International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-9, Sept. 2018 E-ISSN: 2347-2693

Online Handwritten Gujarati Numeral Recognition Using Support Vector Machine

V. A. Naik[1*], A. A. Desai[2]

[1,2]Department of Computer Science, Veer Narmad South Gujarat University, Surat, Gujarat, India

** Corresponding Author: vishalkumar.naik@gmail.com*

Available online at: www.ijcseonline.org

Accepted: 22/Sept/2018, Published: 30/Sept/2018

***Abstract* -** In this paper, online handwritten numeral recognition for Gujarati is proposed. Online handwritten character recognition is in trend for research due to a rapid growth of handheld devices. The authors have compared Support Vector Machine (SVM) with linear, polynomial, and radial basis function kernels. The authors have used hybrid feature set. The authors have used zoning and chain code directional features which are extracted from each stroke. The dataset of the system is of 2000 samples and was collected by 200 writers and tested by 50 writers. The authors have achieved an accuracy of 92.60%, 95%, and 93.80% for linear, polynomial, RBF kernel and an average processing time of 0.13 seconds, 0.15seconds, and 0.18 seconds per stroke for linear, polynomial, RBF kernel.

***Keywords*:** Online Handwritten Character Recognition (OHCR), Handwritten Character Recognition (HCR), Optical Character Recognition (OCR), Support Vector Machine (SVM), Gujarati Numeral, Gujarati Digits

I. INTRODUCTION

Gujarati is an Indo-Aryan language and one of the official languages of India, spoken by people of Gujarat state, union territories of Diu - Daman and Dadra - Nagar Haveli.
In Offline Handwritten Character Recognition, input to a system is an image of handwritten text captured by scanner and system converts it into text. A notable work is done by many researchers on offline handwritten character recognition for Gujarati language.

Online character recognition system records the movement of a pen or other device and converts it into text. It provides a platform to generate computerized handwritten data and helps users to improve interaction with a computer. Recently due to the increasing use of touchscreen devices, many researchers are working in online character recognition in different languages. Some work is available for online recognition of Hindi, Malayalam, Assamese, Tamil, Devanagari, Bangla, Kannada, Telugu, and Gurumukhi languages.

Gujarati characters are different than other sister languages because it does not have Shirolekha over characters. Unlike other languages, Gujarati digits have different types of curves and different writing styles. Figure 1 illustrates various handwritten Gujarati digits.

There are many challenges in recognition of Gujarati digits because of variation in writing style and handwriting. Gujarati digits have more curves than lines and there are similar curves in some characters. Because of such similarities between some characters, there is a higher possibility of misclassification.

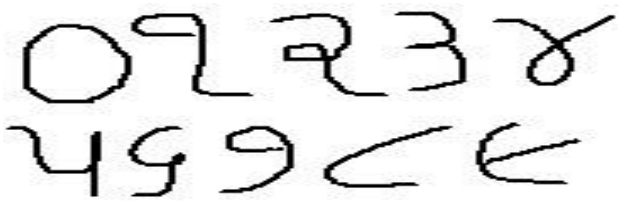

Figure.1 Handwritten Gujarati Numerals

The rest of the paper is organized as follows, section II contains related work, section III contains feature extraction methods, section IV contains classification process., section V contains result discussion, and section VI contains conclusion of the paper.

II. RELATED WORK

Many researchers are working in the area of handwritten character recognition for different Indian languages. Some researchers are working on offline handwritten character recognition for Gujarati.

C. Patel and A. A. Desai [1] have proposed segmentation of text lines into words. They have used projection profile and morphological operations for segmentation. They [2] have proposed zone identification for words. They have used distance transform method for identification of zone like upper, middle, and lower. They [3] have proposed handwritten character recognition system. They have used hybrid classifier using tree and k-NN. They have used structural and statistical features. They have achieved an accuracy of 63%.

A. A. Desai [4] has proposed character segmentation from old documents. He has used some pre-processing methods and Radon transform for segmentation.

He [5] has proposed character recognition system for Gujarati numerals. He has used binarization, size normalization and thinning pre-processing methods. He has used hybrid features like a subdivision of skeletonized image and aspect ratio. He has used k-NN classifier with Euclidean distance method and achieved 96.99% accuracy. He [6] has proposed similar work using profile vector-based features. He has used multilayer feed forward neural network. He has achieved an accuracy of 82%. He [7] has proposed similar work using hybrid feature set that includes aspect ratio, extent, and zoning. He has used SVM with a polynomial kernel for classification and achieved an accuracy of 86.66%. M. Maloo and K. V. Kale [8] have proposed handwritten numeral recognition system for Gujarati. They have used pre-processing methods like binarization, dilation, and skeletonization. They have used affine invariant moments (AMI) for feature extraction and SVM for classification and achieved 91% accuracy. M. B. Mendapara and M. M. Goswami [9] have used binarization, noise removal, and thinning pre-processing methods. They have used stroke based directional feature and used k-NN as a classifier. They have achieved 88% accuracy. R. Nagar and S. Mitra [10] have used binarization and thinning pre-processing methods. They have used orientation estimation features and SVM as a classifier and achieved 98.97% accuracy. A. Vyas and M. Goswami [11] have used binarization, noise removal, and thinning pre-processing methods. They have used modified chain code, Discrete Fourier Transform, and Discrete Cosine Transform as a feature. They have used k-NN, SVM and ANN as a classifier and achieved 85.67%, 93.60%, and 93.00% accuracy respectively. Prutha Y M and Anuradha S G [12] have proposed real time traffic analysis system. They have used different morphological and edge detection techniques.

Many researchers are working in Online Handwritten Character Recognition for different Indian languages, but no work is found for the Gujarati language till date.

In Hindi online handwritten character recognition, M. Abuzaraida, A. Zeki, and A. Zeki [13] have used basic preprocessing methods and used chain code as a structural feature. They have used global alignment algorithm (GAA) for classification and achieved 96% accuracy. S. Belhe et al. [14] have used smoothing using a Gaussian filter and normalization pre-processing methods. They have used a Histogram of oriented gradient (HOG) as a feature. They have used a Gaussian kernel HMM for training symbol and stroke group-based tree structure for word recognition and achieved 89% accuracy.

In Malayalam online handwritten character recognition, S. Joseph and A. Hameed [15] have used basic preprocessing methods and used six-time domain features with directional and curvature features. They have used SVM as a classifier and achieved 95.45% accuracy. Anoop M. Namboodiri [16] have presented work on Malayalam and Telugu language. They have used normalization, resampling using a Gaussian low-pass filter and an equidistant resampling to remove variations in writing speed. They have used moments of the stroke, direction, curvature, length, an area of the stroke, aspect ratio as features. They have used SVM using a Decision Directed Acyclic Graph (DDAG) and discriminative classifier. They have achieved an accuracy of 95.78% on Malayalam and 95.12% on Telugu. Primekumar K.P. and S. Idiculla [17] have used duplicate point elimination, smoothing, normalization, resampling as preprocessing methods. They have used x-y coordinates, angular features, direction, and curvature are extracted. Using HMM classifier, they have used k means using Euclidean distance for training and using SVM classifier, they have used discrete wavelet transform for training. They have achieved an accuracy of 97.97% using SVM and 95.24% using HMM.

S. Amritha, C. Tripti, and V. Govindaru [18] have used noise removal and linking broken character preprocessing methods. They have used low level, high level, and directional features. They have used sigmoid function based neural network and used back propagation method to train it. They have used the disambiguation technique as a post-processing method for confusion sets.

In Assamese online handwritten numerals recognition, G. Siva Reddy et al. [19] have used remove duplicate points, size normalization, smoothing, and linear interpolation and resampling pre-processing methods. They have used preprocessed coordinates as features and used HMM for classification. They have achieved an accuracy of 99.3%.

In Tamil, online handwritten character recognition, A. Bharath and S. Madhavanath [20] have presented work on Tamil and Devanagari. They have used a set of nine features like, writing direction and curvature, aspect, curliness, linearity, and slope. They have used HMM for symbol modeling. They have achieved an accuracy of 91.8% for Tamil and 87.13% for Devanagari.

K.H. Aparna et al. [21] have used preprocessing methods like interpolation, smoothing, and normalization of strokes. They have used 18 shape-based features. They have converted a group of stroke labels into a suitable character code. They have achieved an accuracy of 82.80%.

In Devanagari online handwritten character recognition, H. Swetha Lakshmi et al. [22] have presented work on Devanagari and Telugu. They have used normalization and smoothing preprocessing methods. They have used SVM and HMMs for classification. A. Sharma and K. Dahiya [23] have presented work for Devanagari and Gurumukhi languages. They have used size normalization, interpolating missing points, smoothing, slant detection preprocessing methods. They have used directional, local features. They have used the K-means clustering technique for a classification. They have achieved an accuracy of 94.69% for Gurumukhi and 86.90% for Devanagari.

In Bangla online handwritten character recognition, N. Bhattacharya et al. [24] have used offline horizontal histogram for segmentation. They have used chain code as a feature. They have used SVM as a classifier and achieved an accuracy of 97.45%. S.K.Parui et al. [25] have used noise removal, scaling, smoothing as preprocessing methods. They have used stroke-based features. They have used HMM as a classifier and achieved an accuracy of 87.7%. C. Biswas et al. [26] have removed repeated points, scaling, smoothing and normalization used as a preprocessor and used stroke-based features. They have used HMM for classification and achieved an accuracy of 91.85%.

In Kannada online handwritten character recognition, K. Prasad G. et al. [27] have used principal component analysis (PCA) and dynamic time wrapping (DTW) for classification. They have achieved an accuracy of 88% using PCA and 64% using DTW.

In Telugu online Character Recognition, J. Rajkumar et al. [28] have used normalization, smoothing and interpolation as preprocessing methods. They have used histogram based and SVM based methods for stroke pre-classification. They have used X and Y coordinate values, Fourier transforms, Hilbert transform and Wavelet features. They have used SVM for stroke recognition. They have used a ternary search tree and SVM for character recognition. They have achieved an accuracy of 90.55% using a search tree and 96.42% using SVM.

III. FEATURE EXTRACTION

Each stroke has some unique features. Such unique features of stroke are extracted and used for classification of characters. Different methods can be used to extract various features from a stroke. The feature can be of different types like shape based, intensity based, and texture based [29].

We have used different structural and statistical features. These features are extracted from each stroke. The size of stroke and speed of stroke may vary from user to user that may lead to wrong feature extraction. Proposed features are a size and speed independent.

Each stroke drawn by the user is divided into 9 equal zones according to its length and width. Percentage wise distribution of active pixels from each zone is calculated and considered as features. Figure 2 illustrates the zoning of digit 3 with a percentage of active pixels in each zone. Zone of starting and ending of every stroke is considered as features. Stroke's directional feature describes curvature. The chain code is used to get directional details of a stroke. Stroke's direction is measured after every 10% of total drawn pixels. Figure 3 illustrates different chain code values for a different direction. Figure 4 illustrates stroke direction with its specific chain code. The extracted feature set values for character 3 are 2,2,4,6,8,2,2,4,6, 1,1,3,1,8,7,16,9,15,17,7,8,13. First 9 digit indicates chain code values, next 4 digit indicates the starting and end zone and last 9 digits indicate a percentage of active pixels among 9 zones.

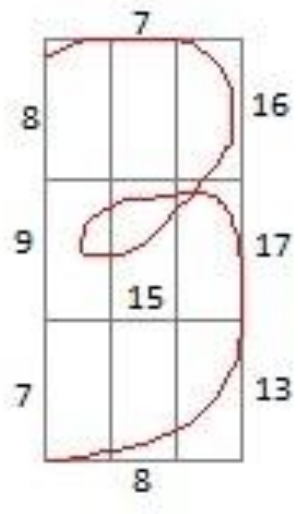

Figure 2. Digit 3 in 9 equal zones with the percentage of active pixels

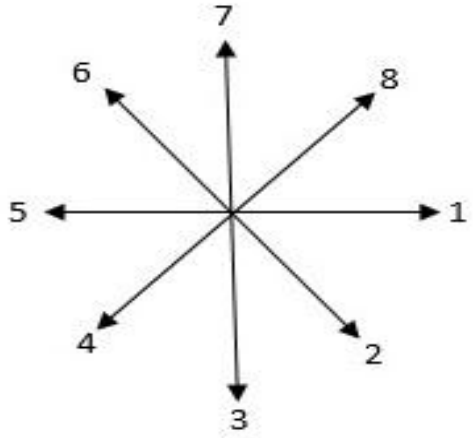

Figure 3. Chain Code for each Direction

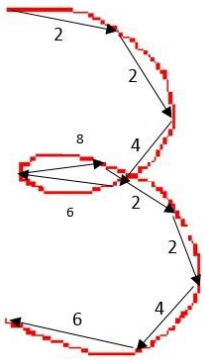

Figure 4. Chain code for Digit 3

IV. CLASSIFICATION

Support Vector Machine (SVM) is a supervised machine learning algorithm that is used as a classifier. It can be used for classification of one or more than one classes.

It uses a different type of kernels to transforms data. This transformed data finds an optimal boundary. SVM can be used with different types of kernels like linear, polynomial, and radial basis function. With the help of these kernels, SVM gain flexibility in the selection of the threshold. SVM finds unique global optimum solution due to quadratic programming and is less prone to overfitting [30]. We have compared linear, polynomial and radial basis function kernels for classification.

V. RESULTS AND DISCUSSION

The proposed system has a training dataset of 2000 samples collected from 200 different writers of different age group and gender, 10 samples were taken from each writer. The proposed System was tested by 50 different writers with 10 samples each. In this system, we have achieved highest accuracy of 95% using the polynomial kernel, 93.8% using RBF kernel, and 92.6% using linear kernel. As illustrated in figure 5, the highest accuracy of 99% for digit 0 using polynomial kernel. An accuracy of digit 0 is 97% using RBF kernel and 95% using a linear kernel. The lowest accuracy of 90% for digit 1 using linear kernel.

Many writers wrote digit 1 in a different style and it has similarities with digit 2 and 7 so digit 1 has minimum accuracy. Digit 9 require multiple strokes to write it so it has minimum accuracy too. Table 1 shows a confusion matrix for SVM Polynomial kernel of all digits of 50 writers.

All the training data and testing data captured by developed GUI system. In this system, we have not used any preprocessing techniques that result in faster execution. The feature set used is size and speed independent. The proposed system took an average processing time of 0.15 seconds per stroke.

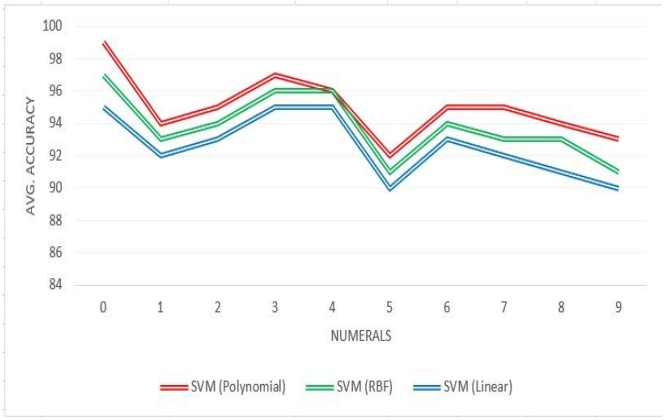

Figure 5. The accuracy comparison for Gujarati Numerals

Table 1. Confusion matrix of Gujarati Numerals

No	0	1	2	3	4	5	6	7	8	9
0	49	0	0	0	0	0	0	1	0	0
1	0	43	4	0	0	1	0	2	0	0
2	0	0	48	2	0	0	0	0	0	0
3	0	0	1	48	0	0	1	0	0	0
4	0	0	2	0	48	0	0	0	0	0
5	0	0	0	0	0	49	1	0	0	0
6	0	1	2	0	0	0	47	0	0	0
7	2	3	0	0	0	0	0	45	0	0
8	0	0	0	0	0	0	0	0	46	4
9	0	0	0	0	0	0	0	0	1	49

VI. CONCLUSION

We proposed an algorithm for online handwritten Gujarati numeral recognition using feature set of 22 different structural and statistical features. Support Vector Machine is used as a classifier with a linear, polynomial, and RBF kernel with a training set of 2000 samples. This work achieved 95% accuracy of character recognition rate using SVM Polynomial with 0.15 seconds of average execution time per stroke.

ACKNOWLEDGMENT

The authors acknowledge the support of the University Grants Commission (UGC), New Delhi, for this research work through project file no. F. 42-127/2013.

REFERENCES

[1] C. Patel and A. Desai, "Segmentation of text lines into words for Gujarati handwritten text," *Proc. 2010 Int. Conf. Signal Image Process. ICSIP 2010*, pp. 130–134, 2010.

[2] C. Patel and A. Desai, "Zone identification for Gujarati handwritten word," *Proc. - 2nd Int. Conf. Emerg. Appl. Inf. Technol. EAIT 2011*, pp. 194–197, 2011.

[3] C. Patel and A. Desai, "Gujarati Handwritten Character Recognition Using Hybrid Method Based on Binary Tree-Classifier And K-Nearest Neighbour," *Int. J. Eng. Res. Technol.*, vol. 2, no. 6, pp. 2337–2345, 2013.

[4] A. Desai, "Segmentation of Characters from old Typewritten Documents using Radon Transform," *Int. J. Comput. Appl.*, vol. 37, no. 9, pp. 10–15, 2012.

[5] A. A. Desai, "Handwritten Gujarati Numeral Optical Character Recognition using Hybrid Feature Extraction Technique," *Int. Conf. Image Process. Comput. Vision, Pattern Recognition, IPCV*, 2010.

[6] A. A. Desai, "Gujarati handwritten numeral optical character reorganization through neural network," *J. Pattern Recognit.*, vol. 43, no. 7, pp. 2582–2589, 2010.

[7] A. a. Desai, "Support vector machine for identification of handwritten Gujarati alphabets using hybrid feature space," *CSI Trans. ICT*, vol. 2, no. January, pp. 235–241, 2015.

[8] M. Maloo, K. V Kale, and I. Technology, "Support Vector Machine Based Gujarati Numeral Recognition," *Int. J. Comput. Sci. Eng. ({IJCSE}), {ISSN} 0975-3397*, vol. 3, no. 7, pp. 2595–2600, 2011.

[9] M. B. Mendapara and M. M. Goswami, "Stroke identification in Gujarati text using directional feature," *Proceeding IEEE Int. Conf. Green Comput. Commun. Electr. Eng. ICGCCEE 2014*, 2014.

[10] N. Rave and S. K. Mitra, "Feature extraction based on stroke orientation estimation technique for handwritten numeral," in *Eighth International Conference on Advances in Pattern Recognition (ICAPR)*, 2015.

[11] A. N. Vyas and M. M. Goswami, "Classification of handwritten Gujarati numerals," *2015 Int. Conf. Adv. Comput. Commun. Informatics, ICACCI 2015*, pp. 1231–1237, 2015.

[12] Y. M. Prutha and S. G. Anuradha, "Morphological Image Processing Approach of Vehicle Detection for Real-Time Traffic Analysis," *Int. J. Comput. Sci. Int. J. Comput. Sci. Eng.*, vol. 3, no. 5, pp. 88–92, 2014.

[13] M. A. Abuzaraida, A. M. Zeki, and A. M. Zeki, "Online recognition system for handwritten hindi digits based on matching alignment algorithm," in *International Conference on Advanced Computer Science Applications and Technologies, ACSAT 2014*, 2014, pp. 168–171.

[14] S. Belhe, C. Paulzagade, A. Deshmukh, S. Jetley, and K. Mehrotra, "Hindi handwritten word recognition using HMM and symbol tree," *Proceeding Work. Doc. Anal. Recognit. - DAR '12*, p. 9, 2012.

[15] S. Joseph and A. Hameed, "Online handwritten malayalam character recognition using LIBSVM in matlab," in *National Conference on Communication, Signal Processing and Networking, NCCSN 2014*, 2015, pp. 1–5.

[16] A. Arora and A. M. Namboodiri, "A hybrid model for recognition of online handwriting in Indian scripts," in *International Conference on Frontiers in Handwriting Recognition, ICFHR 2010*, 2010, pp. 433–438.

[17] K. P. Primekumar and S. M. Idiculla, "On-line Malayalam Handwritten Character Recognition using HMM and SVM," *Int. Conf. Signal Process. , Image Process. Pattern Recognit. [ICSIPR]*, pp. 1–5, 2013.

[18] A. Sampath, C. Tripti, and V. Govindaru, "Online Handwritten Character Recognition for Malayalam," *ACM Int. Conf. Proceeding Ser.*, pp. 661–664, 2012.

[19] G. S. Reddy, P. Sharma, S. R. M. Prasanna, C. Mahanta, and L. N. Sharma, "Combined online and offline assamese handwritten numeral recognizer," in *National Conference on Communications, NCC 2012*, 2012.

[20] A. Bharath and S. Madhvanath, "HMM-based lexicon-driven and lexicon-free word recognition for online handwritten indic scripts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 670–682, 2012.

[21] K. H. Aparna, V. Subramanian, M. Kasirajan, G. V. Prakash, and V. S. Chakravarthy, "Online Handwriting Recognition for Tamil," *Ninth Int. Work. Front. Handwrit. Recognit.*, pp. 438–443, 2004.

[22] H. Swethalakshmi, "Online handwritten character recognition of Devanagari and Telugu Characters using support vector machines," *Tenth Int. Work. Front. Handwrit. Recognit.*, pp. 1–6, 2006.

[23] A. Sharma and K. Dahiya, "Online Handwriting Recognition of Gurmukhi and Devanagari Characters in Mobile Phone Devices," in *International Journal of Computer Applications*, 2012, pp. 37–41.

[24] N. Bhattacharya, U. Pal, and F. Kimura, "A System for Bangla Online Handwritten Text," in *12th International Conference on Document Analysis and Recognition*, 2013, pp. 1335–1339.

[25] S. K. Parui, K. Guin, U. Bhattacharya, and B. B. Chaudhuri, "Online handwritten Bangla character recognition using HMM," in *19th International Conference on Pattern Recognition*, 2008.

[26] C. Biswas, U. Bhattacharya, and S. K. Parui, "HMM based online handwritten Bangla character recognition using Dirichlet distributions," in *International Workshop on Frontiers in Handwriting Recognition, IWFHR*, 2012, pp. 600–605.

[27] G. Keerthi Prasad, I. Khan, N. R. Chanukotimath, and F. Khan, "On-line handwritten character recognition system for Kannada using Principal Component Analysis Approach: For handheld devices," in *2012 World Congress on Information and Communication Technologies*, 2012, pp. 675–678.

[28] J. Rajkumar, K. Mariraja, K. Kanakapriya, S. Nishanthini, and V. S. Chakravarthy, "Two schemas for online character recognition of Telugu script based on Support Vector Machines," in *International Workshop on Frontiers in Handwriting Recognition, IWFHR*, 2012, no. 2004, pp. 565–570.

[29] D. N. Lohare, R. Telgade, and R. R. Manza, "Review Brain Tumor Detection Using Image Processing Techniques," *Int. J. Comput. Sci. Eng.*, vol. 6, no. 5, pp. 735–740, 2018.

[30] K. P. Wu and S. De Wang, "Choosing the kernel parameters for support vector machines by the inter-cluster distance in the feature space," *Pattern Recognit.*, vol. 42, no. 5, pp. 710–717, 2009.

Authors Profile

Mr. V. A. Naik received Master of Science (I.T.) from Veer Narmad South Gujarat University, Surat, Gujarat, India in 2010. He is currently pursuing Ph.D. from Veer Narmad South Gujarat University, Surat, Gujarat, India. His main research work focuses on handwritten character recognition, pattern mataching, and image processing. He has 7 years of teaching experience.

Mr. A. A. Desai received a Ph.D. degree from the South Gujarat University, Gujarat, India in 1997. He is a professor with the Department of Computer Science, Veer Narmad South Gujarat University (VNSGU) India since 2004. He is also Deam of Faculty of Computer Science and Information Technology since 2012. He is involved in many projects during his long academic and research experience of over 25 years. His research activities are in the area of Digital Image Processing, Pattern Recognition, Natural Language Processing, and Data Mining.

