Geometrical features extraction and KNN based Classification of Handwritten Marathi characters

Parshuram M. Kamble
Department of Computer Science
Solapur University, Solpaur
Maharashtra, India-413255
Email: parshu1983@gmail.com

Ravindra S. Hegadi Department
of Computer Science Solapur
University, Solpaur Maharashtra,
India-413255
Email: rshegadi@gmail.com

Abstract—Handwritten character recognition of the Marathi language is a challenging task because characters are complex in structure. 31320 samples of characters from different writers have been collected and database is prepared. Noise is removed by using morphological and thresholding operation. Skewed scanned pages and segmented characters are corrected using Hough Transformation. The characters are segmented from scanned pages by using bounding box techniques. Size variation of each handwritten Marathi characters are normalized in 40×40 pixel size. Here we propose feature extraction from handwritten Marathi characters using connected pixel based features like area, perimeter, eccentricity, orientation and Euler number. The k-nearest neighbor (KNN) algorithm with five fold validation has been used for result preparation. The accuracy of proposed method is 85.88 % obtained.

Keywords-Pre-processing; Geomatrical feature; Marathi Character; KNN Classification

I. Introduction

Handwritten character detection and recognition in general have a list of related applications, such as information retrieval or automatic indexing, such as document indexing, content based image retrieval and postal address recognition application, which further opens up the possibility for advanced system. Marathi language uses 63 phonemic letters, divided into three groups namely swaar (Vowels: 12 letters, as shown in Figure 1 , Vyanjan (Consonants: 38 letters), Ankh (numbers: 10 digits) and Modifiers (Diacritic: 12 letters) [1], [2]. Development of off-line and On-line OCR for (MHC) Marathi handwritten characters is challenging work for researchers because handwriting of each person are mimetic. D. V. Rojatkar et. al. [3] proposed Handwritten Devanagari consonants recognition using Multilayer Probability neural network (MLPNN) with five fold cross validation. In this work handwritten Marathi consonants characters were used for experiment. The neural network is trained three times by varying neurons and classification performance are tested using five fold cross validation. U. Bhattacharya and S. K. Paru proposed a novel scheme for on-line handwritten character recognition based on Levenshtein Distance matric [4]. In this work shape and position of characters were used as a features. The shape information of character was calculated using quantized values of angular displacement between successive sample point along the trajectory of (HC) handwritten characters. Vikas Dongare et.al. [5] proposed (DHNR) Devanagari Handwritten Numeral Recognition using Geometric Features and Statistical Combination Classifier. In this paper they used 17 geometric features based on pixel connectivity, line direction, lines, image area, perimeter, orientation etc. and 5 discriminant functions namely, quadratic linear, Mahalanobis and bi-quadratic distance were used for classification. A similar work was proposed by Kamble et. al. [1] in which features such as eccentricity, orientation and mass of characters were extracted and minimum distance classifier was used for classification. In this work authors calculated the eccentricity, orientation and mass of character feature, minimum distance was used for classification. In another work by Kale et. al. [2] Zernike moment feature extraction for handwritten Devanagari compound character recognition. Here authors proposed Zernike moment based feature descriptor for (HCM) handwritten compound character and Support Vector Machine (SVM), K-NN based classification system. Dixit A. et.al. [6] proposed handwritten Devanagari character recognition using wavelet based feature extraction and classification scheme. In this character image was decomposed using wavelet transform and statistical parameters are calculated as feature vector. This feature vector was used as input for back propagation neural networks during training and testing. Hegadi et. al. [7] proposed recognition of Marathi handwritten numerals using multi-layer feed-forward neural network. In this experiment they segmented each numeral from the document and resized it to 7×5 pixels using cubic interpolation. These resized numerals were converted into a vector with 35 values. These vector values were used as input to train the neural network. N. Sharma et.al. [8] proposed recognition of Off-Line handwritten Devanagari characters using Quadratic Classifier. In this work authors proposed a quadratic classification based scheme for the recognition of off-line Devanagari HC. The directional chain code information of counter point of the characters were extracted as features. Based on the chain code histogram,

978-1-5090-5573-9/17 $31.00 © 2017 IEEE
DOI 10.1109/WCCCT.2016.61

Authorized licensed use limited to: MIT-World Peace University. Downloaded on May 20,2022 at 08:11:49 UTC from IEEE Xplore. Restrictions apply.

they used 64 dimensional features and quadratic classification for recognition. In this experiment they obtained 80.36% recognition accuracy for handwritten Marathi characters. In another work by Kamble et. al. Handwritten Marathi Character Recognition Using (R-HOG) Rectangle based Histogram Oriented Gradients Feature and Artificial Neural Network (ANN) and SVM based classification is proposed [9]. In this paper we propose Local Informative k-nearest neighbor (KNN) algorithm for classification of HMC using features such as area, perimeter, eccentricity, orientation and Euler number. The proposed methodology is discussed in Section II, details discussion on feature extraction and classification techniques are discussed in Section III which are used for this work, The experimental setup and result are discussed in section IV. Finally conclusion of the proposed system is discussed in section V.

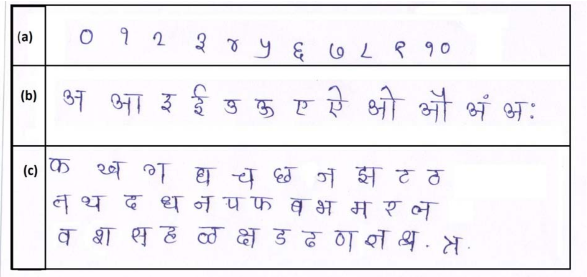

Figure 1. Sample Handwritten Marathi Characters (a) Numerals, (b) Vowels and (c) Consonants.

II. Proposed Method

In this paper, we propose statistical based feature extraction and KNN based classification for the handwritten Marathi characters. Proposed method consists of pre-processing, Segmentation, feature extraction, and classification stages. The handwritten character recognition system have feature extraction and classification are two important stages. Our work the consists of preparation of the standard database for the Marathi handwritten character images and extraction of geometrical features. In pre-processing stage the images are segmented into individual characters and converted in to binary form. Preprocessing of these character images is essential before feature extraction stage. In this stage we extract a set of geometrical features such as centroid, eccentricity, center mass of gravity for each and every character. These features are used in the classification stage. The main objectives of pre-processing are binarization of input image, noise reduction, normalization, skew correction and slant removal. The binarization of image is done by applying Otsu technique [10]. This process converts the image into two components, namely, the object component and background. The object component contains the characters and background contains the noise and other unwanted information. During the scanning of input handwritten Marathi characters document, noise may be generated due to device error, lighting condition and spread of ink in the pen while writing. There are possibility of small breaks and gaps in the characters. We applied smoothing median 3 × 3 pixel size filter and morphological opening with 3 × 3 square shape structure element to remove such kinds of noises and links breaks in the characters. Scanned input document pages have some skew which is removed by using Hough transformation [11]. After this process each character is segmented from scanned document by using bounding box. During the writing of handwritten Marathi characters some text will be in different size. The size normalization task will reduce each character image in to a vertical letter of uniform height and made up one pixel wide stroke. Normalization is applied on each character to bring all the characters to uniform size of 40 × 40 pixels. This process makes recognition operation process independent of the writing size and scanning resolution. Figure 2(a) shows a character from the original document image. The images after scanning will be in the form of gray-scaling, which will be converted to binary form as shown in Figure 2(b). Due to scanning errors small gaps may be produced in the formation of characters. These gaps are removed by morphological techniques the character as shown in Figure 2(c).

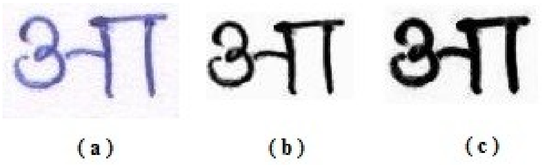

Figure 2. Character in processing stages (a) Character in original form, (b) its binary form and (c) the dilated character.

III. Feature Extraction

After performing different processing steps over MHC, various geometric features are extracted. In this proposed method we calculate area, perimeter, eccentricity, orientation and Euler number on the basis of pixel connected components.

A. Eccentricity

Shape, size and orientation of Marathi characters are heterogeneous. Generally, shape of handwritten Marathi vowels are like an oval shape. We used eccentricity of character as one of the feature for our work. Eccentricity is the ratio of major axis and minor axis of ellipse which covers the entire character. Eccentricity is given by

$$\text{Eccentricity} = \frac{\mathrm{M\,ax}_{\text{axes}}}{\mathrm{M\,in}_{\text{axes}}} \tag{1}$$

Eccentricity is calculated for all characters with connected regions and discarded all regions whose eccentricity is

Authorized licensed use limited to: MIT-World Peace University. Downloaded on May 20,2022 at 08:11:49 UTC from IEEE Xplore. Restrictions apply.

greater than 0.89, since this value corresponds to the noise region. In Figure 3 the doted red line is the ellipse region of handwritten Marathi letter a and the blue lines are the major and minor axes.

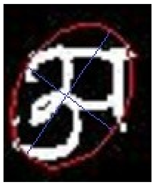

Figure 3. Red line is the ellipse around the letter a and blue lines are major and minor axes

B. Orientation

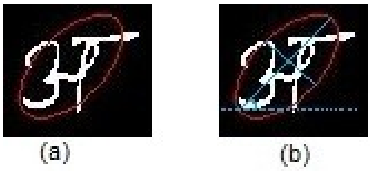

Figure 4. (a) The sample letter a (b) Image showing major axes and doted horizontal blue line as x- axis.

Angle of orientation (in degrees ranging from -90 to 90 degrees) is the angle between major axis of the oval which covers the character and x-axis, as shown in Figure 4. Solid blue lines are axes of the ellipse and red dots are the foci of covered character region. The orientation is the angle between the horizontal dotted line H_{axes} and the major axis M_{axes}, which is calculated by using Equation 2.

$$\tan(\Theta) = \frac{H_{axes} - Max_{axes}}{1 + H_{axes} Max_{axes}} \qquad (2)$$

C. Perimeter

It is the distance around the boundary of the region around the handwritten Marathi character. Following Figure 5 shows the sample Marathi character and red line shows the perimeter.

Figure 5. The sample letter the red line is perimeter.

D. Area

Area of handwritten Marathi character is the actual number of pixel in the region. In binary handwritten Marathi character there are two values for pixels: 0 and 1. 0 represent background region and 1 represent actual character region. The sum of 1's is the area of handwritten Marathi character image.

E. Euler number

It is one of the topological features defined by the number of holes and connected components in the image region. This property will not get affected by character rotation, stretching or transformation. The number of holes H and connected components C can be used to define the Euler number E by Equation (3).

$$E = C - H \qquad (3)$$

F. Local Informative K-Nearest Neighborer Classification

We computed the four features, namely, total mass of character, center of mass, eccentricity and orientation, for one set of characters from a to a, and stored same in database. The above said features were computed for different sets of characters samples and features values of each character is computed and stored in database.

In pattern recognition the k-NN algorithm is one of the methods for classifying objects based on closest training examples in the feature space. k-NN is a type of instance-based learning where the function is only appreciated locally and all computation is deferred until classification [12].

We propose to make use of the distance metric defined in Equation 4 by restricting the computation between the nearest neighbors in an instance query based k-NN classifier. Local Informative KNN retrieves $\mathbf{I}$ locally informative points by first getting the k nearest neighbors (we consider the city block distance here). The entire training set during learning assigns to each query a class represented by the majority label of its k nearest in the training set. We use the following naming conventions: K indicates the k-nearest neighbors according to distance metric, Q denotes the query point and x_i denotes the i's feature vector, x_{ij} its $\mathbf{j}$'s future and y_i, $\mathbf{I}$ denotes most informative points based on equation for each point, its class label and N represent the total number of training points, where each point has P features.

$$\mathbf{I}(x_i \mid Q = x_i) = -\log(1 - \mathbf{P}(x_i \mid Q = x_i))? \\ \mathbf{P}(x_i \mid Q = x_i), \mathbf{j} = 1, ..., N; \mathbf{j} = i \qquad (4)$$

where $\mathbf{P}(x_i|Q = x_i)$ is probability that point x_i is informative with respect to Q defined as:

$$\mathbf{P}(x_i \mid Q = x_i) = \frac{1}{Z} P\,\mathbf{r}(x_i \mid Q = x_i)^{\eta} \\ \prod_{n=1}^{n} (1 - pr(x_j \mid (Q = x_n))I_{[y_j = y_n]} \Big)^{1-\eta}$$

The first term $P\,\mathbf{r}(x_i|Q = x_i)^{\eta}$ in Equation 5 can be interpreted as the likelihood that point x_j is close to the

Authorized licensed use limited to: MIT-World Peace University. Downloaded on May 20,2022 at 08:11:49 UTC from IEEE Xplore. Restrictions apply.

Q while the second part indicates the probability that x_i far apart from dissimilar points. We have proposed to compute the in formativeness of point in the handwritten Marathi character dataset for specific Q. However, considering the high dimensionality and large number of data points, the computational cost could be prohibitively high. Finally confusion matrix are calculated using training and testing datasets.

IV. Experimental Result

The Experimentation is carried out using Matlab 8.0 tool. 31320 different HC of Marathi Language were used for this experimentation, with five-fold validation. The dataset is manually classified in to three sets Vowels, Consonants, Numerals and finally mixed all sets. From the each set the characters of three features, namely, eccentricity, orientation and area of character were obtained and average value is computed for each character. Based on the KNN classifier with k = 4 value, classification is done. Table I shows the classification accuracy for each of these character set. It can be noticed that the vowel character set such as aO, i and ao were classified with very high rate of accuracy, whereas our technique has performed very poor for the characters like e, e and I. The rate of correct classification of e is poor due to the fact that the part of character in upper portion of shirorekha will be disjoint from the remaining part of the character, due to which it will be treated as a separate character. Hence in many cases this character may be falsely classified as e instead of e. In the consonants set few character have small variation in shapes like B, m due to that fact confusion average rate is 8.20 %. When the three sets are combined for this experiment then overall accuracy is 81.48%.

Table I
Classification performance of each character set

Character set	Samples	Correct Classi fication	False Classi fication	Accuracy in % in %
Vowels	4800	3983	87	82.97
Consonants	6400	5402	525	84.40
Numerals	20120	19785	1847	98.33
Mixed	31320	25521	3621	81.48

Average Accuracy of Vowels, Consonants and Numerals is 88.56%.

V. Conclusion

In this paper we have proposed geometrical based feature extraction on Marathi Handwritten character recognition. We can apply two stage recognition approaches to improve the performance of the scheme. The main characteristics of the handwritten Marathi characters is their shapes which are mostly formed with more curves. Most of the failures in recognition are due to either characters with sharp edges and corners, or breaking of a characters making it as separate characters. The post processing can definitely improve the performance which we will undertake in our feature work.

References

[1] P. M. Kamble and R. S. Hegadi, “Handwritten marathi basic character recognition using statistical method,” in Emerging Research in Computing, Information, Communication and Applications, vol. 3. Elsevier, 2014, pp. 28–33.

[2] K. V. Kale, P. D. Deshmukh, S. V. Chavan, and M. M. Kazi, “Zernike moment feature extraction for handwritten devanagari compound character recognition,” Science and Information Conference (SAI), pp. 459–466, 2013.

[3] D. V. Rojatkar, K. D. Chinchkhede, and G. G. Sarate, “Handwritten devnagari consonants recognition using mlpnn with five fold cross validation,” International Conference on Circuits, Power and Computing Technologies, pp. 1222–1226, 2013.

[4] U. Bhattacharya and S. K. Parui, “Online handwriting recognition using levenshtein distance metric,” Document Analysis and Recognition, pp. 79–83, 2013.

[5] D. Vikas and V. Mankar, “Devnagari handwritten numeral recognition using geometric features and statistical combination classifier,” International Journal on Computer Science and Engineering, vol. 5, no. 10, pp. 856–863, 2013.

[6] A. Dixit, A. Navghane, and Y. Dandawate, “Handwritten devanagari character recognition using wavelet based feature extraction and classification scheme,” India Conference (INDICON), pp. 1–4, 2014.

[7] R. S. Hegadi and P. M. Kamble, “Recognition of marathi handwritten numerals using multi-layer feed-forward neural network,” in Computing and Communication Technologies (WCCCT), 2014 World Congress on. IEEE, 2014, pp. 21–24.

[8] N. Sharma, U. Pal, and F. Kimura, “Recognition of off-line handwritten devnagari characters using quadratic classifier,” Computer Vision, Graphics and Image Processing, vol. 4338, pp. 805–816, 2006.

[9] P. M. Kamble and R. S. Hegadi, “Handwritten marathi character recognition using r-hog feature,” Elsevier Procedia Computer Science, vol. 45, pp. 266–274, 2015.

[10] N. Otsu, “A threshold selection method from gray-level histograms,” Automatica, vol. 11, no. 285-296, pp. 23–27, 1975.

[11] T. A. Jundale and R. S. Hegadi, “Skew detection and correction of devanagari script using hough transform,” Elsevier Procedia Computer Science, vol. 45, pp. 305–311, 2015.

[12] Y. Song, J. Huang, D. Zhou, H. Zha, and C. L. Giles, “Iknn: Informative k-nearest neighbor pattern classification,” in Knowledge Discovery in Databases: PKDD 2007. Springer, 2007, pp. 248–264.

Authorized licensed use limited to: MIT-World Peace University. Downloaded on May 20,2022 at 08:11:49 UTC from IEEE Xplore. Restrictions apply.