

Segmentation of Assamese Handwritten Characters based on Projection Profiles

Sagarika Borah
Research Scholar
Tezpur University

ABSTRACT

The most important part of a character recognition system is segmenting the characters properly and selecting the best features from the characters. This paper describes a character segmentation method for an ANN based character recognition system which is used for recognition of optically scanned handwritten Assamese character. The segmentation of characters are done using horizontal and vertical projections of the hand written text document. For feature extraction the system extracts the geometric features of the characters which are consist of basic line types that are used in the formation of the character skeleton. The feature vector of the training set generated by this system is used to train the recognition system using ANN.

Keywords

Direction vector, HCR, Feature vector, zone, starters, intersection points.

1. INTRODUCTION

CHARACTER recognition is an important area in the field of image processing and pattern recognition. Machine simulation of human functions is always been a challenging research area. The main aim of this project is to translate the human handwritings to machine readable. The handwritten character recognition system is of two types off-line and on-line. In case of on-line HCR system the input is taken in pressure sensitive surfaces (digital tablet PCs) from the writer dynamically. On the other hand the off-line system the input is taken in the form of digital image by scanning the handwritten text. Segmentation of text lines is a very difficult job of a hand written character recognition system. It separates the image text documents into lines, words and characters. The image taken as the input for the HCR system is first preprocessed then it is passed through the process of binarization, line segmentation, word segmentation, character segmentation, feature extraction and character recognition.

2. RELATED WORKS

India is a multilingual country of more than 1 billion population with 18 constitutional languages and 10 different scripts. There are a numerous work that is been done on handwritten character recognition. The origin of character recognition was found in 1870 when Carey invented the retina scanner which is an image transmission system using a mosaic of photocells [1]. There is a quality work that is been done towards Indian script. Assamese script is quite similar with the Bangla script. The script is syllabic in nature. It means that text is written using consonants and vowels that together form syllables. There are various classification models like Artificial Neural Network (ANN), Hidden Markov Model (HMM), support vector machines (SVM) etc., for recognition. Most of the pattern recognition techniques used for handwritten character recognition are usually feature

based [3]. The most important aspect of a handwriting recognition scheme is segmentation of the characters from the words properly and the selection of an appropriate feature set which is reasonably invariant with respect to shape variations caused by various writing styles. [6]

Various methods have been proposed based on the global feature extraction approach, of them, the first and foremost was proposed by [7]. The main problem is encountered, while dealing with Assamese characters written by different persons, where the writers represent the same character differently in terms of size and shape. This variation is due to the individuality of the persons, who write the script, apart from the mood and situation of the writer Handwriting text line segmentation approaches can be categorized according to the different strategies used. These strategies are projection based, smearing, grouping, Hough based, graph-based and Cut Text Minimization (CTM) approach [3]. This paper describes a projection-based algorithm proposed in [4] that first obtains an initial set of candidate lines from the piecewise projection profile of the document used in Assamese HCR.

3. METHODOLOGY

A typical HCR system consists of a number of stages:

- 1) *Pre-processing*: The task of preprocessing relates to the removal of noise and variation in handwritten numeric patterns.
- 2) *Segmentation*: Segmentation subdivides an image into its constituent regions or objects.
- 3) *Feature extraction*: This process extracts the geometric features of a numeric which comprises of the basic line types that forms the character skeleton.
- 4) *Training* : In this stage the network is trained with a set of sample of Assamese character collected from different persons at different times.
- 5) *Classification*: During classification the system recognizes a character based on the training database.

A generic view of the proposed system can be shown by the following figure

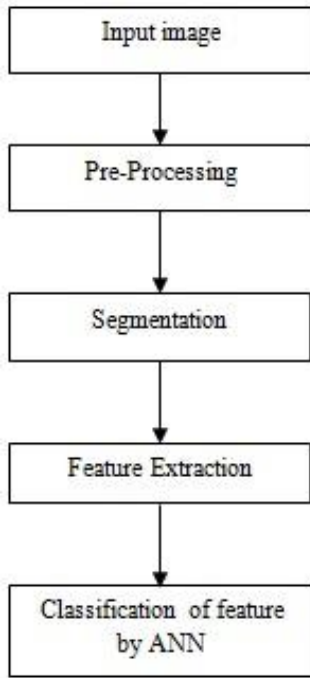


Fig 1: Stepwise overview of the system

A. Overview of the Language script

Assamese is an important language in the North Eastern part of India. There are 40 consonants, 11 vowels, 10 modifiers and over 300 compound characters in the Assamese language. As used in English language there is no use of upper and lower case letters in Assamese. Every individual has a unique style of writing. This style differs from person to person and there different state of mind. The style of writing also changes with the age of the person.

অ	আ	ই	ঈ
উ	ঊ	ঋ	ঌ
এ	ঐ	ও	ঔ
ক	খ	গ	ঘ
ঙ	চ	ছ	জ
ঝ	ট	ঠ	ড
ঢ	ত	থ	দ
ধ	প	ফ	ব
ভ	য	ৰ	ল
ৱ	শ	ষ	স
হ	ক্ষ	য়	ড়
ঢ়	ণ	ৱ	ৱ

Fig 2: Assamese script consonants and vowels

In Assamese script there are some specific characteristics like assamese characters can be divided into three parts horizontally upper, middle and lower parts. Unlike English lines assamese line is ended with a vertical line called as 'dari' and also the upper and middle portion of the divided ba continuous line called as 'matra'.

B. Data collection and pre processing:

Data collection for the training and testing experiment is done on different individuals. A set of samples consisting of about 50 persons were collected. Total 3000 sample characters were collected. Each individual was given an A4 size sheet to write the letters observed in Assamese literature alongwith the numbers from 0 to 9. The collected documents are scanned and the digital images are as binary images.

অ	আ	ই	ঈ	উ	ঊ	ঋ	ঌ	অ	আ	ই	ঈ	উ	ঊ	ঋ	ঌ
এ	ঐ	ও	ঔ					এ	ঐ	ও	ঔ	ন	ব	ল	
ক	খ	গ	ঘ	ঙ	চ	ছ	জ	ক	খ	গ	ঘ	ঙ	চ	ছ	জ
ঝ	ট	ঠ	ড	ঢ	ত	থ	দ	ঝ	ট	ঠ	ড	ঢ	ত	থ	দ
ধ	প	ফ	ব	ভ	য	ৰ	ল	ধ	প	ফ	ব	ভ	য	ৰ	ল
ৱ	শ	ষ	স	হ	ক্ষ	য়	ড়	ৱ	শ	ষ	স	হ	ক্ষ	য়	ড়
ঢ়	ণ	ৱ	ৱ					ঢ়	ণ	ৱ	ৱ				
০	১	২	৩	৪	৫	৬	৭	০	১	২	৩	৪	৫	৬	৭
৮	৯							৮	৯						

Fig 3: Samples collected from the individuals

অই এজন অসমীয়া ।
অসম জাতিৰ উত্তৰ-পূব কোনট অৱস্থিত ।
দিছপুৰ অসমৰ ৰাজধানী ।
অই এজন অসমীয়া ।

Fig 4: Line Samples collected from the individuals

C. Segmentation

Segmentation is one of the most important phases of HCR system. Basically in segmentation, people try to extract basic constituent of the script, which are certainly characters.

Projection technique

The next stage is to extract individual text lines present in the document. In order to extract individual text line, a technique based on projection is used. A projection profile of an image is a histogram giving the number of ON pixels accumulated along parallel lines. Thus a horizontal projection profile is a one-dimensional array where each element denotes the number of ON pixels along a row in the image. Similarly a vertical projection profile gives the column sums [12]. For an uncompressed document of 'm' rows and 'n' columns, the mathematical representation for Vertical Projection Profile (VPP) [18] and Horizontal Projection Profile (HPP) are given below.

$$VPP(y) = \sum_{1 \leq x \leq m} f(x, y) \quad (1)$$

$$HPP(x) = \sum_{1 \leq y \leq n} f(x, y) \quad (2)$$

It is easy to see that one can separate lines by looking for minima in horizontal projection profile of the page and then one can separate words by looking at minima in vertical projection profile of a single line. Such projection profile is used based methods for line, word and character segmentation. To segment the document image into several text lines, the valleys of the horizontal projection computed by a row-wise sum of black pixels. The position between two consecutive horizontal projections where the histogram height is least denotes one boundary line. Using these boundary lines, document image is segmented into several text lines. Similarly, to segment each text line into several text words, the valleys of the vertical projection of each text line are used, which are obtained by computing the column-wise sum of black pixels. The position between two consecutive vertical projections where the histogram height is least denotes one boundary line. Using these boundary lines, every text line is segmented into several text words.

Segmentation of Line: Text lines are detected by horizontal scanning. For segmentation of line, scanned document page horizontally from the top is traversed and find the last row containing all white pixels, before a black pixel is found. Then the first row containing entire white pixel just after the end of black pixels is found out. This process is repeated on entire page to find out all lines.

Word and Character Segmentation: Once the text blocks are detected, the system automatically finds individual text lines, segments the words, and then separates the characters accurately.



Fig 3: Stepwise view of the line segmentation by horizontal projection

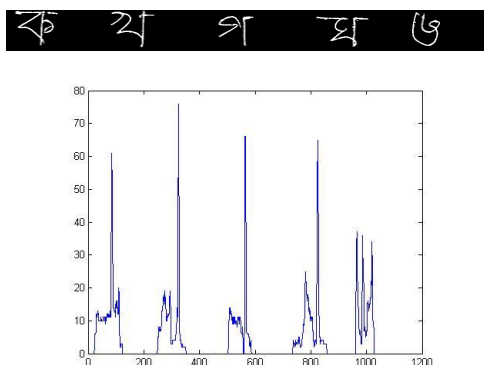


Fig 4: Stepwise view of the letter segmentation by vertical projection of the segmented lines from the document

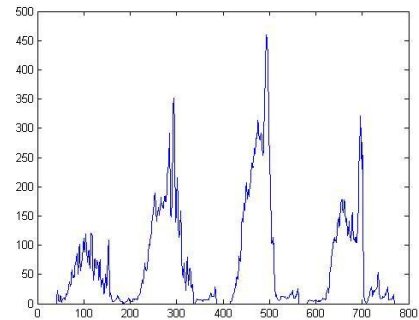
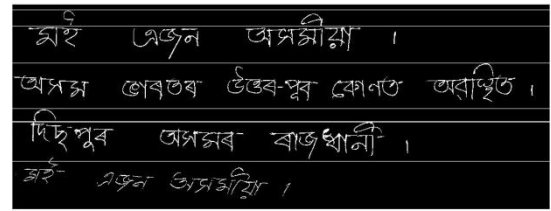


Fig 5: Stepwise view of the line segmentation by horizontal projection of the segmented lines from the document

D. Pre-processing

- **Convert to grayscale:** The colourful RGB image input represented by various colours must be converted to the image with 256 levels of gray scale.
- **Normalization:** Normalization is a process that changes the range of pixel intensity values. Normalization is required as the size of the numeral varies from person to person and even with the same person from time to time. The input numeral image is normalized to size 50x50 after finding the bounding box of each handwritten numeral image.
- **Background noise removal:** The median filter is a nonlinear digital filtering technique, often used to remove noise. It is also useful in preserving edges in an image while reducing random noise.
- **Skeletonisation:** Skeleton (or topological skeleton) of a shape is a thin version of that shape that is equidistant to its boundaries

Input Image	Binarized Image	Normalized image	After noise removal of the image	Skeletonised Image

Fig6: Pre-Processing operations step by step

E. Feature extraction method

For feature extraction geometrical features are extracted which are the basic character image line types and the number of lines appeared in the the zones of the image[2]. So from the set of collected sample letters after cropping and preprocessing is done, the image is divided into windows of

equal size, and the feature extraction is done on individual windows.

i) **Zoning:** For the system implemented, as explained in [5] two types of zoning were used. The character image is divided first into 3 sub images by dividing it vertically in a 1x3 fashion. Then it again divides the original image in a 3x1 fashion and extracts features for each region concatenates this into a single vector. Feature extraction was applied to individual zones rather than the whole image. This gives more information about fine details of character skeleton. The positions of different line segments in a character skeleton become a feature if zoning is used. This is because, a particular line segment of a character occurs in a particular zone in almost cases. To extract different line segments in a particular zone, the entire skeleton in that zone should be traversed. During this traversal some feature elements like starting points of a particular line segment in a zone or the intersection points.

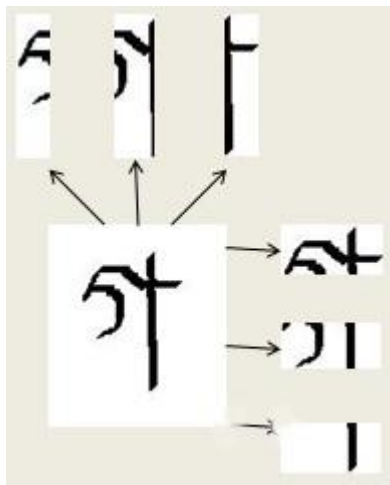


Fig7: extracted zones of a particular letter 'ga'

ii) **Character traversal:** After the letter is zoned for each zone the line segments in there are found out. First the starting points and intersections in the zone are found and then stored in a list. Starters are the pixels having only one neighbor pixel. Then the necessary criterion for a pixel to be a intersection is that it should have more than one neighbors. For this, neighboring pixels are classified into two categories, direct pixels and diagonal pixels. Direct pixels are all those pixels in the neighborhood of the pixel under consideration in the horizontal and vertical directions. Diagonal pixels are the remaining pixels in the neighborhood which are in a diagonal direction to the pixel under consideration. Based on the number of neighboring points the intersection point can be a 3-neighbouring point, a 4-neighbouring point or a 5-neighbouring point. Once all the starting points are processed, the minor starting point obtained so is processed. After that, the algorithm starts with the minor starting points. All the line segments obtained during this process are stored, with the positions of pixels in each line segment. Once all the pixels in the image are visited, the algorithm stops.



Fig8: The starting points of the letter 'ga'

A direction vector is extracted from each line segment which will help in determining each line type.

4	5	6
3	c	7
2	1	8

Fig 9. a 3x3 matrix representation

The line types and the segments are populated by traversing the letter from top to bottom. The algorithm will start with one starter and search out for intersections whenever Any intersection is reached, it will stop and search for the minor starter. And so declares a Line segment. For example in fig7 there is a 5x5 pixel matrix. Here suppose a line segment through (1,1),(2,2),(3,3) ,is present to find this line the algorithm starts from populating the starter pixels, here The starters list consists of (1,1) , (1,5) ,(5,1) , (5,5). Now the algorithm starts the traversal from the pixel (1,1) and checks for the neighboring on pixels and stops when it finds a intersection Pixel (3,3) is a intersection, so the algorithm stops the current segment, and declares all the neighbours as minor starters. So the minor starters list would contain [(4,2),(2,4),(4,4)].

0	0	0	0	1
0	1	0	1	0
0	0	1	0	0
0	1	0	1	0
1	0	0	0	1

Fig10. Line traversal example

iii) **Feature vector formation:** Every zone has a feature vector corresponding to it. From each zone 9 feature element is extracted and from a particular letter get 6 more feature element are generated. So together for a particular image our feature vector will contain 60 feature elements. The contents of each zone feature vector are

1. Number of horizontal lines.
2. Number of vertical lines.
3. Number of Right diagonal lines.
4. Number of Left diagonal lines.
5. Normalized Length of all horizontal lines.
6. Normalized Length of all vertical lines.
7. Normalized Length of all right diagonal lines.
8. Normalized Length of all left diagonal lines.
9. Normalized Area of the Skeleton.
10. Euler Number
11. Filled area
12. Eccentricity
13. Extent
14. Orientation

Besides the line numbers based on their types there are some other features that are included in the feature vector.

a) Euler number: The number of objects in the region minus the number of holes in those objects.

b) Filled area: This specifies the number of on pixels in Filled Image.

c) Eccentricity: It can be thought of as a measure of how much the conic section deviates from being circular. In particular, the eccentricity of a circle is zero.

d) Extent: This is the ratio of pixels in the region to pixels in the total bounding box, computed as the Area divided by the area of the bounding box

e) Orientation: It is the angle (in degrees ranging from -90 to 90 degrees) between the x-axis and the major axis of the ellipse that has the same second-moments as the region.

Normalized length of any particular line type is found using the following method,

$$\text{length} = \frac{\text{Total Pixels in that line type}}{\text{Total zone pixels}} \quad (1)$$

The feature vector explained here is extracted individually for each zone. So if there are N zones, there will be 14N elements in feature vector for each zone.

F. Training & Testing with ANN

Artificial neural network is used in the process of training and testing. Total of 3000 samples were collected from 50 different persons from which 400 samples are scanned and trained. The remaining letters are scanned and used for testing the trained network. For each character there are 10 samples from different person. Feed Forward back propagation model is used for the training. The network consists of 59 feature vector for each input node (corresponding to one feature in each of the 6 zones), 80 neurons in the hidden layer.

4. RESULTS

Each feature vector consisting of 59 element where the first 54 features are obtained from The 6 zones each zone having 9 elements. The last 5th feature element is the Euler number for the image and the other 56th, 57th, 58th, 59th feature elements are filled area, eccentricity, extent and orientation respectively.

At first the method is tried with the trained samples only while testing ; the trained sample are tested and all of them were recognized 100% correctly.

5. CONCLUSION AND FUTURE WORK

This paper describes a segmentation and feature extraction method that can be used to extract features of handwritten scripts. The proposed feature extraction method is implemented for the first time in Assamese script with ANN as classifier. The method was tested with the network using 10 samples for each character and obtained an average of 78% accuracy. As for now the system is tested with isolated characters, but it can be extended to words with modifiers or conjuncts. As in case of Assamese literature modifiers and conjuncts are used in words very often. In assamese literature there a number of elements which are quite similar in structure. So these characters can be misclassified as there features are quite similar. So a similarity analysis between the characters can be done so that they never get misclassified.

6. REFERENCES

- [1] Kaustubh Bhattacharyya and Kandarpa Kumar Sarma "ANN-based Innovative Segmentation Method for Handwritten text in Assamese", IJCSI International Journal6 of Computer Science Issues, Vol. 5, 2009, pp 9-16
- [2] Mohammad Adnan Al-Alaoui, Mohammad Amin Abou Harb, Zeid Abou Chahine, and Elias Yaacoub,"A New Approach for Arabic Offline Handwriting Recognition", IEEE multidisciplinary engineering education magazine, vol. 4, no. 3, september 2009, pp 89-97
- [3] Zaidi Razak, Khansa Zulkiflee , Mohd Yamani Idna Idris, Emran Mohd Tamil, Mohd Noorzaily ,Mohamed Noor, Rosli Salleh, Mohd Yaakob ,Zulkifli Mohd Yusof and Mashkuri Yaacob,"Off-line Handwriting Text Line Segmentation : A Review", IJCSNS International Journal of Computer Science and Network Security, vol.8 No.7, July 2008 pp 12-20
- [4] M. Arivazhagan, H. Srinivasan, S. N. Srihari.2007. A Statistical Approach to Handwritten Line Segmentation. In Proceedings of SPIE Document Recognition and Retrieval XIV , San Jose, CA, February 2007
- [5] Dinesh Dileep," A feature extraction technique based on character geometry for Character recognition"
- [6] Sarma Kandarpa Kumar, Member, IEEE,"Bi-lingual Handwritten Character and Numeral Recognition using Multi-Dimensional Recurrent Neural Networks (MDRNN)," International Journal of Electrical and Electronics Engineering 3:7 2009, pp 441-448
- [7] Yusuf Perwej, Ashish Chaturvedi," Machine Recognition of Hand Written Characters using Neural Networks", International Journal of Computer Applications (0975 – 8887) Volume 14– No.2, January 2011, pp 6-9.
- [8] Belhadeh Hacene, Eutamene Aicha, Kholadi Mohamed Khiredine "Character Recognition Approach Based on Ontology".in press. Pp 160-168
- [9] Krevat Elie, Cuzzillo Elliot, "Improving Off-line Handwritten Character Recognition with Hidden Markov Models,"in press.
- [10] Peyarajan S, "On-line Tamil hand written character recognition using Kohonen neural network," Research Journal of Computer Systems Engineering- An International journal,in press.
- [11] Ranpreet Kaur , Baljit Singh , " A Hybrid Neural Approach For Character Recognition System", International Journal of Computer Science and Information Technologies, Vol. 2 (2) , 2011, pp 721-726
- [12] Fox Richard, Hartmann William, "An Abductive Approach to Hand- written Character Recognition for Multiple Domains".
- [13] Sandhya Arora, Debotosh Bhattacharjee, Mita Nasipuri, Dipak Kumar Basu, Mahantapas Kundu," Combining Multiple Feature Extraction Techniques for Handwritten Devnagari Character Recognition ", 2008 IEEE Region 10 Colloquium and the Third ICIIS, Kharagpur, INDIA December 8-10. Pp 1-6
- [14] Sheetal Dabra, Sunil Agrawal and Rama Krishna Challa, "Novel Feature Set for Recognition Of Similar Shaped

Handwritten Hindi Characters Using Machine Learning”,
Cs & It 02, Pp. 25–35, 2011.





- [15] Bilan Zhu and Masaki Nakagawa, “Online Handwritten Chinese/Japanese Character Recognition”, InTech, 2012
- [16] Sarat Saharia, Prabin K. Bora, Dilip K. Saikia, “Improving Character Recognition Accuracy of Tchebichef Moments by Splitting of Images”, NCC 2009, January 16-18, IIT Guwahati, pp 390-393
- [17] Dipak D. Bage, K. P. Adhiya, Sanjay S. Gharde, ” A New Approach For Recognizing Offline Handwritten

Mathematical Symbols Using Character Geometry”,
International Journal Of Innovative Research In Science,
Engineering And Technology Vol. 2, Issue 7, July 2010

- [18] Mamatha H R and Srikantamurthy K. Morphological Operations and Projection Profiles based Segmentation of Handwritten Kannada Document ,International Journal of Applied Information Systems (IIAIS) – ISSN : 2249-0868 Foundation of Computer Science FCS, New York, USA Volume 4– No.5, October 2012

7. APPENDIX

Table 2: The percentage of recognition accuracy obtained for some of the character samples

Character	Number of samples trained	Number of samples used for testing	Number of epoch	% of Recognition Accuracy
	50	10	5000	88%
	50	10	6000	80%
	50	10	9000	90%
	50	10	10000	78%