A Research Survey on Sanskrit Offline Handwritten Character Recognition

R. Dineshkumar[1], Prof. Dr.J.Suganthi[2]

[1] HOD Incharge - CSE, KTVR Knowledge Park for Engineering and Technology, Coimbatore, Tamilnadu, India
me.dineshkumar@gmail.com
[2] Vice Principal and Head – CSE, Hindusthan College of Engineering and Technology, Coimbatore, Tamilnadu, India.
sugi_jeyan@hotmail.com

Abstract: Sanskrit (Devanagari), an alphabetic script, is used by over 500 million people all over the world. Recognition of Sanskrit (Devanagari) handwritten scripts is complicated compared to other language scripts. However, many researchers have provided real-time solutions for offline Sanskrit character recognition also. Offline Sanskrit handwritten documents recognition still offers many motivating challenges to researchers. Current research offers many solutions on Sanskrit (Devanagari) handwritten documents recognition even then reasonable accuracy and performance has not been achieved. This paper analyses the various approaches and challenges concerning offline Sanskrit (Devanagari) handwritten character recognition.

Index Terms – Binarization, Segmentation, SVM

I. Introduction

The handwritten text written in palm leaves decayed over a period of time. It is very difficult to preserve them in the same form. . This paper analyses the various approaches and challenges concerning offline Sanskrit handwritten character recognition.

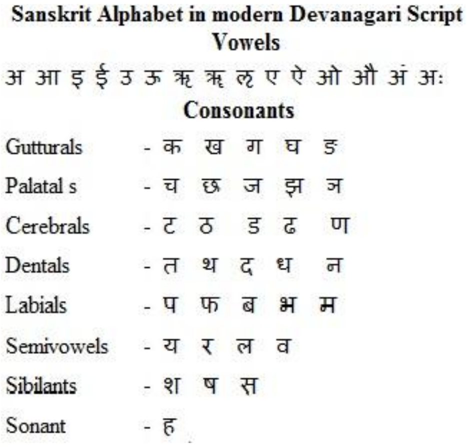
Recognition of characters can be done either from printed documents or from handwritten documents. Handwritten document recognition can be done offline or online. Offline character recognition is more complicated than online. In particular, Sanskrit (Devanagari) handwritten OCR is more complicated than other related works. This is because Sanskrit (Devanagari) letters have more angles and modifiers.

Challenges that researches face during recognition process are due to the curves in the characters, number of strokes and holes, sliding characters, differing writing styles so on.

The steps involved in character recognition comprise pre-processing, segmentation, feature extraction and classification.

1.1 Sanskrit (Devanagari) Language

Although, Sanskrit is an ancient language and no longer spoken, written material still exists. Hindi is world's third most commonly used language after English and Chinese and there are approximately 500 million people all over the world that speak and write in Hindi. Devanagari has about 14 vowels and 34 consonants. It is used as the writing system for over 28 languages including Sanskrit, Hindi, Kashmiri, Marathi and Nepali. The corpus of Sanskrit literature encompasses a rich tradition of poetry and drama as well as scientific, technical, philosophical and dharma texts. Sanskrit continues to be widely used as a ceremonial language in Hindu religious rituals and Buddhist practice in the forms of hymns and mantras.



Sanskrit Alphabet in modern Devanagari Script
Vowels

अ आ इ ई उ ऊ ऋ ॠ ऌ ए ऐ ओ औ अं अः

Consonants

Gutturals	- क ख ग घ ङ
Palatal s	- च छ ज झ ञ
Cerebrals	- ट ठ ड ढ ण
Dentals	- त थ द ध न
Labials	- प फ ब भ म
Semivowels	- य र ल व
Sibilants	- श ष स
Sonant	- ह

Figure 1: vowels and consonants

II. BLOCK DIAGRAM OF RECOGNITION SYSTEM

The schematic block diagram of handwritten Sanskrit (Devanagari) Character Recognition system consists of various stages as shown in figure. They are Preprocessing, Segmentation, Feature Extraction, and Classification.

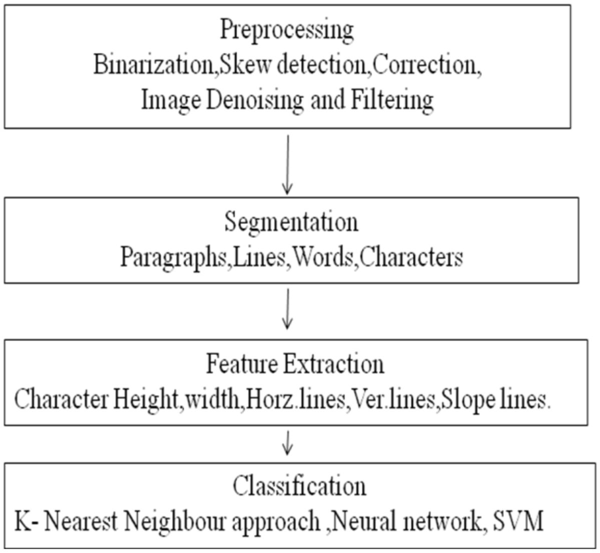

Preprocessing
Binarization,Skew detection,Correction,
Image Denoising and Filtering
Segmentation
Paragraphs,Lines,Words,Characters
Feature Extraction
Character Height,width,Horz.lines,Ver.lines,Slope lines.
Classification
K- Nearest Neighbour approach ,Neural network, SVM

Figure 2. Block diagram of Recognition system

2.1 Preprocessing

This is the first step in the processing of scanned image. The scanned image is pre processed for noise removal. The resultant image is checked for skewing. There are possibilities of image getting skewed with either left or right orientation. Here the image is first brightened and binarized.

The processes which get involved in pre-processing are illustrated below:

1. Binarization
2. Noise reduction
3. Normalization
4. Skew correction, thinning

2.1.1 Binarization

Binarization is a method of transforming a gray scale image into a black and white image through Thresholding. Normally, most researchers use thresholding concepts to extract the foreground image from background image.

2.1.2 Noise Removal

Digital images are prone to many types of noises. Noise in a document image is due to poorly photocopied pages. Median Filtering, Wiener Filtering method and morphological operations can be performed to remove noise. Median filters are used to replace the intensity of the character image, where as Gaussian filters can be used to smoothing the image.

2.1.3 Normalization

Normalization is the process of converting a random sized image into a standard size. The Roi-Extraction [19] method is used to get the single structural element from the image.

2.1.4 Skew correction, Thinning

Thinning is a pre-process which results in single pixel width image to recognize the handwritten character easily. It is applied repeatedly leaving only pixel-wide linear representations of the image characters.

2.2 Segmentation

After pre-processing, the noise free image is passed to the segmentation phase, where the image is decomposed into individual characters. The various steps in segmentation.

Algorithm for segmentation:

(1) The binarized image is checked for Inter line space.

(2) If inter line spaces are detected then the image is segmented into sets of paragraphs across the interline gap.

(3) The lines in the paragraphs are scanned for horizontal space intersection with respect to the background. Histogram of the image is used to detect the width of the horizontal lines. Then the lines are scanned vertically for vertical space intersection. Here histograms are used to detect the width of the words. Then the words are decomposed into characters using character width computation

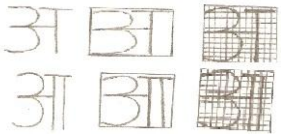


Figure 3. Character Segmentation

2.3 Feature extraction

The feature extraction techniques can be broadly grouped into three classes namely statistical features, structural features and the hybrid features. A statistical technique uses quantitative measurements for feature extraction, whereas structural techniques use qualitative measurements for feature extraction. In hybrid approach, these two techniques are combined and used for recognition.

In feature extraction where individual image glyph is considered and extracted for features. Each character glyph is defined by the following attributes: (1) Height of the character. (2) Width of the character. (3) Numbers of horizontal lines present—short and long. (4) Numbers of vertical lines present—short and long. (5) Numbers of circles present. (6) Numbers of horizontally oriented arcs. (7) Numbers of vertically oriented arcs. (8) Centroid of the image. (9) Position of the various features. (10) Pixels in the various regions

2.4 Classification

The Extracted features are given as the input to the Classification process. A bag-of-key point extracted from the feature extraction approaches are used for classification. There are some approaches are used for classify the character features in the existing systems such as K- Nearest Neighbour approach, Neural network, SVM classifier and so on.

2.4.1 SVM Classifiers

Support vector machines (SVM), when applied to text classification provide excellent precision, but poor recall. One means of customizing SVMs to improve recall, is to adjust the threshold associated with an SVM. Shanahan and Roma described an automatic process for adjusting the thresholds of generic SVM [20] with better results.

SVMs have achieved excellent recognition results in various pattern recognition applications. Some more properties are commonly seen as reasons for the success of SVMs in real-world problems. [20]. Empty area around the decision boundary defined by the distance to the nearest training patterns. These patterns, called support vectors, finally define the classification function.

2.4.2 Neural network Technique

Neural network architectures can be classified as, feed forward and feedback (recurrent) networks. The most common neural networks used in the OCR systems are the multilayer perceptron (MLP) of the feed forward networks and the Kohonen's Self Organizing Map (SOM) of the feedback networks.

III. COMPREHENSIVE STUDY

The table 1 shows the comprehensive study of which has been made on the different OCR’s available for offline character recognition:

Ref. Paper No	Preprocessing	Segmentation	Feature Extraction	Classification
[1]	Binarization (two-level thresholding), using a global or a locally adaptive method conversion to another character representation (e.g. skeleton or contour curve)	Line (imaginary) segment and word and characters segment	Template matching Deformable templates Graph description Discrete features Zoning Fourier descriptors	Graph descriptions or Grammar based descriptions of the characters are well suited for structural or synthetic classifiers. Real-valued feature vectors are ideal for statistical classifiers.
[2]	Image Binarisation Thinning of binarised image Windowing	Character recognition by neural network (Feed Forward Algorithm.	Replacing the recognized characters by standard fonts.(Back Propagation Algorithm)	Assembling all the separated characters in the same order as they appeared in the input image to give final output.
[3]	Convert the character image to bitmap and scale	Chain code histogram(for each segment)	Shadow features are extracted from scaled binarized character image. Chain code Histogram features are extracted by chain coding.	Combined MLP Combined MLP and Minimum Edit distance classifier
[4]	Binary level images, pseudo color and true color images. (preprocessing required is minimal) The next category of images is the pseudo color images.	Line Segmentation Word Segmentation Character Segmentation Maintain the data structure to feed the line, word and character boundaries	-	Normalized feature vector used to identify characters using fuzzy logic
[5]	Median and wiener		Zone based approach is	SVM for

	filtering used for noise removal	Conjunct segmentation algorithm process Structural segmentation algorithm.	used for Feature extraction	Classification Process
[6]	Preprocessing is done to normalize the position and size of the sample and to remove local noise so that the extracted features from the sample become robust.	Horizontal projection file method is used for segmentation(upper line and lower line)	Images scaled into height and width using bilinear interpolation technique Unwanted portion corrected using sobel edge.	well known feed forward algorithm
[7]	Morphological operation are used to noise removal Thinning algorithm is used to remove the distortions Bicubic interpolation are used for standard sized image	Differential distance based technique used for identifying the shirorekha and spine	Top, bottom, left, right or on a combination technique. A single or double vertical line called a Danda (Spine) was traditionally used to indicate the end of phrase or sentence	Preliminary classification is performed for better results. Devnagari hand-printed numeral recognition system based on binary decision tree classifier
[8]	Thresholding method used for Binarization	Lines are segmented by noting the valleys of projection profile	Upper Zones denotes portion above head line, middle Zone denotes portion of basic, lower Zone in the portion below base line. Special feature called hill-valley-distance extracted to estimate the orientation of text block	Statistical analysis, insensibility of machine printed and hand written text classification
[9]	The digitized images are in gray tone uses histogram based thresholding approach to convert them in two-tone images.	Line segmentation uses a horizontal projection profile based technique	Head-line feature horizontal projection profile Robustness, accuracy and speed of computation are the features used in feature extraction	Optical Character Recognition (OCR) used to separate different scripts before feeding them to their individual OCR system.
[10]	Morphological operations are used to achieve Noise removal.	Histogram profile and connected component analysis for line character segmentation	Any interesting character points Scale Invariant Feature Transform(SIFT) method for Feature Extraction K-Means clustering to	Bag-of-key points to count the number of patches K-Nearest Neighbor to recognition Scale.

	Bicubic interpolation is used for standard sized image. Morphological gradient to find character boundaries		create a code book for each characters (two features condition Nearest Neighbour,centroid condition) CB obtained from LBG algorithm	Invariant Feature Transform(SIFE) transfer character image to set of local frame
[11]	Scanned data encapsulated in Glyphs	-	Character height, width	Multilayer perception (MLP) learning algorithm for two hidden layers with back propagation used for final character identification. Hyperbolic tangent function used for activation purpose.
[12]	Median filter used to replace the intensity filter (noise). Thresholding method used for Binarization	Line segmentation used to segment the characters.	Cover vertical projection profile, word profile, background to-link transitions Profile features (vertical features and word profile) approach used for feature extraction.	Hidden Markov model (HMM) method used to recognize character. Recognize the word using Baum-Welch or forward and backward algorithm
[13]	Thinning algorithm used to thin the characters Hilditch's algorithm is used for skeletonization. Thresholding method used for Binarization	Spatial space detection technique. Histogram method used to convert the image to glyph	Character height, width, no. of horizontal and vertical lines (long and short). Horizontal and vertical curves, circles, slope lines image centroid and dots Output of the segmentation part(image glyph) is subjected to feature extraction procedure	Support Vector Machine(SVM) used for classification (vapnik's structural). Type 1 SVM. Self organizing map used to minimize the errors. Neural classification algorithm and Radial-Basis-Function networks, Hybrid Neuro fuzzy systems used for reduce the recognition problems. RCS Algorithm used for better recognition results.
[14]	Gabor Thiresholding and Otsu Thresholding methods(global) are used for Binarization	Horizontal and vertical profile method is used for segmentation (line and characters). Bilinear interpolation	Zone will capture pixel variation Zone based approach is used for Feature	Support vector machine (SVM) method is used for classification. Redial basic

		technique used for normalization	extraction	function(RDF) kernel used for support to SVM to divide the line
[15]	Threshold technique used for preprocessing color image to gray scale. Normalization done using Java Image Class		Row wise, column wise and diagonal wise selection Encoding binary variation method used for extract the features. Then comparing trained text and tested image for recognize the characters	SVM Comparison technique
[16]	Converted to grey image. Noise reduction is done through Nonlinear Algorithm. Grey image is then converted to Black and White image called Binarization	Line segmentation is done through ROWS having black pixel frequency. Word segmentation is done through COLUMN having black pixel frequency. In Letter segmentation MATRA is preserved	Tree Data Structure is used in features. Separate algorithm is used for Line tracing and making Feature tree. Connected components are extracted. For each char component we make Feature tree and finally prefix notation is applied to HMM.	Post processing is done through 2 steps 1. Constructing letter from component. 2.Rearranging the letter
[17]	Scanned document is Filtered and Binarized for both Hindi and Telugu. Projection files in range of + or – 20 degree for Skew correction.	Line and Word segmentation is done through projection files For both Hindi and Telugu. In Telugu characters are split to constituent components.	1. Extraction considers entire image. 2. Considers some selected moment and shape as its dimensionality is reduced by principal components.	SVM-ability to identify decision boundary with minimal margin K-nearest neighbor and neural network classifiers are popular for characters rearrange application.
[18]	Approach that computes the length of input stroke and if it is smaller than set of priority we ignore those words. Operations of pre-processing such as Shift of origin, smoothing, resampling of points.	Analytic approach segmentation is made into smaller components and is identified during recognition stage. No segmentation is done in holistic approach. Recursive contour following approach and certain water reservoir technique was used in segmentation	Preprocessed stroke is divided in to 7 sub strokes.Then its center of gravity (C.G), Histogram is calculated for number of feature components are generated and feature vector for strokes.	QDF classifier used for recognition of offline Hand writing. MQDF is used by considering the principle of Eigen vectors.

IV. CONCLUSION

A lot of research work exists in the survey for Sanskrit (Devanagari) Handwritten recognition. However, there is no standard solution to identify all Sanskrit characters with reasonable accuracy. In this paper, we have projected various aspects of each phase of the offline Sanskrit character recognition process. Researchers have used minimal character set. The following key challenges to be carried out by researchers cursive character, increased number of holes and strokes, mixed words.

REFERENCES

[1] Oivind Due Trier, Anil K Jain, Torfinn Taxt "Feature Extraction Methods For Character Recognition A Survey" ,(1995) Pattern Recognition Vol 29, No. 4 pp. 641-662, 1996.

[2] K. Y. Rajput and Sangeeta Mishra "Recognition and Editing of Devnagari Handwriting Using Neural Network", Proceedings of SPIT-IEEE Colloquium and International Conference, Mumbai, India Vol. 1, 66.

[3] S. Arora, D. Bhattacharjee, M. Nasipuri, D. K. Basu & M. Kundu "Recognition of Non-Compound Handwritten Devnagari Characters using a Combination of MLP and Minimum Edit Distance", International Journal of Computer Science and Security (IJCSS), Vol 29, No. 4 pp. 641-662, 1997.

[4] Script Vijay Kumar, Pankaj K. Sengar "Segmentation of Printed Text in Devanagari Script and Gurmukhi" International Journal of Computer Applications (0975 – 8887) Volume 3 – No.8, June 2010.

[5] Bansal, Veena and R.M.K. Sinha "Segmentation of touching and fused Devanagari characters", Pattern Recognition, volume 35 (2002), number 4 pp. 875-893.

[6] Chandan Biswas, Ujjwal Bhattacharya, Swapan Kumar Parui "HMM Based Online Handwritten Bangla Character Recognition using Dirichlet Distributions", International Conference on Frontiers in Handwriting Recognition 2012.

[7] Sandhya Arora, Debotosh Bhattacharjee, Mita Nasipuri, Latesh Malik "A Two Stage Classification Approach for Handwritten Devanagari Characters" Proceedings of the Fifth International Conference on Document Analysis and Recognition,1999, pp.653-656.

[8] Veena Bansal and R. M. K. Sinha "Integrating Knowledge Sources in Devanagari Text Recognition" IEEE Transactions on Systems, Man, and Cybernetics, Systems and Humans, VOL. 30, NO. 4, JULY 2000

[9] M C Padma and P A Vijaya "Identification of Telugu, Devanagari and English Scripts Using Discriminating Features", International Journal of Computer science & Information Technology (IJCSIT), Vol 1, No 2, November 2009

[10] S. Palit and B.B. Chaudhuri "A feature-based scheme for the machine recognition of printed Devanagari script", In Pattern Recognition, Image processing and Computer Vision, Ed. P. P. Das and B. N. Chatterjee, Narosa Publishing House, 1995, pp. 163-168.

[11] Stuti Asthana, Farha Haneef and Rakesh K Bhujade, "Handwritten Multiscript Numeral Recognition using Artificial Neural Networks", Int. J. of Soft Computing and Engineering ISSN: 2231-2307, Volume-1, Issue-1, March 2011

[12] Sigappi A.N, Palanivel S and Ramalingam V, "Handwritten Document Retrieval System for Tamil Language", Int. J of Computer Application, Vol-31, 2011

[13] Suresh Kumar C and Ravichandran T, "Handwritten Tamil Character Recognition using RCS algorithms", Int. J. of Computer Applications, (0975 – 8887) Volume 8– No.8, October 2010

[14] Shanthi N and Duraiswami K, "A Novel SVM -based Handwritten Tamil character recognition system", springer, Pattern Analysis & Applications,Vol-13, No. 2, 173-180,2010

[15] *U. Garain, B.B. Chaudhuri* "Segmentation of Touching Characters in Printed Devnagari and Bangla Scripts Using Fuzzy Multifactorial Analysis", Proceedings of the 6th International Conference on Document Analysis and Recognition. (ICDAR '01).

[16] Md. Sheemam Monjel and Mumit Khan "Optical Character Recognition For Bangla Documents Using HMM" Proceedings of Int. Conf. on Document Analysis and Recognition, Bangalore, India, September 20-22, 1999.

[17] C. V. Jawahar, M. N. S. S. K. Pavan Kumar, S. S. Ravi Kiran "A Bilingual OCR for Hindi-Telugu Documents and its Applications", Proc. of the 11th ICPR, vol. II, pp. 200-203, 1992.

[18] U. Bhattacharya A. Nigam Y. S. Rawat S. K. Parui "An Analytic Scheme for Online Handwritten Bangla Cursive Word Recognition", Pattern Recognition, vol. 26, no. 3, pp. 451-460, 1993

[19] Xiaoguang Shao, Kun Gao, Guoqiang Ni "A new method for the extraction of Region-Of-Interest based on visual Attention", Proc. of SPIE Vol. 7513 75132X

[20] Sandhya Arora et al., "Performance Comparison of SVM and ANN for Handwritten Devnagari Character Recognition", IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 3, May 2010.

AUTHORS BIOGRAPHY

Mr. R.DINESHKUMAR obtained his B.E degree in CSE from Bharathiar University. He received M.E degree in Computer Science in 2008, from the Anna University, Tamilnadu, India. He is a part-time PhD research scholar in the Department of Information Technology, Anna University, Chennai. His main interests in research are Hand written Character recognition of Sanskrit language. Email address: me.dineshkumar@gmail.com

Prof.Dr.J.SUGANTHI obtained her B.E degree in CSE from Madurai Kamaraj University. PG degree in M.E CSE from Bharathiar University, Coimbatore & further did her PhD in Anna University, Chennai. She is with Hindusthan college of Engineering and Technology as Vice Principal and Head, CSE since Aug. 2008. Her main interests in research are Image processing and Data mining. Email address: sugi_jeyan@hotmail.com