2016 International Conference on Computational Science

Handwritten Marathi Compound Character Segmentation using Minutiae Detection Algorithm

Mrs. Snehal S. Golait\textsuperscript{a}, Dr. Latesh Malik\textsuperscript{b}

\textsuperscript{a}Research Scholar, Department of CSE,
G.H. Raisoni College of Engineering, Nagpur

\textsuperscript{b}Professor & Head, Department of Computer Science & Engineering,
G.H. Raisoni College of Engineering, Nagpur

Abstract

Segmentation process is the heart of handwritten script identification system. Aside from the large variation of individual’s handwriting, many researchers found difficulty to separate characters from scanned word document image. The key factor of selection of segmentation algorithm is used to improve efficiency of character segmentation as well as good feature extraction. One of the feature of Marathi script is Compound Character, derived from Devnagari, occur rarely in the script. Segmentation of such type characters is very difficult due to their complex structure. This paper proposed new technique for segmentation of handwritten Marathi compound characters. The proposed algorithm used the concept of Minutiae extraction for fingerprint for segmenting the compound character. Basically segmentation is carried out using morphological operations such as erosion and dilation. For segmenting the character from compound character our aim to find termination point and bifurcation points. And for finding the termination and bifurcation point proposed algorithm used the morphological operation hit or miss transform. The experimental results shows 90% accuracy in finding termination and bifurcation points.

© 2016 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
Peer-review under responsibility of the Organizing Committee of ICRTCSE 2016

Keywords: Erosion; Segmentation; bifurcation; Morphology;

E-mail address: snehal.golait@gmail.com
1. Introduction

An OCR has variety of physical and commercial applications. It can be used for automatic reading and processing of the forms, bank cheques, and old degraded documents. It can prove as an aid for visually handicapped persons. There are so many scripts and languages in India but very less work is done for recognition of handwritten Marathi Character.

Marathi is the language spoken by the native people of Maharashtra. Marathi script obtained from Devanagari, is an official language of Maharashtra. It is the 4th most spoken language in India and 15th most spoken language in the world. Marathi script consists of 16 vowels and 36 consonants making 52 alphabets. Marathi is written from left to right. Every character has a horizontal line at the top called as Shiorekha. The shiorekha joins the characters in a word. Vowels are combined with consonants with the help of specific characteristic marks. These marks occur in line or at the top or at the bottom of a character in a word. Marathi also has a complex system of compound characters in which two or more consonants are combined forming a new special symbol. Compound characters in Marathi languages obtained from Devanagari.

In optical character recognition, partitioned an image into its constituent regions or objects. That is, it partitions an image into different regions that are meant to correlate strongly with objects or features of interest in the image. The segmentation process is not the easy task, main goal of segmentation is to simplify changes in the representation of an image into meaningful parts which are easier to recognize. Image segmentation is basically used to locate objects and boundaries in scanned Image. Image segmentation is the process to allocate a label to every pixel in an image such that pixels with the same label share certain properties. A proper segmentation of characters is required before individual characters are recognized. This paper gives novel approach for segmenting Compound character for Handwritten Marathi Script. Handwritten character recognition for Indian languages is quite a challenging task for the researchers. This is due to the various characteristics of these scripts like their complex shape, large character, presence of modifiers, presence of compound characters and similarity between characters. The occurrence of compound characters in Marathi is found to be about 15 to 20%. The features of compound characters are consonants are not joined in an arbitrary manner but the combination of some specific characters[1]. The characters set of Marathi Character and sample of Handwritten Marathi document are shown in figure 1.

![Character set and Sample of Handwritten Marathi Document](image)

In this paper, we proposed the novel algorithm used the concept of Minutiae extraction for fingerprint for segmenting the compound character. The concept of minutiae is recognizing fingerprint based on local ridge features.

2. Literature Review

Devanagari is the most widely used script in India. Sanskrit, Nepali, Hindi and Marathi are the devanagari script used by more than 400 million people. Unconstrained Devanagari writing is more complex than English language due to the possible variations in the shape, number and direction of the constituent strokes.

E-mail address: snehal.golait@gmail.com
Devanagari script has 50 characters which can be written as individual symbols in a word. Devanagari Character recognition is complicated process due to presence of multiple conjuncts, loops, lower and upper modifiers and the number of disconnected and multistoried characters, in a word where all characters are connected through Shireorekh. OCR is further complicated by compound characters that make character separation and identification is very difficult.

OCR work on printed Devanagari Script started in early 1970’s. Sinhala and Maharaja published presented a syntactic pattern analysis system with an embedded picture language for the recognition of handwritten and machine printed Devanagari characters. Vienna described Devanagari OCR in her doctoral Thesis. Performance of 93% accuracy at character level is reported after post processing. Pal and Chaudhuri reported a complete OCR system for printed Devanagari with 96% accuracy.

First research report on handwritten Devanagari characters was published in 1977. At present researchers have started to work on handwritten Devanagari characters and few research reports are published recently. Hanmandlu and Murthy proposed a Fuzzy model based recognition of handwritten Hindi numerals and characters and they obtained 92.67% accuracy for Handwritten Devanagari numerals and 90.65% accuracy for Handwritten Devanagari characters. Bajaj et al employed three different kinds of features namely, density features, moment features and descriptive component features for classification of Devanagari Numerals. They proposed multi-classifier connectionist architecture for increasing the recognition reliability and they obtained 89.6% accuracy for handwritten Devanagari numerals. Kumar and Singh proposed a Zernike moment feature based approach for Devanagari handwritten character recognition. They used an artificial neural network for classification. Sethi and Chatterjee proposed a decision tree based approach for recognition of constrained hand printed Devanagari characters using primitive features. Bhattacharya et al proposed a Multi- Layer Perceptron (MLP) neural network based classification approach for the recognition of Devanagari handwritten numerals and obtained 91.28% results. N. Sharma and U. Pal proposed a directional chain code features based quadratic classifier and obtained 80.36% accuracy for handwritten Devanagari characters and 98.86% accuracy for handwritten Devanagari numerals. Dr. Latesh Malik describes recognition of handwritten devanagari Characters with percentage component regular expression matching and classification tree and obtained 95% accuracy. U. pal and Sagarikadutta worked on piece-wise projection method for line segmentation of handwritten Bangla script obtained 99% accuracy and histogram method for word segmentation of handwritten Bangla script obtained 97.8% accuracy. Shubair Abdulla worked on Rotational Invariant Segment method for word segmentation of Arabic handwritten Script obtained 95.66% accuracy. Sushama Shelke worked on handwritten Marathi Compound Character Recognition using Structural feature extraction technique wavelet transform obtained 94.22 % accuracy.[1] As per related research no one is using the minutiae technique to segmenting character. This paper discussed the concept of minutiae used for segmenting Marathi character from handwritten Marathi compound character.

3. Proposed Method

The proposed algorithm used the concept of minutiae extraction for fingerprint for segmenting the compound character. The concept of minutiae is recognizing fingerprint based on local ridge features. Fingerprint identification is commonly employed in forensic science to aid criminal investigations etc. A fingerprint is a unique pattern of ridges and valleys on the surface of finger of an individual. A ridge is defined as a single curved segment, and a valley is the region between two adjacent ridges. Most automatic fingerprint recognition systems are based on local ridge features known as minutiae. There are about 150 different types of minutiae [5] categorized according to their configuration. Among these minutia types, “ridge ending” and “ridge bifurcation” are the most commonly used, since the other types of minutiae can be seen as combinations

E-mail address: snehal.golait@gmail.com
of “ridge endings” and “ridge bifurcations”. Figure 2 illustrates the two most basic types of minutiae: ridge endings and bifurcations. These minutiae points are used for determining uniqueness of a fingerprint.

![Ridge Ending and Ridge Bifurcation](image)

**Figure 2: Ridge Ending and Ridge Bifurcation**

We analyzed the minutiae extraction for fingerprint recognition and used this concept for segmenting character from Marathi compound character. The flowchart for proposed algorithm is shown in the figure 3.

![Flowchart of Proposed Algorithm](image)

**Figure 3: Block Diagram of Proposed Algorithm**

**Input Image**

Input Image is scanned document using any flatbed scanner with resolution of minimum 300 dpi. The document may be scanned as color image or gray image or binary image.

**Binarization**

This is the first step in the processing of scanned image. In this process first digitization of Image is done. For binarization thresholding technique is used. Thresholding is an image processing technique for converting a gray scale or color image to a binary image based upon a threshold value.

**Normalization**

Normalization is one of the important pre-processing factors for character recognition. Normally, in normalization the character image is linearly mapped on to a standard plane by interpolation/extrapolation. The size and position of character is controlled such that the length and width of normalized plane are filled. By linear mapping, the character shape is not only deformed but also the aspect ratio changes.

**Ridge Segmentation**

The need of ridge segmentation is for finding the break points in character. In mathematics the ridges of a smooth function of two variables are a set of curves whose points are local maxima of the function in at least one dimension. This notion captures the intuition of geographical ridges. For a function of N variables, its ridges are a set of curves whose points are local maxima in N-1 dimensions. Correspondingly, the notion of valleys for a function can be defined by replacing the condition of a local maximum with the condition of a local minimum. The union of ridge sets and valley sets, together with a related set of points called the connector set form a connected set of curves that partition, intersect, or meet at the critical points of the function.

**Histogram Equalization**

Apply the Histogram equalization for adjusting image intensities to enhance contrast. In Matlab histeq() function is used to enhanced the image.

**Wavelet Decomposition**

The wavelet decomposition is merely done for the lossless reduction of the size of an image[11]. DWT decomposes the signal into mutually orthogonal set of wavelets. DWT decomposition is used for finding the

---

E-mail address: snehal.golait@gmail.com
pixels at horizontal, vertical and in diagonal direction. In CWT, the wavelets are not orthogonal and the data obtained by this transform are highly correlated. The fourvalues of the decomposition are shown in figure 4.

CA – Accurate Value , CD – Dimension value, CV – Vertical Value, CH – Horizontal Value

The four divisions in the image was the same in all level of the decomposition, instead the levels are theteration. These four divisions of the decomposition levels have their own characteristics and preserved the values of the image. The values are

LL – Smoothing image of the original Image,   LH - Preserves horizontal edge details
HL – Preserves vertical edge details, HH – Preserves diagonal edge

The original image is decomposed into many levels using the dwt function in the Matlab. The db1 waveletis used as wavelet transforms and the four level decomposition has been carried out for the characters.

Figure 4:4 Level Wavelet Decomposition

Gabor Filter

In the literature, there are several image noise removal methods which are applied in spatial, transform, or time-frequency domains. In the spatial domain a small mask is convolved with the image. This mask can be an averaging filter, a mean or Gaussian filter. In the transform domain, first the image is translated to the transform domain, then it is multiplied by a low pass filter and at the end by the inverse transformation, the enhanced image is obtained. In the transform domain, noise in the grey levels of an image contributes heavily to the high frequency components and the most of the image energy is concentrated in the low frequency components. Although, applying a low pass filter to a noisy image in the transform domain reduces the noise, at the same time it could eliminate some high frequency components that are not related to noise and weaken sharp transitions like edges. Furthermore, the transforms which perform on the whole image, do not take into account any spatial information where the frequency components come from. Therefore, noise reduction by low pass filtering in such domains does not preserve the local information of the image very well. Time-frequency transforms combine time-domain and frequency-domain analysis and allow obtaining a revealing picture of the temporal localization of the signal’s spectral components. Due to this problem we consider the Gabor filter as a noise reduction technique.

Thinning

The next step is to thin the processed binary image using the morphological thinning operation. The thinning algorithm removes pixels from ridges until the ridges are one pixel wide.[5]. The thinning of an image I by a structuring element J = (J₁, J₂) is given by

Thin(I,J)= I - (I J)

Finding termination and bifurcation point using Hit or Miss Transform.

After thinning we apply hit or Miss transform to find the termination and bifurcation points. Termination are those pixels in an image which have only one neighbor in a 3X3 neighborhoods. The termination is given by applying Hit or Miss transform on I by J as follows:

M₁ = (I J)

Where, M₁ is the thinned image and J is the sequence of structuring element pairs(J₁, J₂)

I J = (I ⊕ J₁) ∩ (I ⊕ J₂)

Ridge bifurcations are those pixels in an image which have only three neighbors in a 3X3 neighborhoods and these neighbors are not adjacent to each other. The minutiae image M₂ containing ridge terminations is given by

E-mail address: snehal.golait@gmail.com
\[ M_1 = (I \ominus J) \]
Where, \( I \) is the thinned image and \( J \) is the sequence of structuring element pairs \((J_1, J_2)\)

\[ I \ominus J = (I \ominus J_1) \cap (I \ominus J_2) \]

**Experimental Results**

(a)  
(b)  
(c)  
(d)  

a) Input Image  
b) Line Segmentation  
c) Word Segmentation  
d) Output of Minutiae Detection algorithm.

5. Conclusion

In this paper we are proposed algorithm for segmentation of handwritten Marathi Compound Character using Minutiae concept. Using minutiae extraction algorithm and Hit or Miss transform able to get near about 90% accuracy for finding termination and bifurcation points. Experimental result shows best results in getting termination points and bifurcation. Using Minutiae extraction algorithm efficiency of Marathi character recognition is increased up to 95%. This is the novel technique which is not used by any researcher.

**Acknowledgements**

This paper is made possible through the help and valuable support of my guide Dr. Latesh Malik. Finally, I sincerely thanks to my parents and God.

**References**


E-mail address: snehal.golait@gmail.com


E-mail address: snehal.golait@gmail.com