Character Recognition of Offline Handwritten Devanagari Script Using Artificial Neural Network

Naveen Malik[1], Aashdeep Singh[2]

[1]CSE, KUK, India, malikn.10@gmail.com
[2]CSE, KUK, India, aashdeepsingh123@gmail.com

Abstract— **Document segmentation is one of the important phases in machine recognition of any language. Correct segmentation of individual symbols decides the exactness of character recognition technique. It is used to partitioned image of a string of characters into sub images of individual symbols by segmenting lines and words. Devnagari is the most accepted script in India. It is used for lettering Hindi, Marathi, Sanskrit and Nepali languages. Moreover, Hindi is the third most accepted language in the world. Devnagari documents consist of vowels, consonants and various modifiers. Hence perfect segmentation of Devnagari word is challenging. In this paper a bounded box method for segmentation of documents lines, words and characters and proper recognition of Devanagari characters using variation of Gradient, Structural features and artificial neural network (ANN) is proposed.**

Keywords- **Character Segmentation, Character recognition, OCR System, Properties of Devanagari**;

1. INTRODUCTION

Machine simulations for human reading has become a serious topic of research while the introduction of digital computers. The main reason for such an attempt was not only the challenges in simulating human reading but also the possibility of competent applications in which the data present on paper documents has to be transferred into machine-readable format. Devanagari Script is the script used for writing many official languages in India, such as Hindi, Sindhi, Nepali, Marathi, Sanskrit, and Konkani, where Hindi is the national language of the country [1]. Devanagari script has 11 vowels ('svar') shown in Figure.1 and 33 consonants ('vyanjan') shown in Figure.2 [2].

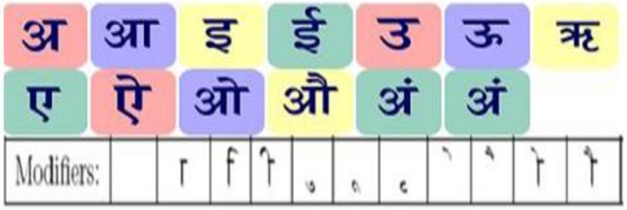
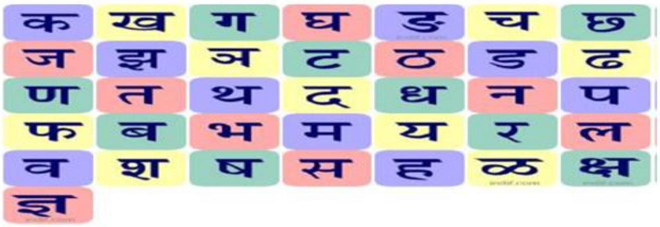

अ आ इ ई उ ऊ ऋ
ए ऐ ओ औ अं अं
Modifiers:
ा ि ी ु ू ृ े ै ो ौ

Figure 1


क ख ग घ ङ च छ
ज झ ञ ट ठ ड ढ
ण त थ द ध न प
फ ब भ म य र ल
व श ष स ह ळ क्ष
ज्ञ

. Figure 2.

There are some special combinations in which a new character or the half forms of consonants may appear in the lower half of the new complex forms. Another unique feature of Devanagari is the presence of a horizontal line on the top of all characters. This line is recognized as header line or "shirorekha"[3]. The words can be separated into three strips: top, core, and bottom, as given in Figure. 3.

1.2 Properties of Devanagari Script:

Devanagari Script is considered as principal script in India, it is utilized inside Hindi, Sindhi, Nepali, Konkani, Marathi, and Sanskrit languages. Center area holds vowels, consonants or synthesis of both and the upper & lower areas hold vowels, modifiers and their parts. Vowels could be formed as free or by using different diacritical imprints, i.e. modifiers or 'matras', which are created above, below, before or after the consonant or vowel[4].

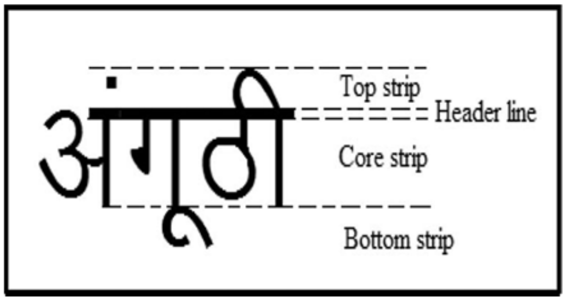


Figure.3: Three strips of a word in Devanagari script

1.3 Character Segmentation:

It is a crucial phase for recognition process, so successful segmentation brings about better detection. Segmentation of Offline Handwritten content is more entangled because of varieties in writing styles of distinctive users, furthermore the complex structure of Hindi characters.

1.4 Character recognition:

Is considered as an important technology for today's world and it is used in many fields such as artificial intelligence, computer vision, pattern matching etc. There are two kinds of character recognition systems.

1.4.1 Optical character recognition: -

It is also known as offline character recognition. In this type either type written, handwritten, or printed text is converted into digital format. It does not have any advantage of recognizing the direction of the movements while writing the character [3].

1.5 OCR System:

1.5.1 Digitization/Data acquisition:

It is the process of converting paper document into electronic form. For this process, handwritten documents are scanned hence an image is processed. This image is fed in to the next preprocessing stage.

1.5.2. Pre-processing:

Pre-processing is the primary stage of character recognition. It includes the following stages:

1.5.2.1 RGB to Gray scale conversion:

The scanned image is stored as BMP, JPEG, and TIFF etc which is in RGB format. Now these images need to be converted into a gray scale image. A gray scale image represent an image in the form of matrix where every element has a value equivalent to how bright or dark the pixel at the corresponding position should be colored.

1.5.2.2 Thresholding/Binarization:

Binarization is a process which uses thresholding procedure to change a gray scale figure into a binary image. Thresholding reduces the storage space requirement and increase the rate of processing by converting the gray scale image into binary image using a threshold value.

1.5.2.3 Noise reduction:

It is introduced by optical scanning devices that causes disconnected line segment, gaps and bumps in lines. These distortions include rounding of corners, local variations, dilation and erosion etc. It is essential to eliminate the limitations. Noise Reduction techniques can be categorized as (a) Filtering (b) Morphological operation (c) Noise modeling [3].

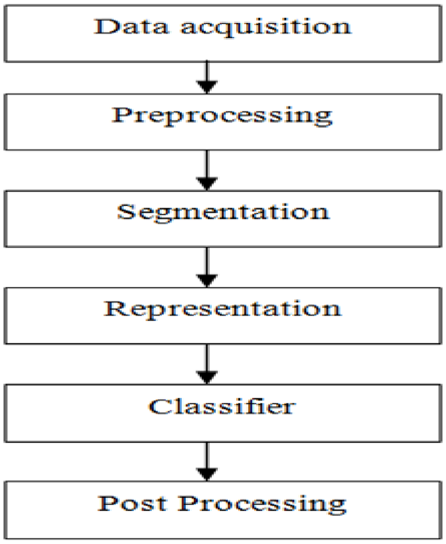


Figure 4. Optical Character Recognition (OCR) consists of following phases [3]

1.5.3. Segmentation:

Segmentation is used to partition an image into various regions. Segmentation is important part of Character recognition system. The process of segmentation works in the following pattern:

i. Identify the text line in the pages.
ii. Identify the word in individual line.
iii. Finally identify individual character in each word**.**

1.5.4. Feature extraction/Representation:

Next step is to extract features from the segmented characters and to differentiate it from other character by comparing it with already stored patterns of all the characters in the library. There are various features used for character recognition. Key features may include height, width, density, loops, lines, stems and other character traits. Chain code histogram feature is used for recognition that is extracted by chain coding the contour points of the scaled character bitmapped image [6].

1.5.5 Applications of OCR:

An OCR can convert the text from the image into text that can be easily edited on the computer. Subsequent are the applications of OCR.

1. Automatic text entry into the computer for desktop publication, library cataloguing, ledgering, etc.
2. Routine reading for sorting of postal mail, bank cheques, postal code reading, business forms reading government records, manuscripts and their archival and other documents,
3. Document facts compression: from document image to ASCII layout,
4. Language processing such as indexing, spell checking, grammar checking, etc.,

5. Multi-media system design, etc [7].

2. RELATED WORK

The research work performed in this area by different researchers is presented as follows:

R. Jayadevanet et al.in 2011[1] More than 300 million people in India use Devanagari script for documentation. There has been a significant progress in the research related to the character recognition of handwritten as well as printed Devanagari text in the last few years. From 1970s state of the art of machine printed and handwritten Devanagari optical character recognition is discussed in this paper.All feature-extraction as well as classification, training and matching techniques helpful for the recognition are discussed here in various sections of the paper. An effort is made to address the most important results reported so far and it is also tried to bring to light the advantageous directions of the research till date. Besides, the paper also contains a comprehensive bibliography of many selected papers appeared in reputed journals and conference proceedings as support for the researchers working in the field of Devanagari OCR.

Dr. Latesh Malik et al. in 2012[2] A holistic graph based approach for handwritten Devanagari Optical Character Recognition is presented as piece of a knowledge-based word interpretation model. This latest method is based on the recognition of sub graphs homeomorphism to formerly defined prototypes of words. In this system the feature graph is uprooted from input word. Then it is extended to net. Net is formed by creating new nodes. Each sub graph that is known is introduced as a node in a directed net that compile different alternatives of interpretation of features in the feature graph. A path in the net represent a reliable sequence of characters in the word. A final search for optimal path under definite criterion gives the best analysis of the word features. The word level accuracy is 84% on the test set. Further the accuracy can be improved with the use of lexical situation to find a meaningful interpretation.

Naresh Kumar Garget al. in 2010[3] the chief purpose of this paper is to offer the new segmentation techniques on the basis of structure technique for Handwritten Hindi text. Segmentation is one of the most important stages of character recognition. The handwritten text is divided into lines then the lines into words and words into characters. The errors in segmentation spread to recognition. The evaluation of performance on handwritten data of 1380 words of 200 lines written by 15 different writers. The overall results of segmentation are very capable.

Saiprakash Palakollu et al.in 2012[4] In this paper we mainly deal with the new methods for line segmentation and character segmentation of overlapping characters of Handwritten Hindi text. The segmentation of text into lines is done and then these lines into words and after that from lines words header lines are detected and transformed as straight lines. Each word is splitted into three parts the upper modifier, consonant and the lower modifier, so that character segmentation became simple. The algorithm is finding the header and base lines by estimating the average line height and rely on it. This algorithm works proficiently on overlapped characters for different text sizes and different resolutions images.

Vedgupt Saraf et al. in 2013[5] Character recognition is an electronic or mechanical translation of scanned images of typewritten or printed, handwritten text into machine encoded text. More than 300 million people use Devanagari script in India for documentation. There has been a considerable enhancement in the research linked to the recognition of handwritten as well as printed Devanagari text in the past few years. In Devanagari script the problem arises in character recognition using quadratic classifier provides less correctness and less efficiency. For this problem and to get better efficiency we use the genetic algorithm. It will give much better results from the above methods. The idea of using genetic algorithm comes from the fact that it can be used as an exceptional means of integrating various styles of writing a character and generates new styles. By closely observing the ability of human mind in the detection of handwriting author find that humans are capable to recognizing characters even though they might be seeing that mode for the first time. This is probable due to their power to visualize parts of the known styles into the unidentified character. We try to signify the same power into the machines.

U. Pal, T. Wakabayashi, F. Kimura et.al in 2009[6] In recent years the interest in research towards handwritten Indian character recognition is increasing. A lot of approaches have been estimated by the researchers in the field of handwritten Indian character recognition and there are many recognition systems made for isolated handwritten characters and numerals that are presented in the literature. To get an idea of the recognition results of various classifiers and to provide new standard for upcoming research, this paper provides a relative study of Devnagari handwritten character recognition using twelve different classifiers and four set of feature is presented. Projection distance, linear discriminant function, subspace method, support vector machines, mirror image learning, modified quadratic discriminant function, Euclidean distance, nearest neighbour, k-Nearest neighbour, compound projection distance, modified projection distance and compound modified quadratic discriminant function are used as diverse classifiers. Feature sets used in the classifiers are computed based on gradient and curvature information procured from binary as well as gray-scale images.

N. Sharma et.al in 2006[7] Recognition of handwritten characters is not an easy task because of the changeability involved in the writing patterns of different individuals. In this paper we plan a quadratic classifier based scheme for the recognition of Devanagari offline handwritten characters. The characteristics used in the classifier are acquired from the directional chain code information of the outline or contour points of the characters. The bounding box of a character is divided into blocks and the chain code histogram is calculated in each of the blocks. On the basis of the chain code histogram there are 64 dimensional features used for recognition. These chain code features feed to the quadratic classifier for the process of recognition. From this projected scheme author have obtained 98.86% and 80.36% recognition accuracy on Devanagari characters and numerals respectively. Fivefold cross-validation technique is used for result calculation

Gaurav Y. Tawde et al.in 2013[8] Optical Character Recognition is an appealing and demanding field of research in pattern recognition, artificial intelligence and machine vision. And it is used in many real life applications. It is a type of document analysis in which a scanned document image that contains either printed or handwritten script is given input to an OCR software engine is converted into editable, machine-readable digital text format. With the spread of computers in public and private sectors and at individual homes mechanical processing of tabular application forms, bank cheques, census forms, tax forms and postal mails has gained importance.Such automation needs development and research of handwritten characters and numerals recognition for different scripts or languages. The field of OCR is categorized into two parts first is recognition of printed characters by machine and the second is detection of handwritten characters. Recognizing handwritten text is a significant area of research due to its various application potentials. Feature extraction is the chief step of the process of OCR. This document gives an examination of relative study of different feature extraction techniques that are used in OCR.

3. PROPOSED WORK

3.1Problem Formulation

Developing the Devanagari Handwritten character recognition and segmentation system has some greater challenge because of the following reasons:-**The character set is very large:**In Devanagari the vowel, consonants, matra, Chandra Bindu, Visarg and many more different symbols are present. **Similarity between the characters is high:**Many similarities in shapes so it is difficult to segment the character and recognize the desired result. **Problem due to shirorekha:**All the characters have a horizontal line at the upper part, known as Shirorekha. In continuous handwriting, from left to right bearing, the shirorekha of one character joins with the shirorekha of the previous or next character of the alike word

3.2Proposed Work

Communication with the machine can be done in many ways and one of them is handwriting. With the help of handwritten input we can give input to the computer. Reason to choose the Devanagari is, it is on the third of the most spoken languages.

To propose a bounded box method for segmentation of documents lines, words and characters. To expand the application which take offline handwritten input and perform proper recognition of Devanagari characters using variation of Gradient, Structural features and artificial neural network (ANN)

4. RESULTS AND ANALYSIS

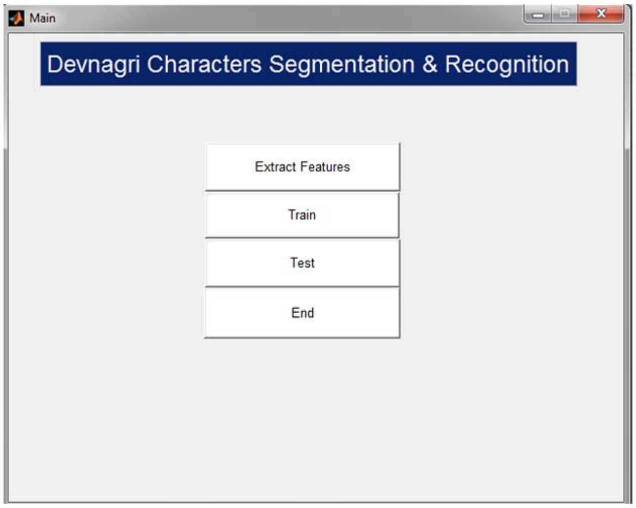
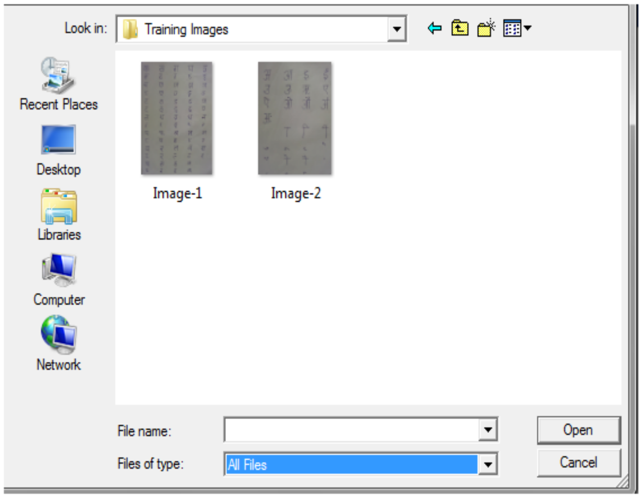


Figure 4.1 GUI of Character Segmentation and Recognition

Figure 4.1 shows the graphical user interface. Various options are Extract features, Training and Testing.


Look in: Training Images

Files of type: All Files

Figure 4.2 Browsing Training Images

Figure 4.2 shows the folder containing images which are used to extract features

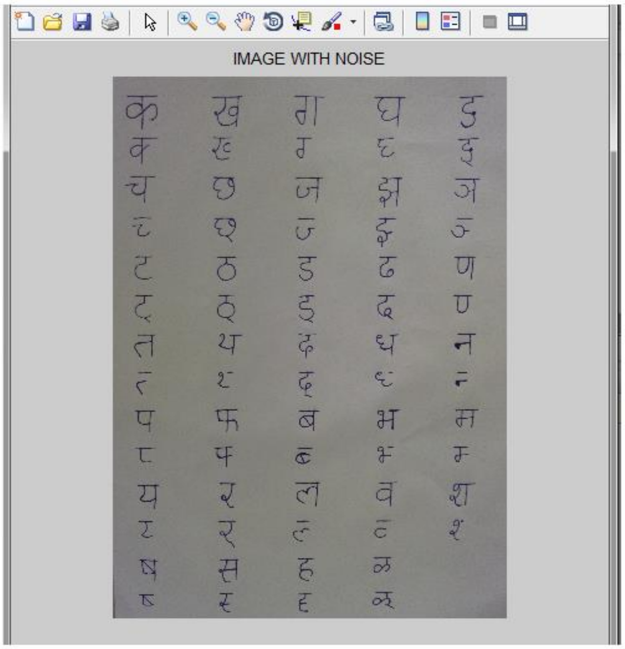


Figure 4.3 Noisy Image

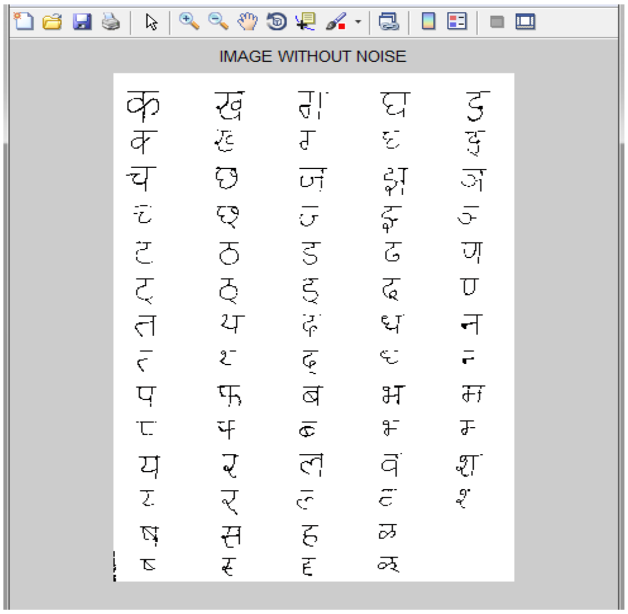


Figure 4.4 Image without Noise

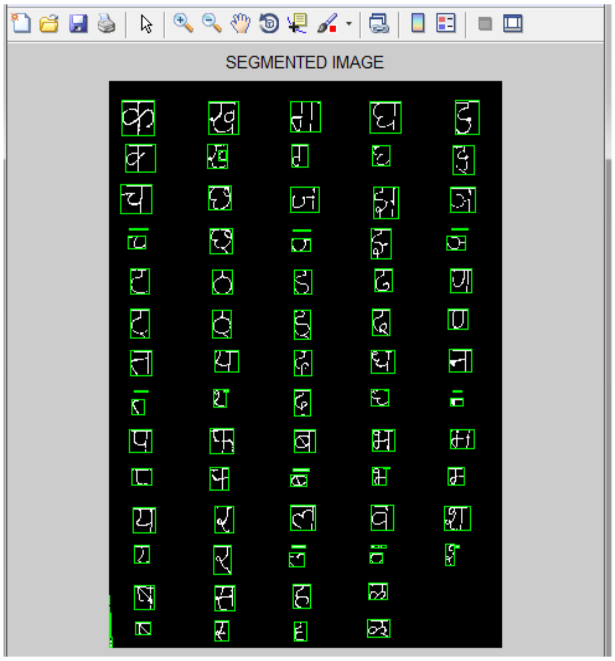
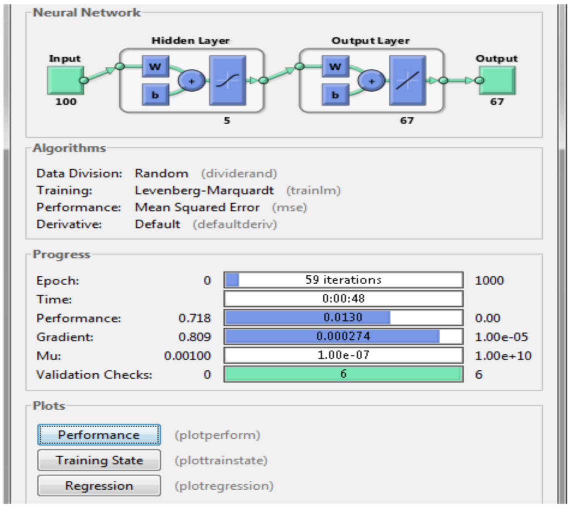


Figure 4.5 Segmented Image


Neural Network
Input
100
Hidden Layer
W
b
5
Output Layer
W
b
67
Output
67

Algorithms

Data Division: Random (dividerand)
Training: Levenberg-Marquardt (trainlm)
Performance: Mean Squared Error (mse)
Derivative: Default (defaultderiv)

Progress

Epoch:	0	59 iterations	1000
Time:		0:00:48	
Performance:	0.718	0.0130	0.00
Gradient:	0.809	0.000274	1.00e-05
Mu:	0.00100	1.00e-07	1.00e+10
Validation Checks:	0	6	6

Plots
Performance (plotperform)
Training State (plottrainstate)
Regression (plotregression)

Figure 4.6 Simulation using Neural Network

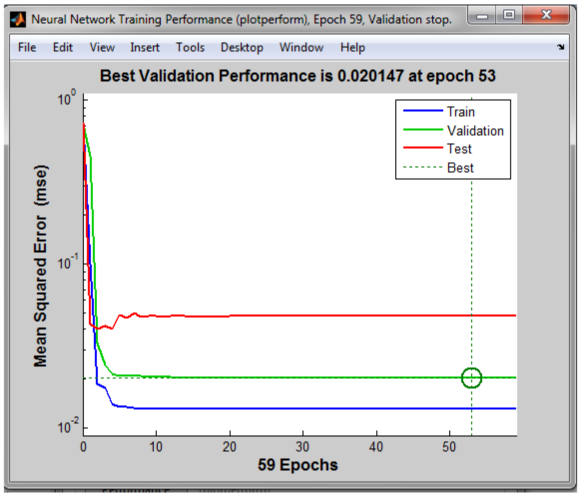
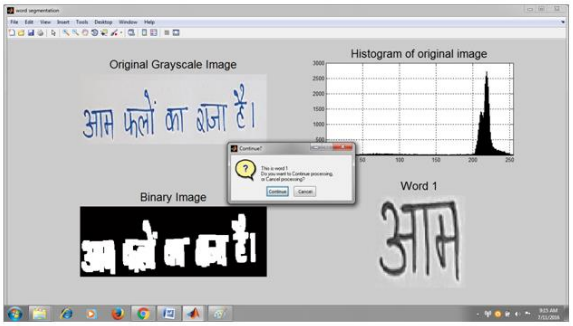
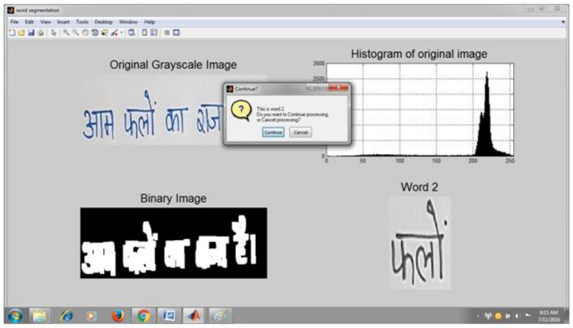


Figure 4.7 Neural Network performances



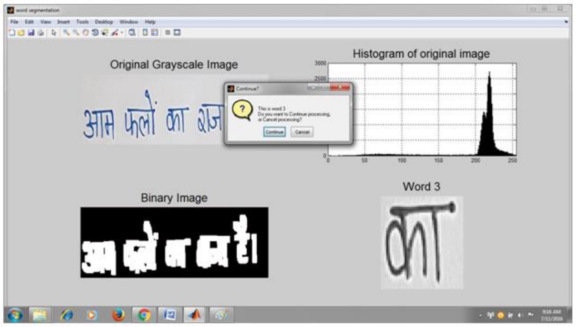
This is word 1
Do you want to Continue processing, or Cancel processing?

Figure 4.8 Segmentation of lines into words



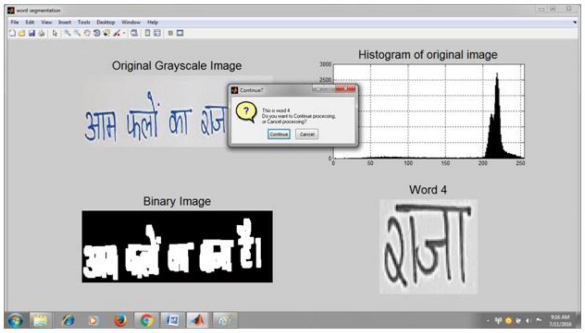
This is word 2
Do you want to Continue processing, or Cancel processing?

Figure 4.9 Segmentation of lines into words



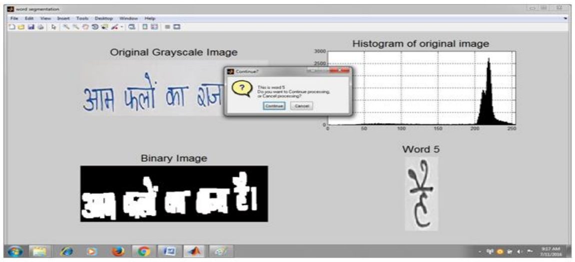
This is word 3
Do you want to Continue processing, or Cancel processing?

Figure 4.10 Segmentation of lines into words



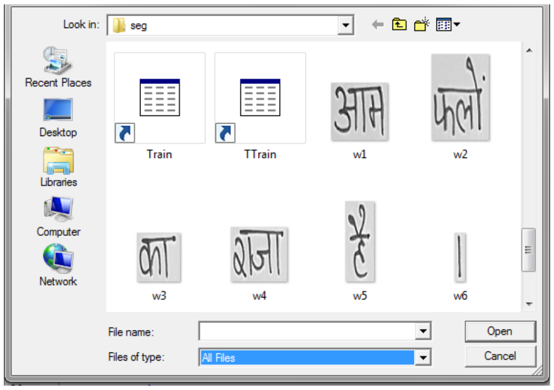
This is word 4
Do you want to Continue processing, or Cancel processing?

Figure 4.11 Segmentation of lines into words



This is word 5
Do you want to Continue processing, or Cancel processing?

Figure 4.12 Segmentation of lines into words


Look in: seg

Files of type: All Files

Figure 4.13 folder containing segmented words

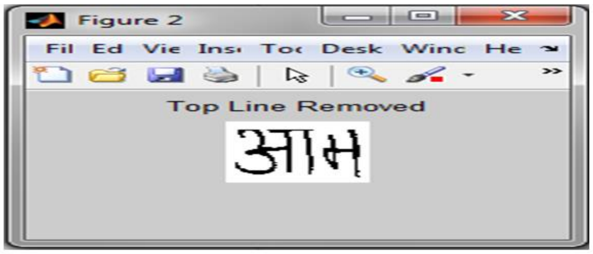


Figure 4.14 Sherorekha Removed

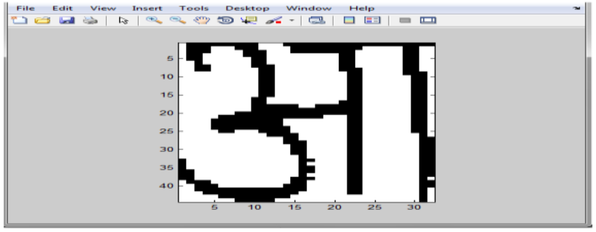


Figure 4.15 Segmentation of words into characters

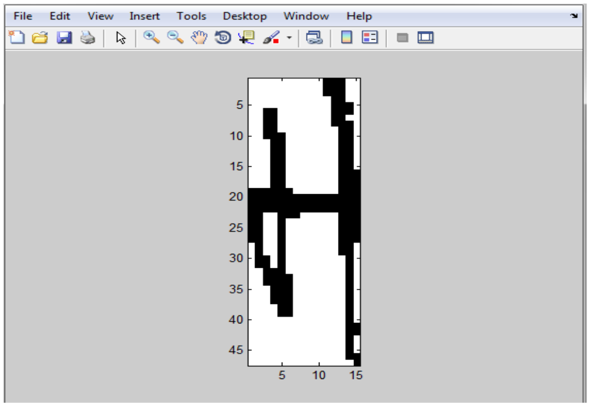
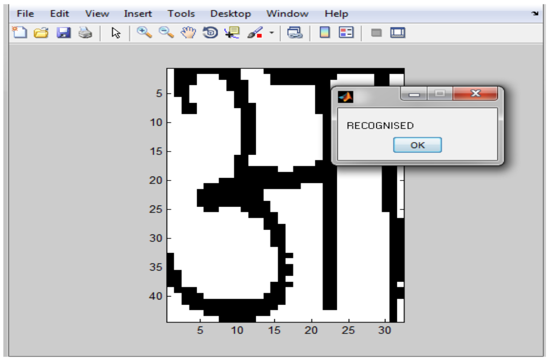
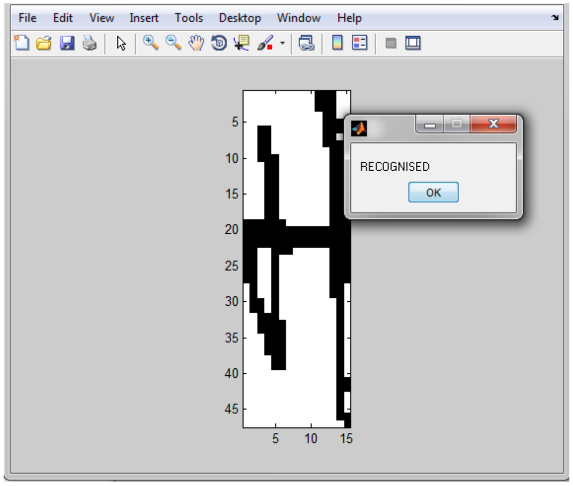


Figure 4.16 Segmentation of words into characters



RECOGNISED

Figure 4.17 Testing Segmented characters



RECOGNISED

Figure 4.18 Testing Segmented characters

Words Segmentation	Words in document	Recognized Words	Accuracy
	60	60	100
Characters Segmentation	Characters in document	Recognized Characters	Accuracy
	167	198	84
Character Recognition	Total Characters	Recognized Characters	Accuracy
	90	82	91.11

Table 4.1 Recognition and Segmentation Accuracy

5. CONCLUSION AND FUTURE SCOPE

In this paper, we have presented a primary effort for segmentation of lines, words and characters of Devnagari script. Nearly 100% efficient segmentation achieved in line and word segmentation but character rank segmentation accuracy is 84% needs more attempt as it is complicated for Devnagari script. The performance of an HCR system hinge on to a great amount on the extracted features. We have used 3 sets of features for comparing the performance of the character recognition system: Projection histograms, Chain code histograms and Histogram of aligned Gradients. Character level recognition accuracy is 91.11%

REFERENCES

[1]. R. Jayadevan, Satish R. Kolhe, Pradeep M. Patil, and U. Pal, “Offline recognition of devanagari script: A Survey”, IEEE Transaction on Systems, Man and Cybernetics-Part C: Applications and Reviews,VOL. 41, No. 6, Nov. 2011.

[2]. Malik, L., G.H. Raisoni "A Graph Based Approach for Handwritten Devanagri Word Recogntion" IEEE,2012.

[3]. Naresh Kumar Garg, Lakhwinder Kaur, M. K. Jindal,” Segmentation of Handwritten Hindi Text”, Volume 1 – No. 4,2010

[4]. Saiprakash Palakollu, Renu Dhir, Rajneesh Rani,”Handwritten Hindi Text Segmentation Techniques for Lines and Characters”, Vol I,2012

[5]. Vedgupt Saraf, D.S. Rao,” Devnagari Script Character Recognition Using Genetic Algorithm for Get Better Efficiency”, Volume-2, Issue-4, April 2013

[6]. U. Pal, T. Wakabayashi, F. Kimura,”Comparative Study of Devnagari Handwritten Character Recognition using Different Feature and Classifiers”,2009

[7]. N. Sharma, U. Pal, F. Kimura, and S. Pal,” Recognition of Off-Line Handwritten Devnagari Characters Using Quadratic Classifier”, pp. 805 – 816, 2006.

[8]. Gaurav Y. Tawde , Mrs. Jayashree M. Kundargi”,An Overview of Feature Extraction Techniques in OCR for Indian Scripts Focused on Offline Handwriting, Vol. 3, Issue 1, January -February 2013, pp.919-926.