An efficient method of segmentation for handwritten devnagari word recognition

Saniya M.Ansari, Dr. Udaysingh Sutar

Abstract— Devnagari is the most popular and widely used script in India. It is used for writing Hindi, Marathi, Sanskrit and Nepali languages. Moreover, Devnagari script consists of vowels, consonants and various modifiers. Detection and extraction of text in images have been used in many applications. Document segmentation is one of the difficult and important phases in machine recognition of any language. The accuracy of character recognition engine depends on the correct segmentation of individual symbols. It is used to segment lines and words into sequence of characters into sub images of individual symbols. Hence proper segmentation of Devnagari word is challenging task. Especially the modifiers (both vowels and consonants) most of the time coincide with the modifying characters. These kinds of non-trivial combinations of characters make the whole process of character segmentation extremely challenging. Besides, some symbols, like Chandra- Bindu, often come between two consecutive characters in a word; then isolating those becomes a tough job.

Index Terms— Preprocessing, Edge Detection, Gaussian Smoothing, Edge detection, Vertical & Horizontal Segmentation

1 INTRODUCTION

The amount of information that can be processed and stored by computers is increasing at a huge rate. Due to this increase in rate the ease with which the information can be exchanged between a computer and a user is creating major problems. In order to fully utilize the tremendous processing capability of the computer, a user interface should not only be efficient but also natural. The user interface must be efficient in order to speed up the communication. The primary modes of data input between a user and a computer are still the conventional input devices such as keyboards and mouse. These devices have certain limitations when compared to the input through natural handwriting. For scripts such as Chinese and Japanese which have a very large alphabet set and due to complex typing nature of scripts such as Devanagari and Gurumukhi, it becomes difficult to input data to the computer through the conventional input devices.

Natural handwriting is one of the easiest ways to exchange information between a human and a computer. Thus, the handwriting recognition field has great potential to improve the communication between the user and the computer. Handwriting recognition is basically the ability of a computer to interpret the data which is basically the intelligible handwritten input from various sources such as paper documents, touch screens and other devices. Handwriting recognition is in research for many decades and many researches on this field are going around all over the world. Great advances have been made in this field and due to this reason, the usage and reliability of online handwriting based devices such as Tablet PC and PDA (Personal digital assistant) has increased a lot.

- *Saniya Ansari is currently pursuing Ph.D. degree in ECE in Karpagam University, Coimbatore, India,. E-mail: Ansari.saniya6@gmail.com*
- *Dr.Udaysingh Sutar is a research guide in Karpagam University, Coimbatore,India,. E-mail: ussutar@aissmscoe.in*

2 LITERATURE SURVEY

Dharam Veer Sharma and Gurpreet Singh Lehal[1] proposed a Segmentation of handwritten text in Gurmukhi script . The proposed technique segments the words by focusing on presence of headline, aspect ratio of characters and vertical and horizontal projection profiles. For testing of the algorithm a set of 1907 handwritten words has been considered, from which 389 sets of connected characters were extracted after first phase of segmentation procedure. These 389 character sets belonged to 234 (20.39% of 1907) words, of these 389 connected character sets 3 sets each were present in 23 words, 2 sets each were present in 109 words and 102 words had only one set each of connected characters. In the next phase, the remaining 264 words, with over-segmented characters are handled and of these 234 (95.12% of 246) words are properly segmented and the remaining (4.88%) error was primarily because of broken characters with gaps greater then the threshold value. The overall successful segmentation achieved through the proposed procedure is 96.22%.

Bikash Shaw, Swapan Kr. Parui et al.[2] presented a novel segmentation based approach for recognition of offline handwritten Devanagari words. A hidden Markov model is used for recognition at pseudocharacter level. The word level

recognition is done on the basis of a string edit distance. The training and test databases here consist respectively of 22500 and 17200 images of handwritten words of 100 word classes collected from 436 different writers. lexicon set is rich in variation in the sense that it covers all individual basic characters (independent vowels and consonants), dependent vowel signs and all commonly occurring compound characters. Also, the word length in the used lexicon covers a wide spectrum (from 2 to 7 characters). The word level accuracy is 84.31% on the test set.

Olarik Surinta and Rapeeporn Chamchong, [3] presented their work on Palm leaf manuscripts ,one of the earliest forms of written media and were used in Southeast Asia to store early written knowledge about subjects such as medicine, Buddhist doctrine and astrology. The author presented an image segmentation of historical handwriting from palm leaf manuscripts. The process is composed of three steps: 1) background elimination to separate text and background by Otsu's algorithm 2) line segmentation and 3) character segmentation by histogram of image. The end result is the character's image. The results from this research may be applied to optical character recognition (OCR) in the future. Their proposed algorithm first converts the color image into a grey image, then converts it into a binary image using Otsu's algorithm. Finally the segmented lines and characters are produced using projection profile analysis.

B.M. Sagar, DR. G. Shobha [4] discussed various character segmentation algorithms for Kannada scripts which is one of the South Indian languages which have 16 vowels and 34 consonants as the basic alphabet of the language. The author explained characteristics of Kannada scripts, two surveyed segmentation algorithms and one implemented segmentation algorithm by brute force approach.

Vassilis Papavassilioua,b, Themos Stafylakisa [5] presented a two novel approaches to extract text lines and words from handwritten document. Viterbi algorithm is used for the line segmentation based on locating the optimal succession of text and gap areas within vertical zones. Then, a text-line separator drawing technique is applied and finally the connected components are assigned to text lines. Word segmentation is based on a gap metric that exploits the objective function of a linear SVM that separates successive connected components. The algorithms tested on datasets of ICDAR07 handwriting segmentation contest and outperformed the participating algorithms. They have presented two effective techniques for segmenting handwritten documents into text lines and words. In word segmentation, a novel metric is used to measure the separability of adjacent CCs.

Muthukrishnan. R. and M. Radha [6], elaborated an attempt to study the performance of most commonly used edge detection techniques for image segmentation and also the comparison of these techniques is carried. In proposed work an attempt is made to review the edge detection techniques which are based on discontinuity intensity levels. It is observed from the results MarrHildreth, LoG and Canny edge detectors shows almost same edge map. Canny result is superior one when compared to all for a selected image since different edge detections work better under different conditions.

Vikas J Dongre [7] proposed a simple histogram based approach to segment Devnagari documents. Various challenges in segmentation of Devnagari script are also discussed. It is observed that line segmentation is done with nearly 100% accuracy. Word segmentation is accurate as long as the document contains characters only. The accuracy of recognition is redused when Devnagari numerals are present in the document. Here each digit is considered as separate word by the proposed algorithm. In the present case it is 91%. They have presented a primary work for segmentation of lines, words and characters of Devnagari script. Nearly 100% successful segmentation achieved in line and word segmentation.

Mamatha H R,[8] Segmentation is an important task of any Optical character recognition system. Segmentation of handwritten text of Indian languages like Kannada, Telugu, Marathi, Hindi is difficult as compare to Latin based languages. Indian languages have structural complexity and increased character set. It contains vowels, consonants and compound characters and overlapped characters. Even though several successful works in OCR all over the world, development of OCR tools in Indian languages is still an ongoing process. The author presented a segmentation scheme for handwritten Kannada scripts into lines, words and characters using morphological operations and projection profiles. The method was tested on totally unconstrained handwritten Kannada scripts. Due to variable inter and intra word gaps an average segmentation rate of 82.35% and 73.08% for words and characters respectively is obtained.

N. Anupama, Ch. Rupa [9] proposed an algorithm based on multiple histogram projections using morphological operators to extract features of the image. Horizontal projection is used on the text image, and then horizontal projection is used for line segments. Threshold is applied to divide the text image into segments. Vertical histogram projections are used for the line segments and decomposed into words using threshold. This approach provides best performance based on the experimental results such as Detection rate DR (98%) and Recognition Accuracy RA (98%).

3 PROPOSED METHODOLOGY

Our proposed system is an online Devnagari word recognition system which is applied on an input taken from i-ball Digital Tablet. The steps that are followed in the proposed work for preprocessing are shown in Figure 1.

As demonstrated in this document, the numbering for sections upper case Arabic numerals, then upper case Arabic numerals, separated by periods. Initial paragraphs after the section title are not indented. Only the initial, introductory paragraph has a drop cap.

3.1 Data collection

The data collection step is responsible to pass user's input data to the preprocessing and subsequently to the feature extraction processes and displays recognized letter. Once the user draw gesture it will pass to the feature extraction phase where

features will extract which is useful for recognition. Proposed system is capable to draw any character or modifier to form a word.

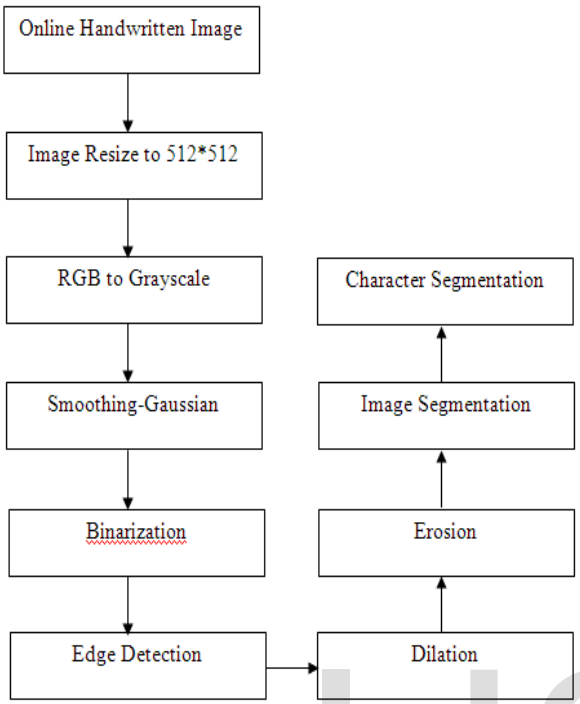


Figure 1. Results of Devnagari Image Preprocessing and Segmentation for 'Pune 'word

Figure 1 In the proposed method we have android based i-ball Digital Tablet as an input device. The user interface of Android is based on direct manipulation, using touch inputs that loosely correspond to real-world actions, like swiping, tapping, pinching and reverse pinching to manipulate on-screen objects. Gesture class is used to capture gesture which user will draw on screen. Compared to the digital tablet, collecting data using mouse is not appropriate because the collected data could be noisy and the trajectory can be more or less jagged. In this proposed work, an effort is made to setup Devnagari online handwriting data set which can be used to train and test the recognition model in the research and can serve as a resource for upcoming researches.

3.2 Preprocessing Of Input Images

Preprocessing is the most important phase in an online handwriting recognition system. The main purpose of preprocessing phase in handwriting recognition is to remove noise or distortions present in input text due to hardware and software limitations and convert it into a smooth handwriting. These noise or distortions include different size of text, missing points in the stroke during pen movement, jitter present in stroke, left or right slant in handwriting and uneven distances of points from adjacent positions. Before features are derived from the handwritten word, the raw data recorded by the hardware goes through several preprocessing steps. The main objective of the preprocessing steps is to normalize word and remove variations that would complicate recognition. Based on the previous work done by various authors it can be easily identified that basic steps for preprocessing in an online handwriting recognition system for any language are almost similar and can be summarized as Interpolation of missing points, Removing duplicate points, elimination of hooks, size normalization, Resampling, slant correction, Smoothing. In online handwriting recognition, preprocessing includes five common steps, namely, size normalization and centering, interpolating missing points, smoothing and resampling of points. In the preprocessing phase of online handwriting recognition system under consideration, we used following steps like Image resize, Smoothing, Binarization, edge detection. [10]

3.2.1 Gray scale conversion:
RGB image is required to be converted gray scale which is done by using rgb2gray function.

3.2.2 Image Resizing
To resize an image, the imresize function is used. An image to be resized using the magnification factor. An image can be enlarged by specifying a magnification factor greater than 1. And reduced by specifying a magnification factor in between 0 and 1.The size of the output image can be specified by passing a vector which contains the number of rows and columns in the output image. More complex variation of scaling algorithms are bilinear, bicubic, spline, sinc, and many others. Nearest neighbor is the simplest and fastest implementation of image scaling technique. In our proposed work we have used 512 * 512 size using imresize function to resize input image.

3.2.3 Image Denoising and Smoothing
Once image is resized and converted into grayscale format, further we have to preprocess it for removal of noise and enhance its contrast by using Gaussian filter. The function h = fspecial ('gaussian',n, sigma) returns a rotationally symmetric Gaussian lowpass filter with standard deviation sigma (in pixels). n is a 1-by-2 vector specifying the number of rows and columns in h. If you do not specify the parameters, fspecial uses the default values of [3 3] for n and 0.5 for sigma. We are using Gaussian filter using sigma 1 and size for filter is [2 2].
☐ Final preprocessed image is generated as output of preprocessing.

3.3 Segmentation

Segmentation is the phase in which data is decomposed at character or stroke level so that nature of each character or stroke can be studied individually. The preprocessing stage provides a "clean" document in the sense that sufficient information of shape, high compression, and low noise on a normalized image is achieved. The next stage is segmenting the input handwritten word into its subcomponents like characters, lower and upper modifiers.
Character segmentation strategies can be classified into three categories as Explicit Segmentation, Implicit Segmentation and Mixed Strategies. In an Explicit Segmentation the seg-

ments are identified based on "character-like" properties. The process of cutting up the image into meaningful components is given a special name: dissection. An Implicit Segmentation is based on recognition. It searches the image for components that match predefined classes. Segmentation is performed by the use of recognition confidence, including syntactic or semantic correctness of the overall result. In this approach, two classes of methods can be employed: 1) methods that make some search process and 2) methods that segment a feature representation of the image. The Mixed Strategies combines explicit and implicit segmentation in a hybrid way. A dissection algorithm is applied to the image, but the intent is to "over segment," i.e., to cut the image in sufficiently many places that the correct segmentation boundaries are included among the cuts made. Once this is assured, the optimal segmentation is sought by evaluation of subsets of the cuts made. Mixed strategies yield better results compared to explicit and implicit segmentation methods. [11]
In our proposed work we have used following steps prior to segmentation.

3.3.1 Binarization

Grayscale preprocessed image is further segmented using thresholding method in which pixels those having intensity value less than 128 are kept as black and rest all kept as white pixels. Output of binarization is segmented binary image i.e. bi-level image by using an optimal threshold. Thus the purpose of binarization is to mark pixels that belong to true foreground regions with a single intensity and background regions with different intensities.

3.3.2 Edge Detection

Edge detection is an image processing technique that used to detect the edges of the picture. There are several types of edge detection algorithms. Edges are the significant local changes of intensity in an image and edges typically occur on the boundary between two different regions in an image. The goal of edge detection is to identify the important feature that will be useful for feature extraction phrases. Sobel Edge detection operator is used to detect the edges on binary image. The Sobel operator performs a 2-D spatial gradient measurement on an image and it is used to find the approximate absolute gradient magnitude at each point in an input grayscale image. The Sobel edge detector uses a pair of 3x3 convolution masks, one calculates the gradient in the x-direction (columns) and the other calculates the gradient in the y-direction (rows). A convolution mask is usually much smaller than the actual image and it is used to slide over the image, manipulating a square of pixels at a time.

3.3.3 Dilation

The basic role of dilation morphological operator is that the value of the output pixel is the maximum value of all the pixels in the input pixel's neighborhood. In a binary image, if any of the pixels is set to the value 1, the output pixel is set to 1. We are applying dilation on edge detected image by using two flat linear structuring elements with angle of 180 and 90.

3.3.4 Clear border:
This is used to suppress light structures connected to image border of dilated image.

3.3.5 Erosion
This morphological function is used after dilation on image in order to erode the image and returning the eroded image using diamond-shaped structuring element. Finally segmented image is generated using finding perimeters objects from image.

3.4 Word Segmentation Algorithms

3.4.1 Vertical Character Segmentation Algorithm

Step 1: Finding white pixel from segmented image column wise
Step 2: Check if column having number of white pixels less than or equal to 10, its true assign 0 value to entire column to make sure that it is represented as black, else keep as it is in output image.
Step 3: Below for loop showing how it is done.

```
for i=1:c
a = nnz (a2 (:,i)); // measure the number of white pixels from each column.
if (a <= 10)
opim (:,i) = 0;
else
opim (:,i) = a2 (:,i);
end
end;
```

Step 4: This is also referred as vertical segmentation.

3.4.2 Horizontal Character Segmentation algorithm

Step 1: Finding white pixel from segmented image row wise
Step 2: Check if row having white pixels and then start counting number of rows those having white pixel using count variable.
Step 3: upper body segmentation
If count == 10,
Insert two black rows after 10th row to represent upper part of devnagari word.
End
Step 4: lower body segmentation, this can be done by just applying reverse process of step 3 above on character segmented image.

4. PERFORMANCE EVALUATION

We have measured performance of preprocessing method used by using PSNR, Mutual information, MSE. Below table 1 showing same for 7 different images. From below table it is clear that our used preprocessing method improves the quality of input image better than existing methods and solutions in literature. This further improves the accuracy of recognition. Another important advantage of this approach is take it takes very less time for preprocessing and image segmentation.

4.1 Mutual Information

Mutual information is the amount of information that one variable contains about the other. It can be considered as a measure of how well one image explains the other. The most commonly used measure of information in image processing is the Shannon- Wiener entropy measure. Given m events occurring with probabilities p1…….pn the Shannon entropy is defined as shown in equation(1)

$$H = \sum_{i=1}^{m} p_i \log 1/p_i = -\sum_{i=1}^{m} p_i \log p_i \quad (1)$$

For an image the entropy is calculated from the image intensity histogram in which the probabilities are the histogram entries. In image registration as there are two images joint entropy will have to be also considered. Joint entropy measures the amount of information we have in the two images combined. The Joint entropy H (I, J) can be calculated using the joint histogram of two images. When the images are more similar, the joint entropy will be lower as compared with the sum of the individual entropies. The joint entropy is defined by using equation (2).

$$H(A, B) \leq H(A) + H(B) \quad (2)$$

The registration of two images can be accomplished by reducing the joint entropy of the images, but mutual information is a better criterion as marginal entropies H (I) and H (J) are taken into account. The mutual information is given by equation (3).

$$MI(A,B) = H(A) + H(B) - H(A,B) \quad (3)$$

4.2 Mean Square Error (MSE):

Mean Square Error can be estimated in one of many ways to quantify the difference between values implied by an estimate and the true quality being certificated. MSE is a risk function corresponding to the expected value of squared error. The MSE is the second moment of error and thus incorporates both the variance of the estimate and its bias. [12]
The MSE of an estimate and is defined by equation (4).

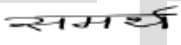
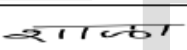
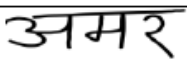
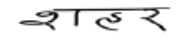
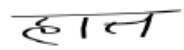
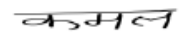
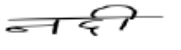
$$MSE = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}(x(i,j) - y(i,j))^2 \quad (4)$$

Where x(i, j) represents the original (reference) image and y (i, j) represents the distorted (modified) image and i and j are the pixel position of the M×N image. MSE is zero when x(i, j) = y(i, j) .Lower the value of MSE higher the quality of image.

4.3 Peak Signal to Noise Ratio:

The PSNR is most commonly used as a measure of quality of reconstruction of lossy compression codecs. The signal in this case is the original data, and the noise is the error introduced by compression. A higher PSNR would normally indicate that the reconstruction is of higher quality. Performance Evaluation of Preprocessed Images using PSNR, MSE and Mutual Information metrics.

Table 1: Perfrmance evaluation based on Mututal information, MSE & PSNR

Devnagari Script	Mutual Information	MSE	PSNR
	0.310	99.28	83.014
	0.25	75.19	86.52
	0.26	71.13	87.46
	0.25	64.59	88.72
	0.26	75.75	86.75
	0.30	94.70	83.63
	0.252	77.08	86.41

4.3 Preprocessing Stage Output results

Figure 2 and 3 below showing the outputs of each step used during preprocessing and segmentation.

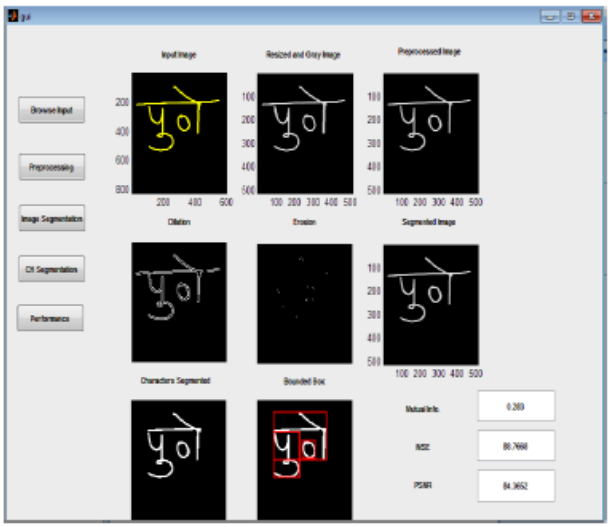


Fig. 1 . Results of Devnagari Image Preprocessing and Segmentation for 'Pune 'word

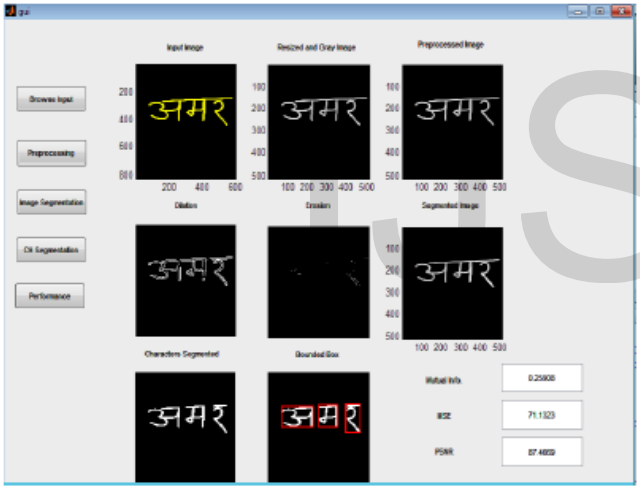


Fig. 2. Results of Devnagari Image Preprocessing and Segmentation for 'Pune 'word

5. CONCLUSION AND FUTURE WORK

Preprocessing is first vital step of any image processing. Use of methods in preprocessing and segmentation defines the efficiency and accuracy of handwritten character recognition method. We have used first Gaussian filter on resized image and then applied thresholding method for binarization, this both approaches not only gives us better outputs but also taking very less time for processing. The image is segmented further using dilation, erosion and perimeter detection morphological operations. Characters from segmented image are separated using vertical segmentation method and then horizontal segmentation method. The performance results such as PSNR, MSE and Mutual information showing very improved quality of preprocessed image as compared to existing solutions.

ACKNOWLEDGMENT

I would like to thanks to support and all the assistance provided by Karpagam University during the research work.

REFERENCES

[1] Dharam Veer Sharma1 and Gurpreet Singh Lehal2 " An Iterative Algorithm for Segmentation of Isolated Handwritten Words in Gurmukhi ScriptThe 18th International Conference on Pattern Recognition (ICPR'06) 0-7695-2521-0/06 $20.00 IEEE computer society© 2006

[2] Bikash Shaw, Swapan Kr. Parui, Malayappan Shridhar "Offline Handwritten Devanagari Word Recognition: A Segmentation Based", ©2008 IEEE ,978-1-4244-2175-6/08

[3] Olarik Surinta and Rapeeporn Chamchong, "Image Segmentation of Historical Handwriting from Palm Leaf Manuscripts" 2008, in IFIP International Federation for Information Processing, Volume 288;Intelligent Information Processing IV; Zhongzhi Shi, E. Mercier-Laurent, D. Leake; (Boston: Springer), pp. 182–189.

[4] B.M. Sagar, DR. G. Shobha, DR. P. Ramakanth Kumar "Character Segmentation Algorithms For Kannada Optical Character Recognition", Proceedings of the 2008 International Conference on Wavelet Analysis and Pattern Recognition, Hong Kong, 30-31 Aug. 2008, 978-1-4244-2239-5/08/©2008 IEEE

[5] Vassilis Papavassilioua,b, Themos Stafylakisa "Handwritten document image segmentation into text lines and words", Pattern Recognition 43 (2010) 369 -- 3770031-3203- © 2009 Elsevier Ltd. All rights reserved. doi:10.1016/j.patcog.2009.05.007

[6] Muthukrishnan.R and M.Radha "Edge Detection Techniques For Image Segmentation", International Journal of Computer Science & Information Technology (IJCSIT) Vol 3, No 6, Dec 2011

[7] Vikas J Dongre 1 Vijay H Mankar "Devnagari Document Segmentation Using Histogram Approach " International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.1, No.3, August 2011

[8] Mamatha H R , Srikantamurthy K , " Segmentation of Handwritten Kannada Document", International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868 Foundation of Computer Science FCS, New York, USA Volume 4– No.5,October 2012

[9] N. Anupama, Ch. Rupa & Prof. E. Sreenivasa Reddy "Character Segmentation for Telugu Image Document using Multiple Histogram Projections" Global Journal of Computer Science and Technology Graphics & Vision Volume 13 Issue 5 Version 1.0 Year 2013, Online ISSN: 0975-4172 & Print ISSN: 0975-4350

[10] Ved Agnihotri, "Offline Handwritten Devnagari Script Recognition", IJCSI International Journal of Computer Science Issues, 2012.

[11] R. Jayadevan, Satish R. Kolhe, Pradeep M. Patil and Umapada Pal, —Offline Recognition of Devanagari Script: A Survey in IEEE Transaction on Systems, Man and Cybernetics –Part C. Applications and Review, vol.41, pp-782-796, 2011.

[12] Pooja Kaushik and Yuvraj Sharma "Comparison Of Different Image Enhancement Techniques Based Upon PSNR& MSE", International Journal of Applied Engineering Research, ISSN 0973-4562 Vol.7 No.11 (2012)