duraiswamy.
[7].   Avanindra and Subhasis Chaudhuri "Robust Detection of Skew in Document Images", IEEE Transaction on Image processing, Vol. 6, No. 2 Feb 1997, p344-p349.
[8].   Changming Sun and Deyi Si, "Skew correction for document images using gradient direction", Fourth ICDAR, Volume 1, p142-146.
[9].   Ming Chen and Xiaoquing Ding, "A Robust skew detection algorithm for grayscale document image".
[10].   Debyan Banerjee, Algorithm for finding skew angle and the de-skewing the image, 14th June 2008.
[11].   U. Pal and B. B. Chauduri, "An improved document skew angle estimation techinque", Pattern recognition letters, Vol 17, No 8, pp.899-904, 1996.
[12].   "Document Image Skew Detection: Survey and annotated bibliography" by Jonathan J. Hull, Ricoh California Research Center, 2882 Sand hill roadm suite 115, Menlo Park, CA 94025.
[13].   "Hierarchial Recognition system for machine printed Kannada Characters by Dinesh Achaya U, N V Subba Reddy and Krishnamoorthi.
[14].   M K Jindal, G S Lehal, R K Sharma, "A study of touching characters in degraded gurmukhi text"
[15].   U. Pal and Anirban Sarkar, "Recognition of printed Urdu Script", Computer vision and pattern recognition unit Indian Statistical Institure, 203 BT Road, Kolkata-108.
[16].   M K Jindal, G S Lehal, R K Sharma, "segmentation of horizontally overlapping lines"
[17].   N Shanthi, K Duraiswamy, "Pre-processing algorithms for the recognition of tamil handwritten characters."
[18].   "Online handwritten recognition for Indian scripts" by Barath A and Sriganesh Madhavanath HP Laboratories, India HPL-2008, May 5-2008.
[19].   "An overview of Character recognition focused on off line handwritten" by Nafiz Arica, Student Member IEEE and Fatos T Yarman-Vural, Senior member IEEE.
[20].   "Research report on the pre processing engine of the optical character recognition system for Sinhala scripts" by dulip hearth and nishantha medagoda language technology research laboratory, University of Colombo School of Computing, Sri Lanka.
[21].   "Skew angle estimation and correction of hand written, testual and larger areas of non-textual document images" A novel approach by D R Ramesh babu , Piyush M. Kumat, Mahesh D Dhannawat.
[22].   G. Ciadiella, G Scafuro, M T Degrandi, M R Spada and M P Roccotelli, "An experimental system for office document handling and text recognition, in proc. Of 9th Int. Conf. Pattern recognition" 1988, p739-743. .
[23].   T.M.Cover and P.E.Hart, "Nearest Neighbour pattern classification," IEEE Trans. Inform. Theory, vol.17, no.1, pp. 15-28, 1978.
[24].   J. M. Keller, M. R. Gray, and J. A. Givens, Jr., "A Fuzzy K-Nearest Neighbor Algorithm", IEEE Transactions on Systems, Man, and Cybernetics, Vol. 15, No. 4, pp. 580-585.