

# Zonal moments based Handwritten Marathi Barakhadi recognition

Shreya N. Patankar Leena R. Ragma

**Abstract** - Handwritten character recognition (HCR) is an important subset within the pattern recognition area. Very little work is happening on Marathi Barakhadi characters which are formed by the combination of one of the 12 vowels and 36 consonants resulting in 432 characters. As the number of characters to be uniquely identified is very large, the proposed method aims at recognizing Marathi language Barakhadi characters by recognizing a vowel and a consonant separately. Based on the Devanagiri characters shape analysis and data set, the whole image is split into top region image with information above the header line and middle region image with information below the header line. The middle region is further processed to detect and separate the side modifiers if any, for vowel recognition. Invariant moment features are extracted from the top region and from the side modifiers and classified using quadratic classifier for recognition of vowel matra. If no vowel matra found, the image is cut by 20-30% from the bottom for detecting the presence of lower modifiers. Invariant moment features are extracted from the cut image and classified using quadratic classifier. Core consonant is divided into various zones and invariant moment features are extracted from each zone. These features are compressed using principle component analysis and classified using quadratic classifier for consonant recognition. These features will be trained and tested for both vowel and consonant recognition using quadratic classifier.

**Keywords**- *Handwritten character recognition; Marathi Barakhadi; zonal moments; classifier; feature extraction.*

## I. INTRODUCTION

Character recognition is becoming more and more important in the modern world. It helps humans ease their jobs and solve more complex problems. Handwritten character recognition is a topic of research in recent years. It aims at automation by reducing the human efforts to a larger extent and to meet various applications like postal automation, office automation etc. Lot of work is being done in this particular area on different Indian languages but the work is limited to basic character set which comprises of vowels and consonants. Researchers have also achieved good recognition accuracy for the basic data set.

Because of the complexity associated with the large data due to the variations in the writing style of different individuals and shape similarity, handwritten character recognition systems are more

complex. Very little work is reported on Marathi language Barakhadi characters to the best of our knowledge. Marathi Barakhadi characters consist of top, side and bottom modifiers with their nature being curved with straight line existing between or to the sides of the consonants. We will be using Marathi Barakhadi characters for the experiment.

Previous research on HCR for Devanagiri language uses various feature extraction methods such as moments for vowel recognition [4], capturing directional information using gradient method [6], chain code histogram and shadow features [3] and [7], connected component labelling [10] etc. Some of these features are also applied on different languages like Bangla [9], kannada[1], Gurumukhi [5] etc. Gradient information is sensitive to noise where as moments are robust to high frequency noises as discussed in [1].

In this paper, we are proposing a method to recognise the vowel and consonant part separately for Marathi Barakhadi character using zonal moments and quadratic classifier.

The paper is organized as follows. Section 2 discusses the Marathi language Barakhadi characters. Section 3 gives the proposed methodology. Section 4 is devoted to feature extraction. Section 5 discusses the classifier used. Section 6 concludes our study.

## II. MARATHI BARAKHADI

Marathi is the language spoken by the native people of Maharashtra. Marathi is an Indo-Aryan language spoken by about 71 million people mainly in the Indian state of Maharashtra and neighbouring states. Marathi is also spoken in Israel and Mauritius. Marathi is thought to be a descendent of Maharashtri, one of the Prakrit languages which was developed from Sanskrit. Marathi first appeared in writing during the 11<sup>th</sup> century in the form of inscriptions on stones and copper. Marathi is written in Devanagiri script which is the most popular script in India.

The Marathi basic character set consist of 12 vowels and 36 consonants. The first 10 vowels are very widely used and the last two are less commonly used. Barakhadi character is a conjunct character formed by combining one of the 12 vowels with each of the 36 basic consonants. Thus

a Marathi Barakhadi has  $36 \times 12 = 432$  characters which comprises of large data set. Figure below shows the basic vowels and consonants and one sample of consonant Barakhadi.

अ आई ई उ ऊ ए ऐ ओ औ □□ □□  
 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 ळ □□□ □□□  
 प □□ □□ □□ □□ □□ □□ □□ □□ □□ □□  
 □□ □□

Figure 1. 12 Vowels, 36 Consonants and Barakhadi

III. PROPOSED METHOD

The proposed method to recognize a handwritten Barakhadi character uses zonal moments. This method tends to recognise a Marathi Barakhadi character by recognising the vowel and consonant parts separately. The steps of handwritten Marathi Barakhadi character recognition is shown in figure 5.

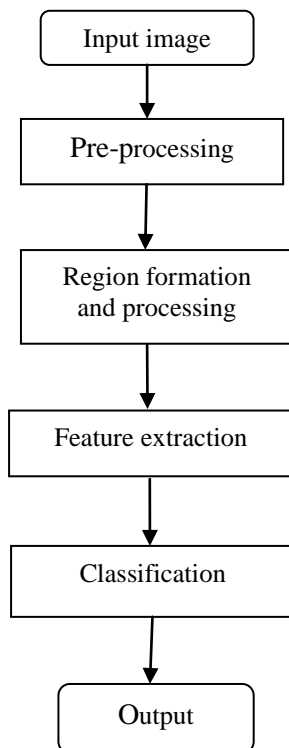


Figure 2. Marathi Barakhadi recognition

Pre-processing begins with thresholding where any character image with given file format is converted into binary image of 0's and 1's. Handwritten characters show various undesirable effects like unwanted strokes, gaps or breaks which occur due to binarization [5]. Many a times when a character is handwritten, it exhibits lesser width at the curvature than at other parts of the character.

This point is more likely to break during binarization. Hence, a 3x3 averaging filter will be applied before binarization, which blurs the image resulting into bridging small gaps and retaining the actual shape of the character. A minimum bounding box is fitted to the character and the character is cropped. To bring uniformity among the characters the cropped character image is normalized to fit into a specific size. After size normalization image is thinned to single pixel width.

The header line is the most distinguishing factor for any Marathi or Hindi language characters which needs to be detected and removed so that the image gets divided into two regions. Hough transformation is used for detection of header line [8]. Shown below is the diagram depicting two regions namely top region above the header line and middle region below the header line.

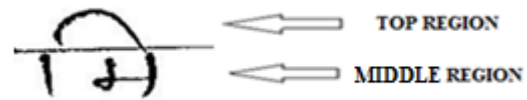


Figure 3. Region formation

Middle region is further processed so that any information present to the sides of the consonant can be detected by taking the vertical histogram of the image. If the side modifier information is present, its position is checked, saved and separated.

For the detection of vowel matra, features are extracted from the top region and side modifier if present. Consonant region is divided into various zones and features are extracted from each zone.

IV. FEATURE EXTRACTION

To recognize the Barakhadi, both vowel and consonant are to be recognized. The problem becomes complicated since separating of vowel and consonant information from a given handwritten Barakhadi character is very difficult due to high writing variations and need very robust set of features. In this paper, we focus on using moments.

Carefully selected moment features can ensure that the extracted features are invariant under translation, rotation and scaling. Also moments are robust to high frequency noise as high order terms are not used for feature formation [1]. More importantly moments can represent each character uniquely regardless of how close the characters are in terms of local features as discussed in [1]. This unique nature makes moments appropriate for handwriting character recognition.

a) Geometric moments

For a digital image with  $f(x,y)$  of size  $M \times N$ , image moments  $M_{ij}$  are calculated by

$$M_{ij} = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} x^i y^j f(x,y)$$

All  $M_{ij}$  with  $i+j \leq n$ , a positive integer, are the geometric moments of order  $i+j$ .

b) *Central moments*

To make features invariant to translation, the  $M \times N$  image plane is to be mapped onto a square defined by  $x \in [-1, +1]$  and  $y \in [-1, +1]$ . Invariance with respect to position of the object in the image can be achieved by calculating the central moments of the mapped digital image.

$$\mu_{ij} = \sum_{x=-1}^{+1} \sum_{y=-1}^{+1} (x - \bar{x})^i (y - \bar{y})^j f(x,y)$$

Where,  $\bar{x} = \frac{M_{10}}{M_{00}}$  and  $\bar{y} = \frac{M_{01}}{M_{00}}$  are the components of the centroid.

c) *Scale invariant moments*

Moments  $\eta_{ij}$  where  $i + j \geq 2$  can be constructed to be invariant to both translation and changes in scale by dividing the corresponding central moment by the properly scaled  $(00)^{\text{th}}$  moment using the following formula.

$$n_{ij} = \frac{\mu_{ij}}{\mu_{00}^{(1+\frac{i+j}{2})}}$$

d) *Rotation invariant moments*

It is possible to calculate moments which are invariant under translation changes in scale and also rotation. Most frequently used are the Hu's set of invariant moments.

$$I_1 = \eta_{20} + \eta_{02}$$

$$I_2 = (\eta_{20} - \eta_{02})^2 + (2\eta_{11})^2$$

$$I_3 = (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2$$

$$I_4 = (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2$$

$$I_5 = (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3\eta_{21} + \eta_{03}] + (\eta_{21} + \eta_{03})^2$$

$$I_6 = (\eta_{20} - \eta_{02})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03})$$

$$I_7 = (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3\eta_{21} + \eta_{03}] + (\eta_{21} + \eta_{03})^2$$

## V. CLASSIFICATION

Features are compressed using principle component analysis and then given as input to the classifier, one for vowel recognition and the other for consonant recognition. The job of classifier is to correctly classify the input into one of the several classes. In this paper, the proposed method uses Quadratic classifier which is based on quadratic discriminant analysis as shown below.

$$d_k(X) = (X - \mu_k)^T \Sigma_k^{-1} (X - \mu_k) + \log |\Sigma_k| - 2 \log \pi_k$$

Where,  $\mu_k$  and  $\Sigma_k$  are the class  $k$  mean vector and covariance matrix.  $X$  represents feature vector. And to the classification rule

$$d_k(X) = \min_{1 \leq k \leq K} d_k(X) \Leftrightarrow \max_{1 \leq k \leq K} p(k|X)$$

The classifier used for recognition will take input as the feature vector formed by extracting moment features. The extracted features will undergo two phases namely training and testing phase as shown in figure 4. Few of the extracted features of various samples of each character will be trained to recognize a particular character and a knowledge base will be prepared and kept in the database. Remaining samples will be used for testing the character by comparing the character with the knowledge base for recognition.

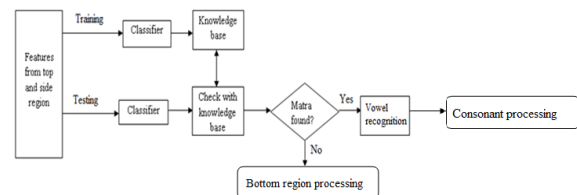


Figure 4. Training and testing phases

Moments features are extracted from the top and side regions to detect the presence of any vowel matra information. If any matra is not detected at the top or side or in both regions, then bottom region is processed to detect the presence of lower modifier. Whole image below the header line is cut from the bottom by 20-30%.

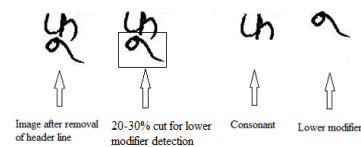


Figure 5. Bottom region processing

Moments features are extracted from the cut image and sent to the classifier for detecting the presence of lower modifiers. After detecting and separating the modifier information if any, the

consonant present in the middle region is divided into various zones. Features will be extracted from each zone and will undergo training and testing phases for recognition of consonant.

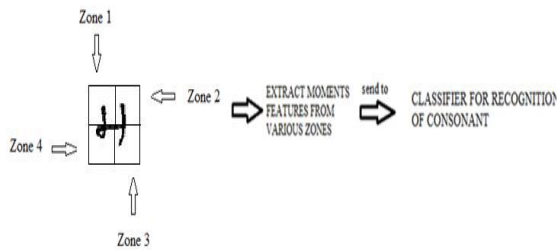


Figure 6. Consonant into zones

The extracted features for consonant recognition are compressed using principle component analysis and send to the classifier for recognition. The classifier recognizes the vowel and consonant part of the character image separately and the expected output is as shown in figure 9.

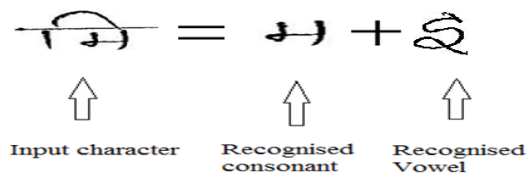


Figure 7. Expected Output

## VI. CONCLUSION

A method is proposed which focuses on recognition of handwritten Barakhadi recognition for Marathi language characters using zonal moments. Pre-processing followed by removal of header line helps to divide the image into two regions for further processing. Moments features are extracted from both the regions. Extracted features will be sent to the quadratic classifier for recognition of vowel and consonant part separately.

The Barakhadi recognition can be done by individual vowel and consonant recognition rather than as a Barakhadi character. This reduces the number of characters to be recognized from 432 to just 36 consonants and 12 vowels. That is a total of  $36+12=48$  unique shapes need to be identified.

The proposed methodology will be helpful to the researchers for the future work in handwritten recognition of isolated characters of any Indian language script.

## REFERENCES

- [1] Ragha L., and Sasikumar M., 2011, "Feature analysis for handwritten kannada kagunita recognition", *International Journal of Computer theory and engineering*, Vol. 3, No. 1.
- [2] Dhandra B., Hangarge M., and Mukarambi G., 2010, "Spatial features for handwritten kannada and English character recognition", *IJCA special issue on Recent trends in image processing and pattern recognition*, pp. 146-151.
- [3] Arora S., Bhattacharjee D., Nasipuri M., Basu D., and Kundu M., 2010, "Recognition of non-compound handwritten Devanagiri characters using a combination of MLP and minimum edit distance", *International journal of computer science and security*, Vol 04, No. 01, pp. 107-120.
- [4] Ramtake R., 2010, "Invariant moments based feature extraction for handwritten Devanagiri vowels recognition", *International Journal of computer applications*, Vol. 01, No.18, pp.1-5.
- [5] Lehal G., and Singh C., 2009, "Feature extraction and classification for OCR of Gurumukhi script", *International conference on Pattern recognition*, pp. 1-10.
- [6] Pal U., Wakabayashi T., and Kimura F., 2009, "Comparative study of Devanagiri handwritten character recognition using different feature and classifiers", *IEEE International conference on document analysis and recognition*, pp. 1111-1115.
- [7] Arora S., Bhattacharjee D., Nasipuri M., Basu D., and Kundu M., 2008, "Combining multiple feature extraction techniques for handwritten Devanagiri character recognition", *IEEE, Third International conference on Industrial and information systems*, pp. 1-6.
- [8] Singh C., Bhatia N., and Kaur A. , 2008, "Hough transform based fast skew detection and accurate skew correction methods", *Science direct, Pattern recognition*, pp. 3528-3546.
- [9] Pal U., Wakabayashi T., and Kimura F., 2007, "Handwritten Bangla compound character recognition using gradient feature", *IEEE International conference on information technology*, pp. 208-213.
- [10] Deshpande P., Malik L., and Arora S., 2007, "Handwritten Devanagiri character recognition using connected segments and minimum edit distance", *IEEE, Region 10 conference*, pp. 1-4.