International Journal of Trend in Scientific Research and Development (IJTSRD)
International Open Access Journal
ISSN No: 2456 - 6470 | www.ijtsrd.com | Volume - 2 | Issue – 3

Pattern Recognition of Jharkhand Tribal Language

Ritushree Narayan
Department of Computing and Information Technology, Usha Martin University, Ranchi, India

Puja Mishra
Department of Management, Usha Martin University, Ranchi, India

ABSTRACT

Image processing has wide area for processing various functionality of image. Image with any pattern comes under the categories of pattern recognition where recognizing of the pattern can be any character, symbol, numeral or it can be any image also. Character Recognition (CR) has broad area of research in Devnagri script. Devnagri script has complicated structure. so, this script has not progressed well. Devanagari character recognition provides less correctness and efficiency. To recognize Devanagari script, various development done which is discuss in detail. Developers used to recognize the pattern with their structure, template, and graph. Some developers use classifiers to segmenting the characters.

Keywords: *Devanagari Script, Offline Character recognition, Feature Extraction, Segmentation, Neural Network*

INTRODUCTION

"Aendra manja, aekase lagei?" In Kurukh, the language spoken widely among the Oraons, it means "what's happened? how's your health?"
"*Raure man kaisen hi ?"* in Nagpuri(sadari) , the language spoken widely among local people,it means "How are you?"
The Chotanagpur region of Jharkhand has many tribal groups living close to each other. A unique phenomenon of this region is the emergence of a hybrid language called "Nagpuri" or "Sadri", which is used as lingua franca. It is a mix of many tribal languages and Hindi. It's a bit like the Creole used among migrants in some areas of the world. In Santal Parganas (Dumka) region, there are two main tribal groups – the Santal and the Paharia. The Santals see themselves as dominant and do not intermarry with Paharias, though they do intermarry with Oraon and Munda groups (which are the most advanced among Chotanagpur tribes). There are other small groups also in Chotanagpur like the Birhor of Netarhat and the Chik Baraik. The Birhor and Paharia communities are among the "Particularly Vulnerable Tribal Groups" (PVTGs) identified by the government. Handwriting Recognition (HWR) is one of the most engrossing and challenging research areas in the field of image processing and pattern recognition. Handwriting recognition System is a system by which a computer system can recognize the characters and the symbols written by hand in natural language like English, Devanagari, and Gurumukhi etc. The HWR system is basically of two types: Online HWR System and Offline HWR System. In Online HWR system, the text to be recognized is given as input to the system using a stylus or digitizer. Then the data signals undergo some filtration process, the data signal is then normalized to normal size and the slant and slope is corrected. After normalization, the text is divided into segments, and each segment is classified and labelled. Then using a search algorithm the most appropriate path is sent back to the user as output. In Offline HWR system, the text to be recognized is given as input in the form of scanned text, camera pictures etc. The Optical Character Recognition (OCR) is a type of Offline HWR system. In OCR, the input data is segmented into pieces using different algorithms. After the data is segmented into pieces, the text is further segmented into words or characters and sent to the recognition system. In the engine, the skeletonization and preprocessing is applied on the segmented text. Then different classifiers are applied

which extract certain features and creates a character hypothesis list. Then a search algorithm is used to search the most appropriate path in conjunction with language models and sent as output to the user.

A. Pattern Recognition

It is defined as the field concerned with machine recognition of meaningful regularities in noisy and complex environments. Pattern recognition techniques associated symbolic value with the image of the pattern. Pattern recognition applicable in character recognition, online signature verification, and face recognition and so on Four significant approaches to PR have evolved. Which are Statistical based, Syntactic based, neural network and Knowledge-based.

B. *Character Recognition*

Character recognition is part of pattern recognition field in which images of characters from a text image are recognized and recognition result as character codes are returned. It is the process of recognizing typed, printed or handwritten characters and converting into machine readable code. Process of converting printed or handwritten scanned documents into their corresponding ASCII characters that system can recognize converting image of documents into digital textual equivalent. Character recognition can be used for automatic number plate recognition, converting handwriting in real time to control a computer, as a reading aid for blind etc.

1) Offline System:

It is based on the type of the text which is printed or hand written. In this type of character recognition as handwritten, type written or printed text is well transformed into digital format. There does not exist benefits of recognizing direction of the movements while writing the character. The typewritten or handwritten character is normally scanned in form of paper document and store it in codes of a binary/gray scale image to the recognition algorithm.

2) Online System:

It is the two dimensional coordinates represented as function of time and order of strokes. The online methods superior to off-line in recognizing handwritten characters as temporal information available. Online character recognition is much easier and achieved better results than offline character recognition because more information may be captured in online like direction, speed and order of strokes of the handwriting.

II. DEVANAGARI SCRIPT

A. Devanagari script

It is the basic script of various languages which are speaking in India, such as Hindi and Sanskrit. This script is composition of symbols in two dimensions. In script, horizontal writing style, from left to right and characters do not have any uppercase/lowercase. Devanagari is a phonetic and syllabic script. Devanagari is phonetic, as words are written exactly as they are pronounced and syllabic means that text is written using consonants and vowels that together form syllables. Devnagari script has 13 vowels and 36 consonants and 11 modifiers. Devanagari has some features as:

B. Vowels

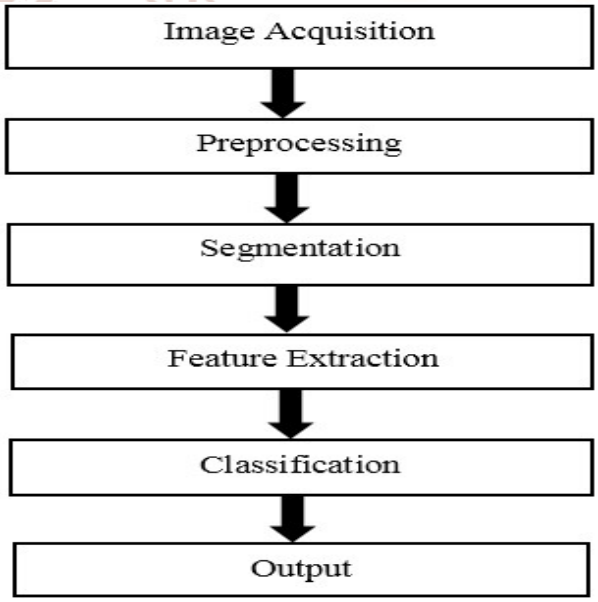

Figure 1: Steps of HWR system

The vowels use in two ways in both English and Hindi: to produce their own sounds and to modify sound of consonant; too this, an appropriate modifier The flowchart in Figure. 1 shows the steps that are performed during the recognition. The first step is the image acquisition in which the image consisting of the words to be recognized is input to the system. Then the image is preprocessed by performing the operations such as noise removal, normalization, skew detection and correction, greyscale and binarization etc. After pre-processing, the word in the image is segmented into individual characters for recognition. Then the key features are extracted from the image and these key features include height, width, density, loop, lines, stems and other character traits. This step is known as feature extraction. In the classification step, the methodologies of pattern recognition are

used for assigning an unknown sample to a predefined class. Then the character is compared with the character in the trained system and an output is obtained in form of character.

2. CHARACTERISTICS OF DEVANAGARI SCRIPT

In India, there are officially eighteen languages and Devanagari is one of them. The Devanagari script is used for writing Sanskrit and other Indian other languages. It is written from left to right, lacks distinct letter cases and is recognizable by the horizontal line running along the tops of the letters that links them together known as "Shirorekha" or headline. It consist of 11 vowels and 33 consonants. Vowels can be written as independent letters or by using them above, below, before or after the consonant they belong to. When the vowels are written in this way they are known as modifiers and the characters so formed are known as conjuncts. Two or more consonants can be combined together to form compound characters.

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

Table 1: Vowels and Corresponding Modifiers [2]

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

Table 2: Consonants [2]

क	ख	ग	घ		च		ज	झ	ञ	
			ण	त	थ		ध	न	प	फ
ब	भ	म	य		ल	व	श	ष	स	

Table 3: Half Form Consonants With Vertical Bar [2]

क क्क	क ल क्ल	घ न घ्न	च ञ च्छ	ज च ज्ञ	त न त्न	प त प्त	प ल प्ल
व व व्व	भ न भ्न	म ल म्ल	ल ल ल्ल	श न श्न	श व श्व	श ल श्ल	स न स्न

Table 4: Example of Combination Of Half-Consonants And Consonants [2]

क ष क्ष	ज ञ ज्ञ	ट ट ट्ट	ट ठ ट्ठ	त र त्र	द द द्द
द ध द्ध	द व द्व	द व र द्व	श र श्र	द भ द्भ	द य द्य

Table 5: Example Of Special Combination Of Half-Consonant And Consonant [2]

क़	ख़	ग़	ज़	फ़	ड़	ढ़	ॅ	ॉ	ः	।‌	ऽ	॑

Table 6: Special Symbols [2]

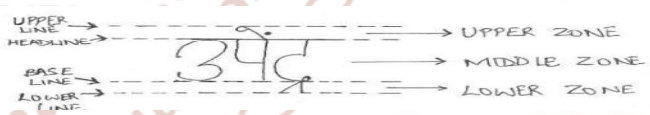

Figure 2: Different zones of Devanagari text

In Devanagari script, there are four imaginary lines that are drawn for a word, Headline which is also the header line, Baseline where the characters complete without modifiers, Upperline which the line above the Headline after above modifiers, and Lowerline which after the below modifiers. The text is partitioned into three zone: Upper zone between Headline and Upperline, Middle zone between Headline and Baseline, and the Lower zone between Baseline and Lowerline.

3. COMPLEXITY OF DEVANAGARI SCRIPT

While segmenting the Devanagari character, there are many problems that occurs due to complexity of the

script. Following are some of the complexities of the script:

i. Problem of broken character
ii. Problem of overlapped characters
iii. Problem of touching characters
iv. Problem of skewed characters
v. Problem of irregular intensity with character
vi. Problems due to presence of upper and lower modifiers
vii. Problems while separating the word and "Shirorekha"
viii. Problems of recognition special symbols

The HWR system should be developed such that it deals with all these complexities and recognizes the word properly.

4. IMAGE ACQUISITION

Image acquisition is the first step of the HWR system in which the image consisting of handwritten Devanagari characters are input to the system. Either a scanner or the digital camera can be used to take the image of the document consisting of the Devanagari script. The quality of the camera and the scanner effects the quality of the image.

5. PRE-PROCESSING

After the image is input to the system, the image is preprocessed by applying some steps which will help to improve the quality of the image for the recognition process.Some of the steps of the preprocessing are the following:'

5.1. Grayscale conversion

The image input to the system can be in rgb format which is a three dimensional image, but segmentation and recognition process can be applied only on a two dimensional image, so the colored image is converted into grayscale image.

5.2.Binarization

Binarization is a process of converting a grayscale image into binary form i.e. 0's and 1's representation, in which 0 represents black color and 1 represents white. Binarization separates the background and the foreground objects which is useful in segmentation. One of the most widely use method for Binarization is Ostu's algorithm. This method uses a threshold value to minimize the intra-class variance between the background and the foreground objects.

5.3. Skew correction

Different people have different writing style. It may be possible that the handwritten text in a document will not be in a straight line, this requires for the skew detection and correction. This is because if the skew will not be corrected then there will be problem in segmentation. First the skew is detected and then corrected. To detect a skew, the angle of the page is detected first and then the angle of lines are compared with it and the skew is detected. Then the skew correction algorithm is applied and the skew is removed.

6. SEGMENTATION

Segmentation is the process of breaking the connected word into individual character for the recognition process. The text in the document is first segmented into lines and then into words and finally into characters. To segment the Devanagari words into characters, the first step if to remove the Headline or the Shirorekha and then segment the word into individual characters. Bounding box technique, graph search and Ostu's threshold technique are some of the techniques used for segmentation.

7. FEATURE EXTRACTION

The process of extracting the useful information from raw data to minimize the intra-class pattern variability and maximize the inter-class pattern variability is known as feature extraction. Features are extracted from the segmented words so that to differentiate between classes. Some of the methods for feature extraction include: zone and count metric based system , feature extraction using MDRNN , statistical features etc.

8. CONCLUSION

In the fast growing era of technology, there has been a drastic increase in the research field of Devanagari Character recognition system. The recognition of the Devanagari character is a difficult task due to the "Shirorekha", upper and lower modifiers, and also due to the complexity of the characters. The errors relates to improper, skewed, broken letter, zig-zag letter images should be considered and removed to obtain a better accuracy during recognition. The features extracted should improve the process of recognition. Most of the research have used the concept of neural networks for classification, but there are many other techniques which can be used for classification. Shirorekha extraction and after extraction, recognition of the word properly. But since

only few research have been reported in this area, so different techniques can be applied for Shirorekha extraction and recognition of word properly. Devanagari characters also consist of lower and upper modifiers, but since this increases the complexity of the characters so there are very less innovation related to it. Different techniques should be applied on the Devanagari characters consisting of modifiers and the system should recognize the word correctly and properly with modifiers also. From the survey, it have been noted that the major problems in the recognition of Devanagari character include Shirorekha extraction, broken characters, skewed characters, errors generated by the scanner while scanning the images and many more. Proper techniques should be applied to handle all such errors and create an efficient and effective hand written recognition system.

Reference:

1) Zulfiqar Ali et al. Syed Khurram Shahzad, and Waseem Shahzad Performance Analysis of Statistical Pattern Recognition Methods in KEEL Procedia Computer Science 112 (2017) 2022–2030

2) T.M. Rajisha et al. Performance Analysis of Malayalam Language Speech Emotion Recognition System using ANN/SVM / Procedia Technology 24 (2016) 1097 – 1104

3) Li, Y., et al. Multi-view face detection using support vector machines and eigenspace modelling. In Knowledge-Based Intelligent Engineering Systems and Allied Technologies, 2000. Proceedings. Fourth International Conference on. 2000. IEEE.

4) Gutta, S., et al., Mixture of experts for classification of gender, ethnic origin, and pose of human faces. IEEE Transactions on neural networks, 2000. 11(4): p. 948-960.

5) Rameshwar S Mohite et al, Challenging Issues in Devanagari Script RecognitionInt.J.Computer Technology & Applications,Vol 5 (3),947-952

6) U. Pal, N. Sharma, T. Wakabayashi, F. Kimura, "Offline Handwritten Character Recognition of Devanagari Script", 9th International Conference on Document Analysis and Recognition, 1:496-500, 2007.

7) Niranjan Joshi, G. Sita, A.G. Ramakrishnan, Deepu V., Sriganesh Madhvanath, "Machine Recognition of Online Handwritten Devanagari Characters ", HP Laboratories India HPL-2005-224 November 19, 2007.

8) H. Swethalakshmi, Anitha Jayaraman, V. Srinivasa Chakravarthy, C. Chandra Sekhar, "Online Handwritten Character Recognition of Devanagari and Telugu Characters using Support Vector Machines", IEEE Region Colloquium and the Third , Kharagpur, 2008.

9) Vikas J Dongre, Vijay H Mankar, "Review of Reasearch on Devanagri Character Recognition", International Journal of Computer Applications (0975 – 8887) Volume 12– No.2, November 2010.

10) Amit Durandhar, Kartik Shankarnarayanan, Rakesh awale, "Robust Pattern Recognition Scheme for Devanagri Script" Lecture Notes in Computer Science,2005,volume 3801/2005,1021-1026,DOI:10,1007 /11596448-152.