IJETAE
EXPLORING RESEARCH AND INNOVATIONS

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 5, Issue 8, August 2015)

Handwritten Text Extraction from Documents.

Nikita V Borse[1], Prof. I. R. Shaikh[2]

[1,2]*Department of CSE, S.N.D.E.R.C Yeola, Dist-Nasik, Savitribai Phule University of Pune, Maharashtra, India.*

***Abstract*— Handwritten character recognition and extraction is an most important part of image processing and pattern recognition. Handwritten Devnagari Characters are more difficult for recognition than any other language or English characters. Devnagari character has so many possible variations in order, direction and shape of the constituent strokes. Each and every person has its own handwriting style. Those are written in very bad manner, such words cannot be easily read by a machine. Due to very bad styles of writing, a lot of difficulties are faced in recognition process.**

We have developed a system to extract text from handwritten documents. We used 2 languages Marathi and English language. The main purpose is to introduce a method for recognition and extraction of handwritten Devnagari characters using segmentation and recognition. The whole process of text extraction includes main phases- segmentation of characters into line, lines into words and words into characters and then recognition through neural network.

***Keywords*— Marathi and English Handwritten scan images, Binarization, Segmentation, Feature extraction method, Trained Dataset.**

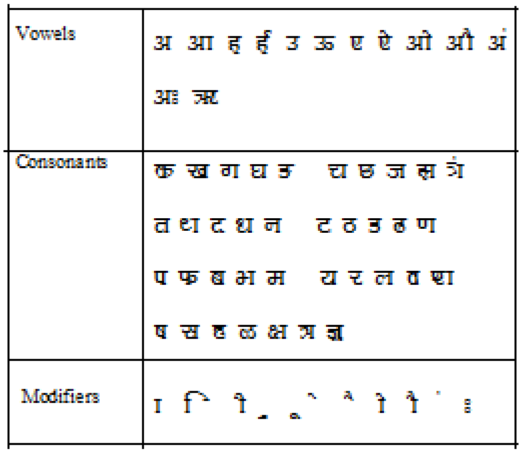
I. INTRODUCTION

Text extraction in document images is an important phase for various document image processing tasks such as layout analysis and optical character recognition. Therefore, there have been so many researches in this area, and lot of algorithms has been proposed for the extraction of text-lines in machine-printed document images. However, text extraction in handwritten documents is still considered a challenging problem because the scale and orientation of characters are spatially varying, inter-line distances are irregular, and characters may touch across words and/or text-lines. Handwriting recognition is comparatively difficult, because different people have different handwriting style.

We are using Marathi and English language for extraction of text. Marathi script derived from Devnagari is an official language of Maharashtra. Marathi script consists of 13 vowels and 37 consonants making 50 alphabets. Marathi is written from left to right. It has no capital and small case characters. There is a horizontal line at the top of character called as the header line. The header line joins the characters to make a word.

Vowels can be written as independent letters, or by using a variety of diacritical marks which are written above, below, before or after the consonant they belong to. When vowels are written in this way they are known as modifiers and the characters so formed are called conjuncts. Sometimes two or more consonants can combine and take new shapes. These new shape clusters are known as compound characters.

Table I:
Marathi Characters



Vowels	अ आ इ ई उ ऊ ए ऐ ओ औ अं अः ऋ
Consonants	क ख ग घ ङ च छ ज झ ञ त थ द ध न ट ठ ड ढ ण प फ ब भ म य र ल व श ष स ह ळ क्ष त्र ज्ञ
Modifiers	ा ि ी ु ू े ै ो ौ ं ः

In Handwritten character recognition, accuracy of character recognition mostly depends on accuracy of segmentation. If segmentation gets correct output or segmented characters properly then it will give easy and accurate text extraction. Incorrect segmentation gets incorrect character recognition process. Segmentation process divides text into line, word, and character segmentation. Text extraction in digital image processing is a difficult area for handwritten document analysis and character recognition. These problems are common in handwritten documents as compared to printed documents because of every individual person have its handwriting styles. Researchers are continuously solving these problems for different languages.

Text extraction in handwritten documents is an essential step for document image understanding, we develop a language-independent handwritten text extraction system.

Our method works only on handwritten text, so finally we translate the handwritten text image into digital text documents.

II. Related Work

The various difficulty related with text extraction are mainly analyzed in the literature survey and these help the researchers to understand and carry out the work further in this field.

Different techniques has been reported for text extraction methods from handwritten documents such as projection profiles, Hough transform, Used KNN classifier, SVM technique and many others.

Nicolaou et al. (2009) proposed technique to segment handwritten document images into text lines by shredding their surface with local minima tracer. It is assumed that there exists a path from one side of the image to other that traverses only one text line. Image is blurred first and then uses tracers to follow the white-most and black-most paths from both left to right and right to left direction in order to shred the image into text line areas.

G. Louloudis et al. (2008) presented a text line detection method for handwritten documents. The proposed technique is based on a approach that consists of three distinct steps. The first step includes image pre-processing and connected component extraction, division of the connected component domain into three spatial sub-domains and average character height estimation. Secondly, author used a block-based Hough transform for the detection of potential text lines while third step is to correct feasible splitting, to detect text lines that the previous step did not expose and, finally, to disconnect vertically connected characters and assigns them to text lines.

Yi Li et al. (2008) proposed an approach based on density estimation and a state-of-the-art image segmentation technique, the level set method. A probability map is estimated from an input document image where each element represents the probability of the underlying pixel belonging to a text line. Then level set method is developed to determine the boundary of neighboring text lines by evolving an initial estimate.

N. Sharma, U. Pal, F. Kimura, and S. Pal have proposed a quadratic classifier based scheme for the recognition of offline Devnagari handwritten characters. The features used in the classifier are obtained from the directional chain code information of the contour points of the characters. The bounding box of a character is segmented into blocks and the chain code histogram is computed in each of the blocks.

Based on the chain code histogram, 64 dimensional features are used for recognition. These chain code features are fed to the quadratic classifier for recognition. In this system fivefold cross-validation technique is used for result computation.

Shreya N. Patankar, Leena R. Ragha, proposed Zonal moments based Handwritten Marathi Barakhadi recognition. They recognise a Marathi Barakhadi character by recognising the vowel and consonant parts separately.

III. Our Approach

We proposed a language-independent text extraction algorithm for the processing of Marathi handwritten document images as well as English. We divide under-segmented CCs into Line, Words and Characters so that we can have better representations for text components. We have implemented a system that will trace out trained data from input file. Input file is consisting of handwritten text only. This image first to be scan and then used as input. After input image is taken then image is converted into gray scale image, then remove noise from that image. Binarization takes place on that noise free image. We get binarized image. After that Segmentation Algorithm is applied. Text is separated using CC segment into lines, words and characters. We are using Feature extraction algorithm. We extract features from each and every character and these features store into its specified folder. Finally we get the accurate result.

IV. System Design

The fig. 1 shows fundamental phases involved in handwritten character recognition system.

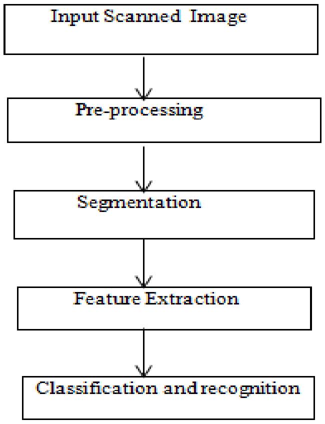


Figure I: Handwritten Character Recognition.

A. Input Scanned Image

The input handwritten images are collected and passed through a scanner. Whenever an image is acquired, there will be some variations and noise gets added to the image. Hence pre-processing is required for adjusting the intensity levels, improve the quality of image and to de-noise the image.

B. Pre-Processing

Pre-processing is the most active part of a better performance of system. In this stage, the input image is processed to remove any noise. A coloured image then it will be converted in to a gray image. The noised free image is then converted to a binary image. Gray scale image are converted into binary image using threshold value removal of noise having less than 30 pixels, Binarization can be made simple and more accurate. Pre-processing aims to produce data that are easy for the HCR system to operate accurately. After pre-processing phase, a cleaned image is available that goes to the segmentation phase.

C. Segmentation

Images are segmented into line, word and character for the given pre-processing input image.

- Line segmentation -To segmenting the text image into Lines, compute horizontal projection profiles.
- Word segmentation -To segmenting the text-line into words, so need compute vertical projection profiles.
- Character Segmentation - First horizontal scan of individual word is computed we get the rows having highest black pixel or highest projection is consider as a header line and removed for further character segmentation. After removal vertical scanning of individual character is computed.

D. Feature Extraction

The selection of good feature set is the most the important aspect of handwritten character recognition. This method provides the ease of implementation and good recognition. We have computed the centroid of image. For feature extraction we will use Zone based approach which is work with image centroid zone. In image centroid zone character is divided into n equal zone or grid and then computed image centroid and then calculate average distance from character centroid to each zones/grid/boxes present in image.

V. ALGORITHMIC STRATEGY TO BE USED

For feature extraction Image centroid Zone (ICZ) based Distance metric feature extraction system is used. Algorithm for that is as follows -

Algorithm1- Image centroid Zone feature extraction Algorithm

Input: Image (Character/Numeral) is Pre-processed.

Output: Extract Features for Classification and Recognition.

Algorithm:

```
Step 1: Compute the input image centroid.
Step 2: Divide the input image into n equal zones.
Step 3: Compute the distance between the image centroid
        to each pixel present in the zone.
Step 4: Repeat the step 3 for the entire pixel present in the
        zone.
Step 5: compute average distance between these points.
Step 6: Repeat this procedure sequentially for the entire
        zone.
Step7:  Finally n such features will be obtained for
        classification and recognition
End
```

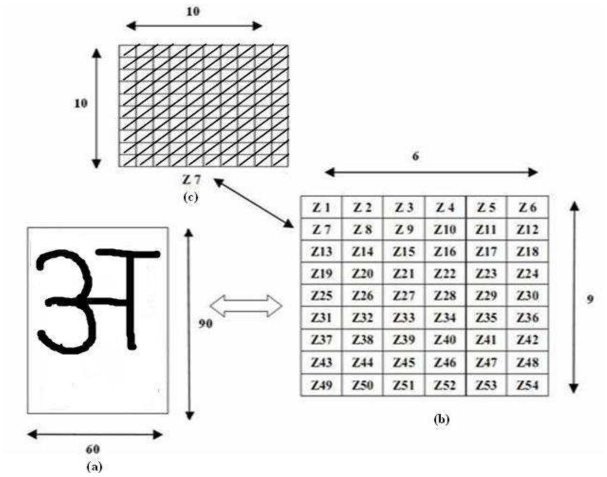


Figure II: Feature Extraction from Devnagari Marathi Character Image "A"

VI. MATHEMATICAL MODEL

Problem Definition –

1) Scanned Document Image.
2) Image Preprocessing.
3) Segmentation.
4) Feature Extraction.
5) Recognition.

Let the system be described by S,

S= {I, SI, IP, SE, FE, RC}

Where,

S: is a System.

I: Set of Input Image.

SI: Scanned Document Image.

IP: Image Preprocessing.

SE: Segmentation.

FE: Feature Extraction.

RC: Recognition.

Activity-

I= {i1, i2… in}

F= {f1, f2… fn}

Y= {SI, IP, SE, FE, RC}

Where,

I is the set of Input Image.

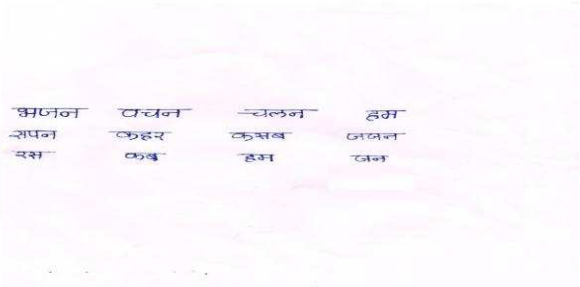
F is set of Functions.

Y is a set of techniques use for Text Extraction Algorithm for Handwritten Documents.

VII. SYSTEM IMPLEMENTATION

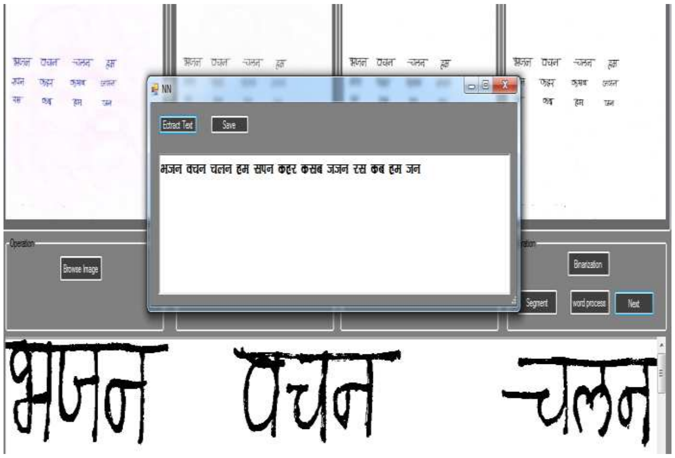
Text extraction from handwritten characters is efficiently carried out with help of this system. We take number of samples of Marathi text as well as English text as input. Handwritten document images are obtained by scanner then these images is passed to the pre-processing stage. In colored or gray images converted into Black and white image, then noise gets removed. Then this cleared binary image passed for segmentation. Here text is segmented into lines, words and character. Then we applied Feature Extraction method, we get the feature these features are stored into its specified folder for extraction, thus we get final result.

If we select a Marathi text as an input as



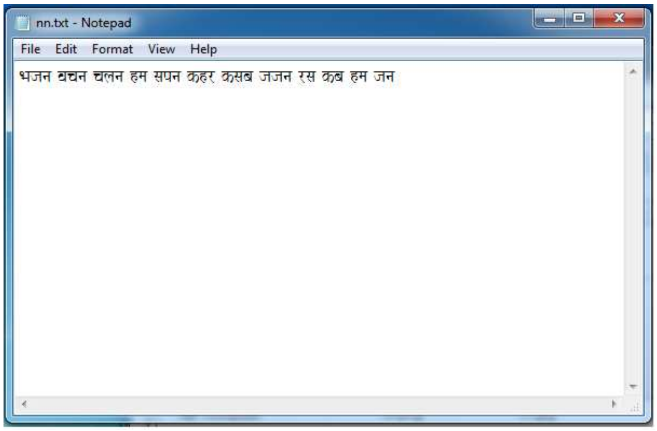
भजन वचन चलन हम
सपन कहर कसब जजन
रस कब हम जन

Figure III: Input image



भजन वचन चलन हम सपन कहर कसब जजन रस कब हम जन

Figure IV: Extraction of text from input image.



भजन वचन चलन हम सपन कहर कसब जजन रस कब हम जन

Figure V: Final result

Thus we get exact and accurate extraction of handwritten text. Thus handwritten input image is converted into text file.

VIII. Performance And Results

We also calculate performance evaluation by using precision and recall. We need to train and test the bulk of character set.

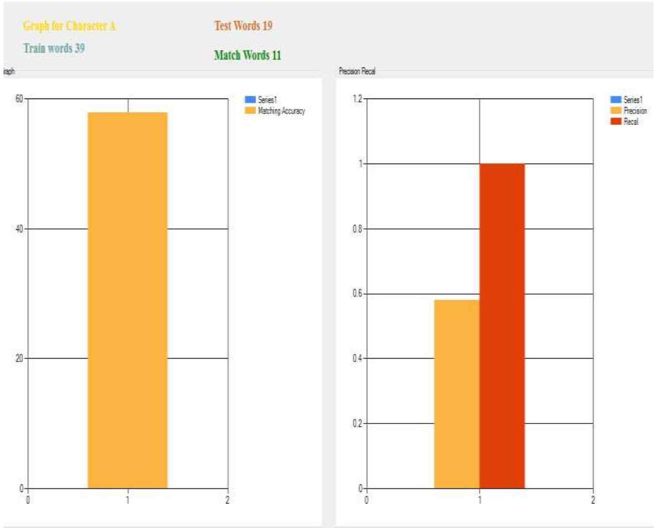


Figure VI: Resultant Graph using Precision and recall.

In above graph we train 39 Character set ‘A’ and Test 19 Characters set that contains only 11 ‘A’ characters and remaining other characters. Finally we get the accurate matching accuracy result and precision and recall.

IX. Conclusion

We have implemented a text-line extraction algorithm for the processing of handwritten document images. We have presented a system for recognizing and extracting a handwritten character. This system works on handwritten Marathi as well as English character recognition. The model starts with pre-processing. The pre-processing stage involves all of the operations to produce a clean character image, so that it is can be used for segmentation and efficiently by the feature extraction stage. Thus finally translate the handwritten text image into text documents with the help of text extraction. Thus this system increased recognition rate for Marathi and English script. Our future work aims to improve classifier for better recognition rate which provides efficient results.

REFERENCES

[1] Jewoong Ryu, Hyung Il Koo, Member, IEEE, and Nam Ik Cho, Senior Member, IEEE “Language-Independent Text-Line Extraction Algorithm for Handwritten Documents” IEEE SIGNAL PROCESSING LETTERS VOL. 21, NO. 9, SEPTEMBER 2014.

[2] A. Nicolaou, B. Gatos, “Handwritten Text Line Segmentation by Shredding Text into its Lines,” 10th International Conference on Document Analysis and Recognition, IEEE Computer society, 2009, pp. 626-630

[3] G. louloudis, B. Gatos, I. Pratikakis, C.Halatsis, ”Text Line Detection in handwritten documents,” Pattern Recognition vol.41, pp. 3758 – 3772, 2008.

[4] Yi Li, Yefeng Zheng, David Doermann, Stefan Jaeger,” Script-Independent Text Line Segmentation in Freestyle Handwritten Documents.” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, no. 8, Aug.2008.

[5] Fei Yin, Cheng-Lin Liu, ”Handwritten Chinese text line segmentation by clustering with distance metric learning,” Pattern Recognition 42, pp. 3146 – 3157, 2009.

[6] N. Otsu. (1979): “A threshold selection method from gray-level histograms”, IEEE transactions on systems, Man and Cybernetics, Vol. Smc-9, No. 1.

[7] Rajean Plamondon, Sargur N. Srihari, “On-Line and Off-Line Handwriting Recognition: A Comprehensive Survey”,Pattern Analysis,Vol.22,No.1, January 2000.

[8] Raid Saabni, Jihad El-Sana, “Language-Independent Text Lines Extraction Using Seam Carving”

[9] B. Gatos, A. Antonacopoulos,N. Stamatopoulos1 Handwritten Segmentation Contest, in Int. Conf. Document Analysis and Recognition (ICDAR), 2007, pp. 626630

[10] Bolan Su, Shijian Lu, and Chew Lim Tan,“Robust Document Image Bi-narization Technique for Degraded Document Images”,Image Processing, vol.22,No.4,April 2013.

[11] S. Bukhari, F. Shafait, and T. Breuel, Text-Line Extraction using a Convolution of Isotropic Gaussian Filter with a Set of Line Filters in Int. Conf. Document Analysis and Recognition (ICDAR), 2009, pp. 446450.

[12] Ram Sarkari, Sougata Halder, Samir Malakar, Nibaran Dasl, Subhadip Basul, Mita Nasipuri,“Text line extraction from handwritten document pages based online contour estimation”, IEEE-20180

[13] Alaei, P. Nagabhushan, and U. Pal, “A new text-line alignment approach based on piece-wise painting algorithm for handwritten documents,”in *Int. Conf. Document Analysis and Recognition (ICDAR)*, 2011, pp. 324–328.

[14] H. I. Koo and N. I. Cho, “Text-line extraction in handwritten Chinese documents based on an energy minimization framework,” *IEEE Trans. Image Process.*, vol. 21, no. 3, pp. 1169–75, Mar. 2012.

[15] T. Stafylakis, V. Papavassiliou, V. Katsouros, and G. Carayannis, “Robust text-line and word segmentation for handwritten documents images,”in *IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2008, pp. 3393–3396.

[16] U. Pal, N. Sharma, T. Wakabayashi, F. Kimura, “Recognition of Off-Line Handwritten Devnagari Characters Using Quadratic Classifier”. ICVGIP 2006, LNCS 4338, pp. 805 -816, 2006.

[17] V. Rajashekararadhya, P. Vanaja ranjan, “Handwritten Numeral/Mixed Numerals Recognition of South Indian: Zonal based Feature Extraction Method”, 2005 - 2008 JATIT.

[18] Mahesh Jangid Kartar Singh, Renu Dhir Rajneesh Rani “Performance Comparison on Devanagari Handwritten Numeral recognition” International Journel of Computer Application (0975-8887) volume-22 No.-1, May 2011 .

[19] Shreya N. Patankar Leena R. Ragha, ”Zonal moments based Handwritten Marathi Barakhadi recognition” In Proc. International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 6, August – 2012.

[20] Vikas J. Dongre, Vijay H Mankar, ―A Review of Research on Devnagari Character, International Journal of Computer Applications (0975 – 8887) Volume 12– No.2, November

[21] Dhaval Salvi, Jun Zhou, Jarrell Waggoner, Song Wang, ”Handwritten Text Segmentation using Average Longest Path Algorithm,” Applications of Computer Vision(WACV), IEEE Workshop, pp. 505-512, 2013.