RESEARCH ARTICLE OPEN ACCESS

Recognition of Handwritten Devnagari Numerals

Shaina Gupta[1], Daulat Sihag[2]
Student, Assistant Professor
Department of Computer Science and Engineering
Jan Nayak Ch. Devi Lal Memorial Engineering College, Sirsa
Haryana-India

ABSTRACT

Natural language processing is a field of science and linguistics concerned with the interaction between computers and human languages. Natural language generation systems convert information from computer databases into readable human language. The term "natural" language refers to the languages that people speak, like English and Japanese and Hindi, as opposed to artificial languages like programming languages or logic. "Natural Language processing", programs that deal with natural language in some way or another. Character identification is one of the important subjects in the field of document Analysis and detection. Character identification can be performed on printed text or handwritten text. Printed text can be from good quality documents or degraded documents. The performance of any OCR system heavily depends upon printing quality of the input document. Little reported work has been bringing into being on the detection of degraded Devnagari Numerals. In this paper, we have predictable consider already isolated handwritten devnagari numerals on which we apply Binirazation techniques. This work is performed over 10 Devnagari numerals only. we have used structural and statistical features like Zoning, Transition features, Distance Profile features and Neighbor pixel zone etc. for generating feature sets that are used for recognizing printed Devnagari numerals by using K-NN classifiers and Parameters used for testing have achieved maximum accuracy of **90%** approximate, Squared Correlation Coefficient to get out results with **0.82** approximate with combined (Grad+ Sobel's +Laplacian) feature vector using KNN.

Keywords:- OCR, printed, Feature Extraction, Zoning, Classifer, Squared Correlation Coefficient.

I. INTRODUCTION

The study of human languages developed the concept of communicating with non-human devices. NLP deals with the Artificial Intelligence under the main discipline of Computer Science. The goal of NLP is to design and build software that will analyze, understand and generate languages that humans use naturally. There are many applications of Natural Language processing developed over the years. The main are text-based applications, which involves applications such as searching for a certain topic or a keyword in a large document, translating one language to another or summarizing text for different purposes. The first conceptual idea of OCR is due to Tauschek in 1929 and handle in 1933. Tauschek obtained a patent on OCR in Germany, followed by handle who obtained a U.S. patent on OCR in U.S.A in 1933.Tauschek was also granted a U.S patent on his method 1935.Machine was mechanical device that used templates. The first commercial system was installed at the Reader's Digest in 1955, which, many years later, was donated by Reader Digest to the Smithsonian, here it was put on display. The United States Postal Service has been using OCR machines to sort mail since 1965 based on technology devised primarily by the prolific inventor Jacob Rabinow. In 1974, Ray Kurzweil started the company Kurzweil Computer Products, Inc. and led development of the first omni-font optical character recognition – a computer program capable of recognizing text printed in any normal font. He decided that the best application of this technology of this technology would be to create a reading machine for the blind, which would blind people to understand written text by having a computer read it to them out loud.

Handwritten Documents

Handwritten documents are a kind of degraded documents. Handwritten are widely used in the government offices in India, Libraries, Museums. It's a mechanical device that, when given command, causes Numerals to be written on a paper. Recognition of Handwritten documents is itself a challenge as a typewritten document contains many problems such as broken Numerals, broken headlines, shaping problem, etc.

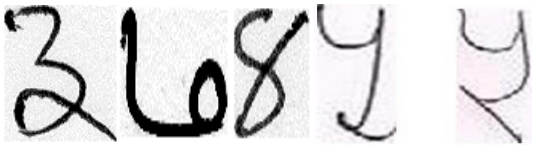

Figure 1.1: Handwritten document in Devnagari Numeral

Introduction to Devnagari Numerals

Devnagari Numerals is used primarily for Hindi language, which is the world's most widely spoken language. Following are the properties of Devnagari Numerals are:

i. Writing style is from left to right.
ii. No concept of upper and lower case Numerals.
iii. Devnagari Numerals is cursive.

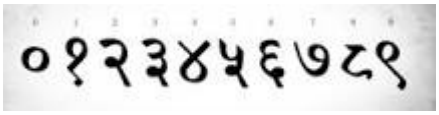

Figure 1.2: Devnagari Numeral

Devnagari Numerals has following challenges [17]:
i. Variability of writing style, both between different writers and between separate examples from the same writer overtime.
ii. Similarity of some Numerals.
iii. Low quality of text images
iv. Unavoidable presence of background noise and various kinds of distortions.

On analysis of the typewritten documents we have observed following problems during character recognition [11, 12]:

1. Touching character problem.
2. Broken character problem.
3. Broken headline problem.
4. Heavy printing problem.
5. Skewness problem.
6. Spacing problem.
7. Shape variance.
8. Background noise problem.

Thus binarization algorithm must perform well to remove these problems. Binarization process deals with the extraction of foreground and background from document image. Whereas, to solve scanning problem image enhancement can be done and some noise is also removed in binarization process.

II. LITERATURE REVIEW

S.Palakollu, R.dhir, R.Rani (2012), This paper mainly deals with the new methods for line segmentation and character segmentation of overlapping characters of Handwritten Hindi text. Algorithm is finding the header lines and base lines by estimating the average line height and based on it. This algorithm works efficiently on overlapped characters for different text sizes and different resolutions images. The new method for line segmentation is working efficiently in the cases of different text sizes and different resolution images. The method which is used for straightening header line is working fine. These methods are also applicable for printed Hindi text. The lines which have broken parts in upper modifiers are not correctly recognized. The lines with thick parts in upper modifiers also not correctly recognized. Touching lines are not correctly recognized [10]. **Ashwin S Ramteke, Milind E Rane(2012),** The process of Segmentation is a vital phase in the recognition of text. Devanagari is very useful Script in India. The segmentation of devanagari words is very difficult due to the presence of large character set that include consonants, vowels and modifiers. In this paper the major focus was on the segmentation of line, word and characters. Before the segmentation of an image some preprocessing of the image is done using the median filter and it also includes the binarization and scaling of image. After this preprocessing the segmentation is done. For the Segmentation of handwritten Devanagari script the histrogram of input image is generated that shows the space b/w the characters so from this the characters can be segmented. The handwritten data set are collected from different users of different background on blank bank cheque format as well as on plane papers. The algorithm is implemented in the Matlab. The algorithm is tested with the large number of input images. The segmentation accuracy for this implementation depends upon the proper writing i.e. non-overlapping or characters, proper space between words and characters, proper connection of characters through shirorekha. The segmentation for word gives 98% of accuracy, for characters 97% of accuracy. The implementation not gives that much accurate result for the broken characters. For numerical segmentation implementation gives the 100% accurate result **[12].**
In this paper we present a system towards the segmentation of Hindi Handwritten Devnagari Text. Segmentation of script is essential for handwritten script recognition. This system deals with segmentation of modifiers (matras) and fused characters in handwritten Devnagari word. Also describe carries out segmentation in hierarchical order. First the header line is identified and segmented. Segmentation of modifiers consists of segmentation of top as well as bottom modifiers. In the last step the fused characters are also a segmented. In recent years, OCR (Optical Character Recognition) technology has been applied throughout the entire spectrum of industries, revolutionizing the document management process. Segmentation is the most important step before recognition. The accuracy of the recognition depends on the segmentation. Better is the segmentation less is the ambiguity encountered in recognition. Segmentation of handwritten text is very critical due to the irregularities in shape and size of the handwritten characters. Very little work is reported towards the segmentation of handwritten text. Proposed approach is suitable for segmentation of handwritten Devnagari (Hindi) text. Very little database is available online for Devnagari .So this filed is still one of the prominent filed for research [7]. **N.K Garg, L. Kaur & M.K. Jindal(2011),** OCR is used to recognize the sccaned text that can be in the form of handwritten or typed form. Segmentation is the important phase in the character recognition that can improve/decrease the accuracy of character recognition. Segmentation of printed words is quite easy as compare to handwritten words because of the various problems that will occur in the segmentation of handwritten text. There are two types of problems that can occur in the segmentation of handwritten text: The Problems that can be ignored (Like the problems due to speed of writing). The problems that can be ignored. The problems explained above are very useful for complete segmentation of handwritten Hindi text. Some problems can be removed if writer uses the better material and write patiently. To solve the problems related with writer's natural handwriting efficient algorithms

are to be designed to segment the text. The study may be carried out in future in the following direction: The efforts should be made to solve the above problems. It is very difficult to determine the presence of lower modifier and to determine the presence of conjuncts in middle region of the word. The writer's can be given instructions to write patiently to minimize some of the problems of segmentation like touching characters. The algorithms used in other Indian scripts for similar problems can be tried on handwritten Hindi text [6]. **N.k garg, L kaur, M.K Jindal(2010),** The main purpose of this paper is to provide the new segmentation technique based on structure approach for Handwritten Hindi text. The handwritten text is separated into lines, lines into words and words into characters. The errors in segmentation propagate to recognition.

The performance is evaluated on handwritten data of 1380 words of 200 lines written by 15 different writers. The overall results of segmentation are very promising. this is the first paper with complete handwritten Hindi text segmentation. The paper is organized as we have discussed the creation of database used for the experimental purposes, includes the discussion about the characteristics of Hindi language, The segmentation technique used for segmenting the handwritten Hindi text are segmentation of lines from the text, words from the lines and last one is segmentation of character from the word.

The result of these segmentation techniques are accuracy of text line and word segmentation is (91.5, 98.1).Then the segmentation problem occurs in ascenders when two ascenders touch each other. Some errors in word separation occur due to incorrect line segmentation. The errors which occurs in text line segmentation also creates problem in word segmentation and character segmentation [4].

III. METHODOLOGY

Algorithm for Creating Character Image

Step 1: Read Image: In this step we will read image one by one from database I(x, y).

Step 2: Pre-processing: Here, we convert the image I(x, y) into gray level image G(x,y) if the image is in RGB format and convert into logical format.

G(x, y) =logical(rgb2gray(I(x, y)));

Step 3: **Binarization:** Now, the gray level image G(x ,y) is binarized by using different binarization techniques such as OTSU Method, Local Method, and Entropy Based Method to get the image connected completely i.e. Proper binarization B(x,y) of image G(x,y).

Step 4: In above applied step, we remove small connected components whose threshold connective pixels are 8.

Step 5: Resize Image: We will resize image by using Bi-cubic method to generate 25 x 25 images.

Step 6: Universal Discourse: We clipped the character images by removing extra white spaced rows and columns residing in four sides of image.

Step 7: Matrix Generation: As, after step 6 we have to generate image of 25 X25 pixels size for this we added half numbers of rows top and bottom and columns on left and right side of image.

Step 8: Erosion Method: After Step 1 to 4, we implemented Structure elements of square with value 2 for Erode, as for removing broken character which is connected.

Step 9: After applying step 1 to 6, refinement of background and shape.

Erosion Process: is similar to dilation, but we turn pixels to 'white', not 'black'. As before, slide the structuring element across the image and then follow these steps:

1. If the origin of the structuring element coincides with a 'white' pixel in the image, there is no change; move to the next pixel.
2. If the origin of the structuring element coincides with a 'black' pixel in the image, and at least one of the 'black' pixels in the structuring element falls over a white pixel in the image, then change the 'black' pixel in the image (corresponding to the position on which the center of the structuring element falls) from 'black' to a 'white'.

Algorithm for Zoning:

The frame containing the character is divided into several overlapping or non-overlapping zones. The densities of the points or some features in different regions are analyzed.

Step 1: Creating Zones: For, creating zones find height and width of the zones by dividing rows and columns by the NXM zones resp.

Step 2: Zones: Now, define each matrix with resp. to image by using zone height and zone width values and determine each zone matrix of N X M zones.

zone11=image (1:zone_**height**, 1:zone_width);

The goal of zoning is to obtain the local characteristics instead of global characteristics. We have created 25 (5*5) zones.

Algorithms for Zoning Density

Step 1: Divide the input image in to 25 equal zones.
Step 2: Compute no. of total black pixels in each zone.
Step 3: compute the average of step 2. (One feature)
Step 4: Repeat step 2 & 3 for each zone. (25 feature)

Algorithms for zoning pixel distance Metric:

Step 1: Divide the input image in to 25 equal zones.
Step 2: Compute pixel distance present in the zone column in VDD
Step 3: Repeat the step 2 for the entire pixels present in the zone column.
Step 4: Compute average pixel distance in zone column (one feature).
Step 5: Repeat the steps 2 to 4 for the entire zone columns present in the zone (10 features)
Step 6: Compute pixel distance present in the zone column in VUD
Step 7: Repeat the step 6 for the entire pixels present in the zone column.

Step 8: Compute average pixel distance in zone column (one feature).
Step 9: Repeat the steps 6 to 8 for the entire zone columns present in the zone (10 features)
Step 10: Compute pixel distance present in the zone row in HRD
Step 11: Repeat the step 10 for the entire pixels present in the zone row.
Step 12: Compute average pixel distance in zone row (one feature).
Step 13: Repeat the steps 10 to 12 for the entire zone rows present in the zone (20 features)
Step 14: Compute pixel distance present in the zone row in HLD
Step 15: Repeat the step 14 for the entire pixels present in the zone row.
Step 16: Compute average pixel distance in zone row (one feature).
Step 17: Repeat the steps 14 to 16 for the entire zone rows present in the zone (10 features)
Step 18: Repeat the steps 5, 9, 13, 17 sequentially for the entire zone present in the image.
Step 19: Finally 500 features are extracted for classification and recognition.

Algorithm Gradient: Directional Features

As for finding the gradients, we had applied Sobel's and canny mask as to calculate the horizontal gradient (gx) and vertical gradient (gy) components as shown in Figure below:
We had calculated the gradient of a pixel (i, j) by using following formula:

$$G_x = gv(i,j) = f(i-1,j+1) + 2f(i,j+1) - f(i-1,j-1) - 2f(i,j-1) - f(i+1,j-1)$$

(Eq-1)

$$G_y = gh(i,j) = f(i-1,j-1) + 2f(i-1,j) + f(i-1,j+1) - f(i+1,j-1) - 2f(i+1,j-1) - f(i+1,j+1)$$

(Eq-2)

$$grad = G_y / G_x = \tan^{-1}\left[gh(i,j) / gv(i,j) \right]$$

After computing the gradient of each pixel of the character, we map these gradient values onto 12 direction values with angle span of 45 degree between any two adjacent direction values. The orientations of these 8 directional values. The mapping of gradient values on 12 directional values can be calculated by generalized formula as given below:
Direction[n(i,j)]=$(45(n-1)^0 <= grad(i, j) < (45n^0)$

If a pixel is surrounded by all the pixels having values zero then is its gradient assigned as 1. During the calculation of directional feature if gradient values are -1 then its directional feature values are assigned the values zeros(0s). Here we get out 300 features by gradient.

Size of Features Vectors

Feature vector are the values which are used for recognizing isolated Devnagari numerals. We have used various techniques to obtain these feature vector values each techniques give a different set of feature vector. The various techniques (individual or combinations) and there no. of feature vector for each technique is as follows:

IV. RESULTS

The results of knn using various kinds of structural and statistical features of and options performed by us are presented in Table 1–4. Table 1-4 show the results for only enhanced binarized samples for eroded sample set used for recovering of broken samples. Parameters used for testing are Accuracy (values has been normalized to [0 1]), Mean Square Ratio, and Squared Correlation Coefficient. Table 1 shows the recognition accuracy on odd trained and even tested by using kNN as a linear; from Table 1 it can be observed that we get maximum accuracy of 78%and maximum Squared Correlation Coefficient of 0.74 with combined techniques (GRAD+Sobel +Laplician) means all combined, feature vector (F8) whereas Mean Squared Ratio is less in Zoning Centroid Zone Technique (ZCZ), feature vector (F7).

Table 1: Recognition KNN Parameters values for various Feature extraction Techniques.

S.No.	Odd trained even tested	Parameters		
		Accuracy	Mean sq. ratio	Sq. col. cof.
1.	projection	0.7	0.68	0.66
2.	Zpd	0.62	0.66	0.66
3.	Zad	0.74	0.76	0.76
4.	Gmadsob	0.7	0.74	0.6
5.	Gmadlap	0.72	0.74	0.72
6.	zpd+zad	0.62	0.66	0.66
7.	gmadsoblap	0.8	0.78	0.74
8.	all combined	0.78	0.74	0.74

Table 2 shows the recognition accuracy by training even data samples and testing over odd data samples by using KNN type kernel is as linear; from Table 2 it can be observed that we get maximum accuracy of 88% and maximum Squared Correlation Coefficient of 0.8 and Mean Squared Ratio is less with combined techniques (GRAD+Sobel+Laplician), feature vector (F8).

Table 2: Recognition KNN Parameters values for various Feature extraction Techniques.

S. No.	Even trained odd tested	Parameters		
		Accuracy	Mean sq. ratio	Sq. col. cof.
1.	projection	0.78	0.8	0.78
2.	Zpd	0.74	0.72	0.74
3.	Zad	0.8	0.82	0.8
4.	gmadsob	0.82	0.72	0.78
5.	gmadlap	0.82	0.82	0.74

6.	zpd+zad	0.74	0.72	0.74
7.	gmadsoblap	0.84	0.8	0.78
8.	all combined	0.88	0.88	0.8

Table 3 shows the recognition accuracy on First half samples are trained and last half samples are tested by using KNN type kernel is linear; from Table 3. it can be observed that we get maximum accuracy of 90% and maximum Squared Correlation Coefficient of 0.82 and Mean Squared Ratio is less with combined techniques (GRAD+Sobel+Laplician), or (GRAD+Sobel) feature vector (F8).

Table 3: Recognition KNN Parameters values for various Feature extraction Techniques.

S .No.	First trained Last tested	Parameters		
		Accuracy	Mean sq. ratio	Sq. col. cof.
1.	Projection	0.84	0.82	0.78
2.	Zpd	0.78	0.84	0.8
3.	Zad	0.84	0.82	0.86
4.	Gmadsob	0.84	0.8	0.7
5.	Gmadlap	0.86	0.82	0.8
6.	zpd+zad	0.78	0.84	0.8
7.	gmadsoblap	0.86	0.8	0.78
8.	all combined	0.9	0.86	0.82

Table 4 shows the recognition accuracy by training even data samples and testing over odd data samples by using KNN type kernel is aslinear; from Table 4 it can be observed that we get maximum accuracy of 84% with combined techniques (ZDF+ZPA), feature vector (F4) whereas maximum Squared Correlation Coefficient of 0.959 and Mean Squared Ratio is less with combined techniques (GRAD+ZCI+ZCZ), feature vector (F8). Figure 10 shows the results of Table 5 in graphical bar format by showing different colors for each Technique.

Table 4: Recognition KNN Parameters values for various Feature extraction Techniques.

S.No.	Last trained First tested	Parameters		
		Accuracy	Mean sq. ratio	Sq. col. cof.
1.	Projection	0.86	0.82	0.84
2.	Zpd	0.78	0.74	0.68
3.	Zad	0.82	0.8	0.82
4.	Gmadsob	0.8	0.72	0.76
5.	Gmadlap	0.82	0.78	0.78
6.	zpd+zad	0.78	0.74	0.68
7.	Gmadsoblap	0.84	0.78	0.82
8.	all combined	0.86	0.84	0.78

V. CONCLUSION & FUTURE WORK

This dissertation presents recognition of isolated typewritten Devnagari characters. Mainly two stages including feature extraction and classification are carried out in detail. Major problem in this recognition from above described problems which are been undertaken for recognition are broken character, heavy printed, broken headline, and shape variance characters. The structural and statistical features selected for recognition of Handwritten Devnagari numerals which are robust to noise are used such as Neighbour Pixel Zone, Zoning Density and Transitions Feature. K-NN have been used for classification purpose. By analysing, results we can conclude that K-NN shows best recognition for Handwritten Devnagari numerals.

The work can also be extended to recognition of other typewritten Indian scripts containing these kinds of degradations and recognition of these documents can be enhanced by creating a complete Optical Character Recognition system as little work has been proposed before and a wide scope of work to be done for enhancing the Printed Devnagari script documents containing these kinds of degradation, and subsequently recognizing them.

REFERENCES

[1] Pal U., Chaudhuri B. B., Indian Script character recognition: a survey, *International Journal of Pattern Recognition*, 37 (2004), pp. 1887-1899.

[2] Shi Z., Setlur S., Govindaraju V., Text extraction from gray scale historical document images using Adaptive Local Connectivity Map, in proceedings of ICDAR, 2 (2005), pp. 794-798.

[3] Lemaitre A., Camillerapp J., Text line extraction in handwritten document with Kalman Filter applied on low resolution image, in proceedings of international conference on DIAL, (2006), pp. 38-45

[4] Garg , Naresh kumar, kaur, Lakwinder and jindal, M.K. 2010. A Segmentation of Handwritten Hindi text. In International Journal of Computer Applications (0975 – 8887) Volume 1 – No. 4

[5] Vikas j Dondre and Vijay H Mankar 2010. A Review of research devnagari character recognition.

[6] N.k Garg, L.kaur, and M.K jindal 2011. The hazards in segmentation of handwritten Hindi text. In international journal of computer applications (0975-8887) volume 29-No.2.

[7] Mr. Sandip N.Kamble, Prof.Mrs. Megha Kamble 2011. Morphological Approach for Segmentation of Scanned Handwritten Devnagari Text. In International Journal of Emerging trends in Engineering and Development ISSN 2249-6149 Issue1,Vol.3.

[8] Aarti desai, latesh malik, rashmi welekar 2011. a new methodology for devnagari character recognition. in jmijit ,volume -1 issue 1 ©jm academy issn: print 2229-6115.

[9] Mr. Dipak V. Koshti, Mrs. Sharvari Govilkar. The segmentation of touching characters in handwritten devnagari script. In IJACEE Volume 2: Issue 2 [ISSN 2250 - 3765].

[10] Saiprakash Palakollu, Renu Dhir, Rajneesh Rani 2012. Handwritten Hindi text segmentation techniques for lines and characters. In Proceedings of the World Congress on Engineering and Computer Science 2012 Volume IWCECS 2012, San Francisco, USA.

[11] Segmentation of Handwritten Hindi Text: A Structural Approach M. Hanmandlu and Pooja Agrawal.

[12] Ashwin S Ramteke, Milind E Rane,"Offline Handwritten Devanagari Script Segmentation" in 2012.

[13] Shailedra Kumar Shrivastava, Sanjay S. Gharde," Support Vector Machine for Handwritten Devanagari Numeral Recognition", International Journal of Computer Applications (0975 – 8887), Volume 7– No.11, October 2010.

[14] Mahesh Jangid Kartar Singh, et .al, "Performance Comparison of Devanagari Handwritten Numerals Recognition", International Journal of Computer Applications (0975 – 8887), Volume 22– No.1, May 2011.