Comparative Study of Handwritten Marathi Characters Recognition Based on KNN and SVM Classifier

Parshuram M. Kamble(✉) and Ravindra S. Hegadi

Department of Computer Science, Solapur University,
Solapur 413255, India
parshu1983@gmail.com, rshegadi@gmail.com

Abstract. Robust handwritten Marathi character recognition is essential to the proper function in document analysis field. Many researches in OCR have been dealing with the complex challenges of the high variation in character shape, structure and document noise. In proposed system, 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 modified k-nearest neighbor (KNN) and SVM algorithm with five fold validation has been used for result preparation. The comparative accuracy of proposed methods are recorded. In this experiment modified SVM obtained high accuracy as compared with KNN classifier.

Keywords: Geometrical feature · Marathi character · KNN and SVM classification · Feature extraction

1 Introduction

Automatic object detection and recognition from a digital image are the field of pattern recognition. In the pattern recognition feature extraction and classification are two important stages, in pattern recognition area researchers worked on various field like face recognition, Content based image retrieval, biometric and Optical character recognition (OCR). The OCR again divided in two parts machine printed and hand written. 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. Optical character detection and recognition is used in various applications, such as document indexing, postal address recognition, number plate recognition, information retrieval and office automation. Marathi language is belongs to Devanagari script, it have 63 phonic

K.C. Santosh et al. (Eds.): RTIP2R 2016, CCIS 709, pp. 93–101, 2017.
DOI: 10.1007/978-981-10-4859-3_9

letters, further they subdivided into three groups namely (vowels: 12 letters, as shown in Fig. 1, Vyanjan (Consonants: 38 letters), Ankh (numbers: 10 digits) and Modifiers (Diacritic: 12 letters) [6,8].

U. Bhattacharya and S.K. Paru proposed a novel approach Levenshtein Distance metric for on-line handwritten character recognition [1]. 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. They formulated a distance function based on Levenshtein metric to compute the similarity between unknown sample and training sample. D.V. Rojatkar et al. [10] 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.

Vikas Dongare et al. [13] proposed based on Geometric Features and Statistical Combination Classifier for (DHNR) Devanagari Handwritten Numeral Recognition. 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. [8] in which features such as eccentricity, orientation and mass of characters were extracted and minimum distance classifier was used for classification. In another work by Kale et al. [6] Zernike moment based 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. [3] proposed schemes for handwritten Devanagari character recognition based on 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. [4] proposed sytem for Marathi handwritten numerals recognition based on multi-layer feed-forward neural network and cubic interpolation feature. N. Sharma et al. [11] proposed recognition of Off-Line handwritten Devanagari characters using Quadratic Classifier. They 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, 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 [7].

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 Sect. 2, details discussion on feature extraction and classification techniques are discussed in Sect. 3 which are used for this work. The experimental setup and result are discussed in Sect. 4. Finally conclusion of the proposed system is discussed in Sect. 5.

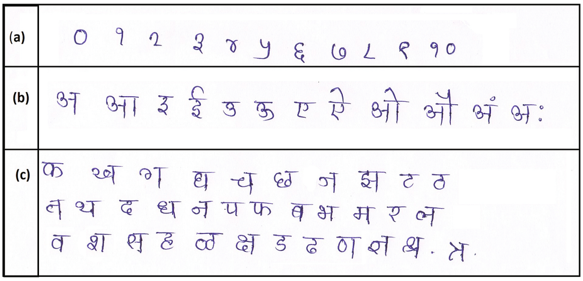

Fig. 1. Sample handwritten Marathi characters (a) Numerals, (b) Vowels and (c) Consonants.

2 Proposed Method

We propose statistical based feature extraction and KNN, SVM 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. Pre-processing 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 [9]. In this process image converts into two components, object components and background. In the object components contains actual object and background contains 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 [5]. 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. After this stage each character is normalized in 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 Fig. 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 Fig. 2(c).

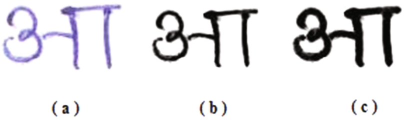

Fig. 2. Shows character pre-processing stages (a) Original character sample (b) Binary (c) Dilated character

3 Feature Extraction

After performing different pre-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.

3.1 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

$$Eccentricity = \frac{Max_{axes}}{Min_{axes}} \tag{1}$$

Eccentricity is calculated for all characters with connected regions and discarded all regions whose eccentricity is greater than 0.89, since this value corresponds to the noise region. In Fig. 3 the doted red line is the ellipse region of handwritten Marathi letter अ and the blue lines are the major and minor axes.

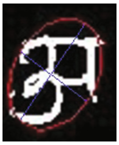

Fig. 3. Red line is the ellipse around the letter अ and blue lines are major and minor axes (Color figure online)

3.2 Orientation

Angle of orientation (in degrees ranging from −90 to 90°) is the angle between major axis of the oval which covers the character and x-axis, as shown in Fig. 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 Max_{axes}, which is calculated by using Eq. 2.

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

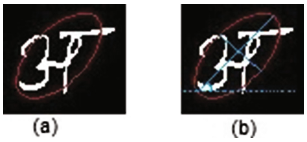

Fig. 4. (a) The sample letter अ (b) Image showing major axes and doted horizontal blue line as x-axis. (Color figure online)

3.3 Perimeter

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

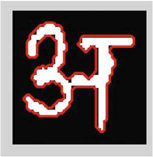

Fig. 5. The sample letter the red line is perimeter. (Color figure online)

3.4 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.

3.5 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 Eq. (3).

$$E = C - H \tag{3}$$

3.6 K-Nearest Neighborer Classification

We computed the four features, namely, total mass of character, center of mass, eccentricity and orientation 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].

Classification of object is the heart of any pattern recognition system. The K-NN classification model widely used in various pattern recognition system like OCR and Biometric. The KNN classify the object based on the feature space. In this proposed model we trained the datasets for this algorithms with respect to feature vector and class label. In classification process the query sample assigned to the label of its k nearest neighbors.

The query sample classifies based on the labels of its k nearest neighbors by majority vote. The different distance functions like Euclidean, Manhattan, Mankowski and Chebyshev are used to result preparation.

3.7 Support Vector Machine Classification

In supervised learning model support vector machine algorithms is linear machine learning approach, it also know as support vector network that are used for classification and regression. The SVM are modified in non-linear classification based on kernel function. The kernel is the adjustable parameter of SVM. Given training set of instance label pairs $(C_i, L_i), i = 1 \ldots n$ where C_i and L_i are histograms from training images. The function θ maps training vectors x_i into a higher dimensional feature space while $C > 0$ is the penalty parameter for the error term. We used basic linear and radial basis function (RBF). The radial basis and linear function are given as:

$$K(x_i, x_i) = x_i^T x_i \quad And \quad \exp(\gamma||x_i - x_i||^2), \gamma > 0 \tag{4}$$

The Radial Basis Function (RBF) provides good performance on handwritten character images. After configuring the kernel function and its parameter, the SVM is applied to classify the trained datasets.

4 Experimental Result

The Experimentation is carried out using Matlab 8.0 tool with Intel core $i5$, 16 GB RAM machine. Bhattacharya and B.B. Chaudhuri [2] prepared 17271 handwritten numeral datasets and 31320 (our datasets) 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, Figure shows the sample of datasets (Fig. 6).

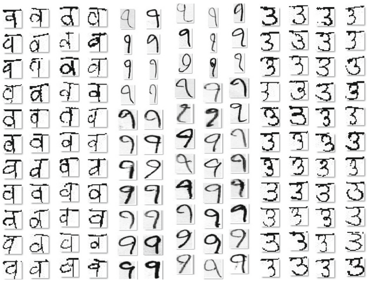

Fig. 6. Marathi handwritten dataset sample.

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 and SVM classifier with diffrent kernel function, classification is done. Table 1 shows the classification accuracy for each of these character set. It can be noticed that the vowel character set such as औ,इ and ओ were classified with very high rate of accuracy, whereas our technique has performed very poor for the characters like ए, ऐ and ई. The rate of correct classification of ऐ 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 ए instead of ऐ. In the consonants set few character have small variation in shapes like भ, म due to that fact confusion average rate is 8.20%. When the three sets are combined for this experiment then overall accuracy with respect to KNN 88.53% and SVM 80.25%.

Average Accuracy of Vowels, Consonants and Numerals with respect to SVM is 95.35 and KNN 91.52%.

Table 1. Classification performance of each character set with respect to KNN and SVM classifier

Character set	Samples	SVM	KNN %
Vowels	4800	94.26	88.77
Consonants	6400	93.24	90.12
Numerals	20120	98.56	95.67
Mixed	31320	88.53	80.25
Numerals (U. Bhattacharya and B.B. Chaudhuri)	17271	98.52	97.00

5 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. Bhattacharya, U., Parui, S.K.: Online handwriting recognition using levenshtein distance metric. In: Document Analysis and Recognition, pp. 79–83 (2013)
2. Bhattacharya, U., Chaudhuri, B.B.: Handwritten numeral databases of Indian scripts and multistage recognition of mixed numerals. IEEE Trans. Pattern Anal. Mach. Intell. **31**(3), 444–457 (2009)
3. Dixit, A., Navghane, A., Dandawate, Y.: Handwritten Devanagari character recognition using wavelet based feature extraction and classification scheme. In: India Conference (INDICON), pp. 1–4 (2014)
4. Hegadi, R.S., Kamble, P.M.: Recognition of Marathi handwritten numerals using multi-layer feed-forward neural network. In: World Congress on Computing and Communication Technologies (WCCCT), pp. 21–24. IEEE (2014)
5. Jundale, T.A., Hegadi, R.S.: Skew detection and correction of Devanagari script using hough transform. Elsevier Procedia Comput. Sci. **45**, 305–311 (2015)
6. Kale, K.V., Deshmukh, P.D., Chavan, S.V., Kazi, M.M.: Zernike moment feature extraction for handwritten Devanagari compound character recognition. In: Science and Information Conference (SAI), pp. 459–466 (2013)
7. Kamble, P.M., Hegadi, R.S.: Handwritten Marathi character recognition using r-hog feature. Elsevier Procedia Comput. Sci. **45**, 266–274 (2015)
8. Kamble, P.M., Hegadi, R.S.: Handwritten Marathi basic character recognition using statistical method. In: Emerging Research in Computing, Information, Communication and Applications, vol. 3, pp. 28–33. Elsevier (2014)
9. Otsu, N.: A threshold selection method from gray-level histograms. Automatica **11**(285–296), 23–27 (1975)

10. Rojatkar, D.V., Chinchkhede, K.D., Sarate, G.G.: Handwritten Devnagari consonants recognition using mlpnn with five fold cross validation. In: International Conference on Circuits, Power and Computing Technologies, pp. 1222–1226 (2013)
11. Sharma, N., Pal, U., Kimura, F., Pal, S.: Recognition of off-line handwritten Devnagari characters using quadratic classifier. In: Kalra, P.K., Peleg, S. (eds.) ICVGIP 2006. LNCS, vol. 4338, pp. 805–816. Springer, Heidelberg (2006). doi:10.1007/11949619_72
12. Song, Y., Huang, J., Zhou, D., Zha, H., Giles, C.L.: IKNN: informative k-nearest neighbor pattern classification. In: Kok, J.N., Koronacki, J., Lopez de Mantaras, R., Matwin, S., Mladenič, D., Skowron, A. (eds.) PKDD 2007. LNCS (LNAI), vol. 4702, pp. 248–264. Springer, Heidelberg (2007). doi:10.1007/978-3-540-74976-9_25
13. Vikas, D., Mankar, V.: Devnagari handwritten numeral recognition using geometric features and statistical combination classifier. Int. J. Comput. Sci. Eng. **5**(10), 856–863 (2013)