Features Fusion based Approach for Handwritten Gujarati Character Recognition

Ankit Sharma, Priyank Thakkar, Dipak M. Adhyaru, and Tanish H. Zaveri

***Abstract*—Handwritten character recognition is a challenging area of research. Lots of research activities in the area of character recognition are already done for Indian languages such as Hindi, Bangla, Kannada, Tamil and Telugu. Literature review on handwritten character recognition indicates that in comparison with other Indian scripts research activities on Gujarati handwritten character recognition are very less. This paper aims to bring Gujarati character recognition in attention. Recognition of isolated Gujarati handwritten characters is proposed using three different kinds of features and their fusion. Chain code based, zone based and projection profiles based features are utilized as individual features. One of the significant contribution of proposed work is towards the generation of large and representative dataset of 88,000 handwritten Gujarati characters. Experiments are carried out on this developed dataset. Support Vector Machine (SVM) and Naive Bayes (NB) classifier based methods are implemented for handwritten Gujarati character recognition. Experimental results show substantial enhancement over state-of-the-art and authenticate our proposals.**

***Index Terms*—Gujarati handwritten characters, Naive Bayes classification, Support Vector Machine.**

I. Introduction

HANDWRITTEN character recognition is a process of identifying characters from handwritten scanned documents. Handwritten character recognition is a challenging area for research. Gujarati script is part of the Brahmic family. Gujarati is the mother tongue of Gujarat state in India. All over the world more than 65 million people use Gujarati language for their communication purpose. As Gujarat is one of the eminent state of India, Gujarati is a well-known and culturally rich language. Gujarati Character Recognition offers more difficulties like most other Indian languages relative to the western languages due to these reasons: (a) Number of classes are higher, (b) Structure of characters in Indian scripts contains curves, holes and strokes which make large variation in writing style by different persons, (c) Presence of similar looking characters (d) Unavailability of standard data set for testing and experimentation for script like Gujarati.

Research in the field of Gujarati character recognition is still in budding stage. From the literature review it can be noticed that compare to many other scripts OCR activities for Gujarati script is very less. One significant reason for the lack of research activities in the area of Gujarati handwritten character recognition is the unavailability of benchmark dataset.

In this work, a large dataset of 88,000 isolated handwritten Gujarati characters is collected. Three different feature extraction techniques and their fusions are implemented. Methods based on Support Vector Machine (SVM) and Naive Bayes (NB) classifier are utilized for classification purpose.

The rest of the paper is organized as follow. Section 2 presents review of earlier work and section 3 describes the generated handwritten Gujarati character dataset. Pre-processing algorithm is discussed in section 4 and Feature extraction algorithm is described in section 5. Prediction models and proposed approach are discussed in section 6 and 7 respectively. Details about experimentation evaluation is presented in section 8 and paper is concluded in section 9.

II. Review of Related Work

In other Indian languages like Hindi, Kannada, Bangla, Tamil, Telugu there has been tremendous progress in the field of OCR as compare to Gujarati language. This section intends to provide a brief description of research efforts in the area of printed and handwritten Gujarati character recognition. The section initiates with the review of literature in the area of printed Gujarati character recognition followed by the review of the literature on handwritten Gujarati character recognition.

First contribution in the area of printed Gujarati character recognition is made by Antani and Agnihotri [1] in 1999. They utilized K-Nearest Neighbor (KNN) and minimum hamming distance classifiers with regular and invariant moments. The accuracy achieved was 67% and 48% with KNN and minimum hamming distance classifiers respectively. Prof S K Shah and A Sharma [2] implemented a template matching system for printed Gujarati character recognition. Fringe distance is used for the comparison of input image with the template. Experiment is performed on small dataset of 1375 images. They achieved the overall accuracy of 72.3%. Jignesh Dholakia,

A. Sharma is with the Department Instrumentation and Control Engineering, Institute of Technology, Nirma University, Gujarat, India.
e-mail: (ankit.sharma@nirmauni.ac.in).

P. Thakkar is with the Department of Computer Science and Engineering, Institute of Technology, Nirma University, Gujarat, India.
e-mail: (priyank.thakkar@nirmauni.ac.in).

Dipak M. Adhyaru is with the Department of Instrumentation and Control Engineering, Institute of Technology, Nirma University, Gujarat, India.
e-mail:(dipak.adhyaru@nirmauni.ac.in).

Tanish H. Zaveri is with the Department of Electronics and Communication Engineering, Institute of Technology, Nirma University, Gujarat, India.
e-mail:(ztanish@nirmauni.ac.in).

Archit Yajnik and Atul Negi [3] designed feature vector from Daubechies D4 wavelet coefficients. GRNN architecture and nearest neighbor classifier were used for printed Gujarati character recognition. They have used dataset of 4173 characters and achieved 97.59% and 96.71% accuracy with GRNN and nearest neighbor classifier respectively. Archit Yajnik and Dharmendra Singh [4] utilized discrete wavelet transform based feature extraction technique. Mandar Chaudhary et al. [5] used extended version of Supervised Locality Preserving Projection (ESLPP) coefficients as features and MLP with BPNN is used as a classifier for similar looking Gujarati character recognition. They have used dataset of 80 to 100 images for each character and achieved accuracy of 96%. Structural feature extraction based method for printed Gujarati characters is proposed by Mukesh Goswami and Suman Mitra [6]. They identified total 30 strokes which formulates almost all printed Gujarati character set. Proposed method was tested on dataset of 4000 printed characters. They achieved 95% accuracy for printed Gujarati characters.

Apurva A. Desai [7] has used profile based feature extraction method for Gujarati handwritten numerals recognition. An accuracy of 81.66% was achieved using multi layered feed forward neural network classifier. Chhaya Patel and Apurva Desai [8] utilized structural and statistical features with KNN classifier for handwritten Gujarati character recognition. They have achieved accuracy of 63%. Lipi Shah et al. [9] used radial histogram as feature extraction technique and Euclidean distance classifier for recognition purpose of handwritten Gujarati characters. They have used dataset of 11720 characters and achieved accuracy of 26.86%. Hetal R. Thaker and C. K. Kumbharana [10] utilized structural features like connected and disconnected components, number of end point, number of close loop for recognition of five isolated handwritten Gujarati characters. Decision tree classifier was used for classification. They have used dataset of 750 characters and achieved 88.78 % accuracy. Hybrid features based on aspect ratio, extent and zone density is used by Apurva Desai [11]. Accuracy of 86.66% is achieved with SVM classifier for handwritten Gujarati characters.

III. Gujarati Handwritten Dataset Generation

To recognize handwritten Gujarati characters, important step is collection of dataset. However, to the best of our knowledge, no such benchmark dataset is available for handwritten Gujarati characters. In this proposed work, dataset was collected from people of different age groups, different educational background and of different professions. The purpose of data collection was not disclosed to them so that they could produce samples of dataset with their natural handwriting styles. All filled handwritten forms used for database collection are scanned with HP flatbed scanner at 300 dpi resolution in color format and all the forms were saved in JPEG format. Total 88,000 isolated character images are generated from these forms, which are divided into 44 classes as shown in Fig. 1. Here, few isolated symbol together generate the complete character. For example class 3 and class 4 symbol together form the character 'Ga'. Similarly class 28 and class 29 together form the character 'La'. Total 2000 images of each symbol is generated. Dataset is collected from 2000 writers. These Gujarati dataset images are used for the training and testing purpose.

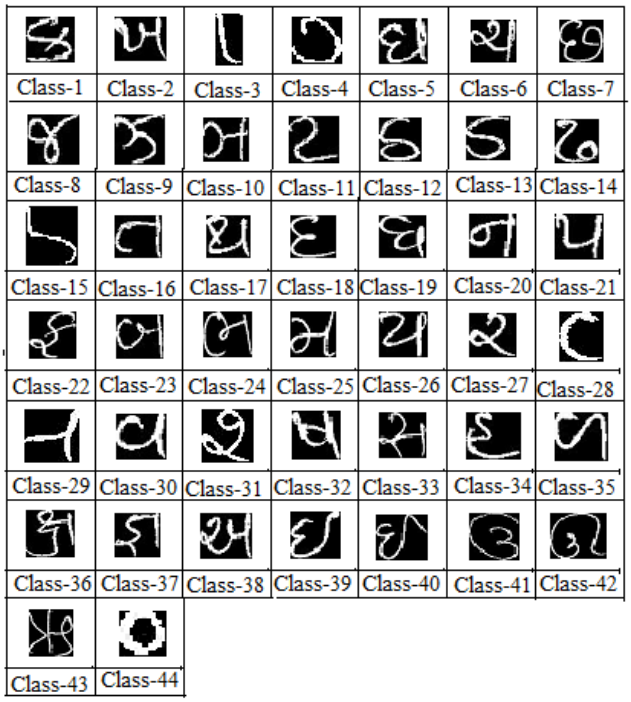

Fig.1. Handwritten Gujarati character dataset divided into 44 classes

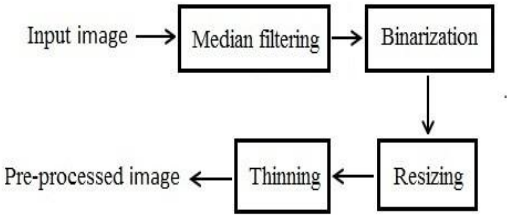

Fig. 2. Preprocessing steps for segmented character images

IV. Pre-processing

Preprocessing of the isolated handwritten character images is required before going for the feature extraction and classification. Size and style of handwritten numeral may vary from person to person, and for recognition purpose preprocessing steps are adapted in order to convert all the numeral images into uniform form. Preprocessing includes median filtering, binarization, resizing and thinning operations. Median filtering is applied in order to remove any salt and pepper noise from the image and fills holes in the object region then, these images are converted into binary images using Otsu's thresholding algorithm [23].

All the numeral images are resized to the size of 16x16 pixels with nearest neighborhood interpolation algorithm. After resizing, one pixel wide thinned image is obtained by using morphological thinning operation.

V. FEATURE EXTRACTION

Feature extraction is a crucial step for object classification. Feature extraction is a process of identifying characteristics of the given image. These characteristics are then converted into classifier acceptable format, so that classifier can be applied to classify the given input image. In this proposed work, three dissimilar types of feature extraction techniques and their fusion was implemented.

First of the three feature extraction techniques was a zone based feature extraction technique [24]. For feature vector calculation bounding box of preprocessed numeral image was divided into uniform 64 zones. The size of the feature vector representing a character with this technique was exactly equal to the 64. Each zone contributed a value to the feature vector. Number of pixels which were part of the character encompassed by a zone determined this value. The idea for zone based feature vector generation is depicted in Fig. 3.

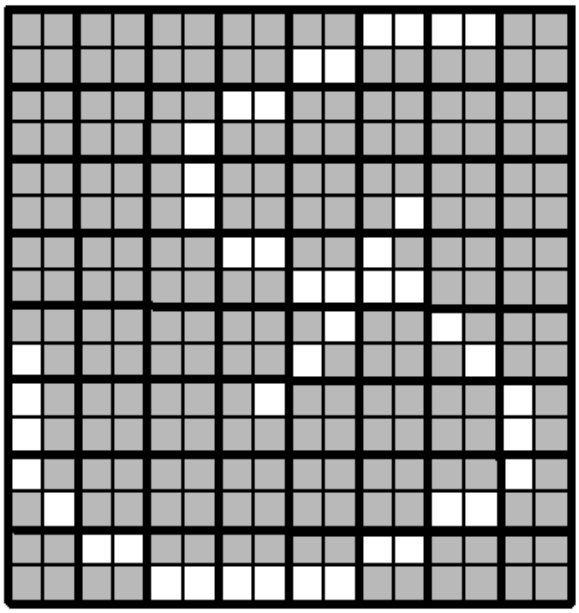

Fig. 3. Character image divided in 64 zones.

The second method proposed for the feature extraction was based on the projection profiles of the character image. Four different profiles were used for feature extraction. Horizontal, vertical, right diagonal and left diagonal projection profiles were calculated. In Fig.5, four profiles are shown for a 3x3 box. Summation of all the pixels which are representing the character in the image in different profile were calculated and used as the feature vector. Size of the feature vector was 94, where 16 elements were corresponding to the horizontal profile, 16 elements were corresponding to the vertical profile, 31 elements were corresponding to the right diagonal profile and 31 elements were corresponding to the left diagonal profile. Vectors of four profiles were combined together, which generate finally 94 elements feature vector.

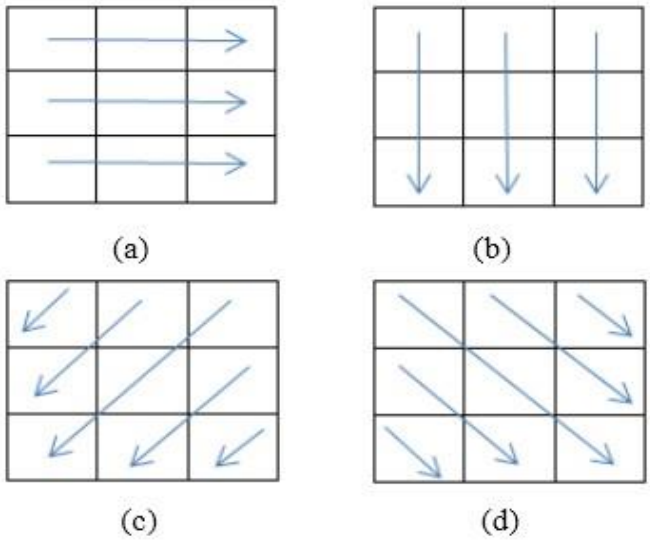

Fig. 4. (a) Represents horizontal profile. (b) Represents vertical profile. (c) and (d) represent left diagonal and right diagonal profiles respectively.

Directional chain code was considered as the third component of fusion features. Chain code was first proposed by H. Freeman. It is also referred as Freeman code or Freeman chain code [25]. Chain code represents the movements of boundary segments in terms of integer numbers. In this proposed work, a method based on horizontal scanning of a numeral image is proposed to identify the beginning of chain code sequence. The idea is depicted in Fig. 5. Rows were scanned from bottom of the image to find out the first pixel which was the part of the numeral's skeleton. Element of chain code corresponding to this first pixel was considered as the beginning of the chain code sequence. Remaining elements of chain code sequence were generated by following the skeleton in a clockwise direction from the identified beginning. The length of chain code sequence was considered as 100 since it was observed that the length of chain code for any digit was never more than 100. If the length of the chain code sequence was obtained less than 100 then remaining elements were filled with value zero. Finally, the feature vector containing 100 elements was generated. Hence, number of features extracted through zoning, projection profiles and chain code were 64, 94 and 100 respectively.

VI. PREDICTION MODELS

Naive Bayes, ANN and SVM predication models were used in this study.

A. Naive Bayes Classifier

Bayesian classifier utilizes Baye's theorem and predict the probability of a test observation belonging to a particular class. Posterior probability, $P(C|X)$, can be calculated from $P(C)$, $P(X|C)$ and $P(X)$ by using Bayes' theorem as per below mentioned Equation (1).

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)} \quad (1)$$

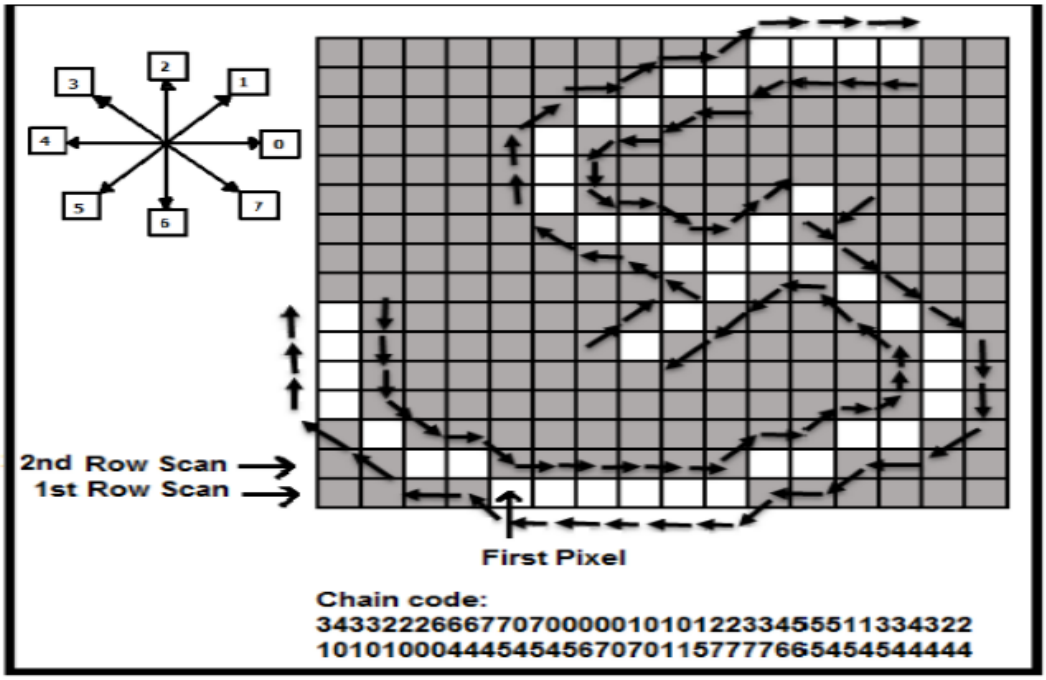

Fig. 5. Chain code obtained by finding the starting point through horizontally scanning.

The Probability of hypothesis *C* being true given that event *X* has occurred is denoted as *P*(*C*|*X*). Here, hypothesis *C* denoted a class from the set of probable classes 0,1,2,...,44 and an event *X* was a test image. Conditional probability of occurrence of event *X* given that hypothesis *C* is true is denoted as *P*(*X*|*C*) and it can be determined from the training data.

B. Support Vector Machine

SVM work on the principal of identification of the maximum margin hyper-plane as the final decision boundary to separate positive and negative classes. SVM is basically a binary classifier. SVM can be utilized for multiclass classification using one-versus-all approach. Degree (*d*) of a kernel function and regularization constant (*c*) were considered as design parameters of polynomial SVM. In case of linear SVM, only regularization constant (*c*) was considered as the design parameter. Here, 5-fold cross-validation of training set was used in order to decide degree (*d*) of a kernel function and regularization constant (*c*). Three values of *d* and five values of *c* were tried during the 5-fold cross validation of training set. These parameters and their values which were tested are summarized in Table 1. The combination of parameters that resulted into best cross-validation performance of training set was considered for designing the SVM for test set images.

TABLE I
SVM DESIGN PARAMETERS AND THEIR VALUES TESTED IN CROSS VALIDATION OF TRAINING SET

Parameters	Values
Degree of Kernel Function (*d*)	2,3,4
Regularization Parameter (*c*)	0.01, 0.1, 1, 10, 100

VII. PROPOSED APPROACH

Gujarati handwritten character recognition is implemented through prediction models learnt through individual features as well as through fusion features. The overall process adapted for handwritten Gujarati character recognition is summarized in Fig. 6. All feature sets that were implemented are shown in Table II.

VIII. EXPERIMENTAL EVALUATION

Performance of proposed model was evaluated on the basis of accuracy and f-measure. Equations (2) and (3), representing the equations used for Accuracy and F-measure calculation.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (2)$$

$$F-measure = \frac{2 \times Precision \times Recall}{Precision+Recall} \quad (3)$$

Accuracy is computed through True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN) values. Precision and Recall are required in order to compute f-measure.

Bayes classifier and Support Vector Machines prediction models were used for experimentation purpose. Linear and polynomial kernel based SVM classifier is use. Experimentations are performed using 3 individual feature sets and 4 fusion feature sets, as represented in Table II.

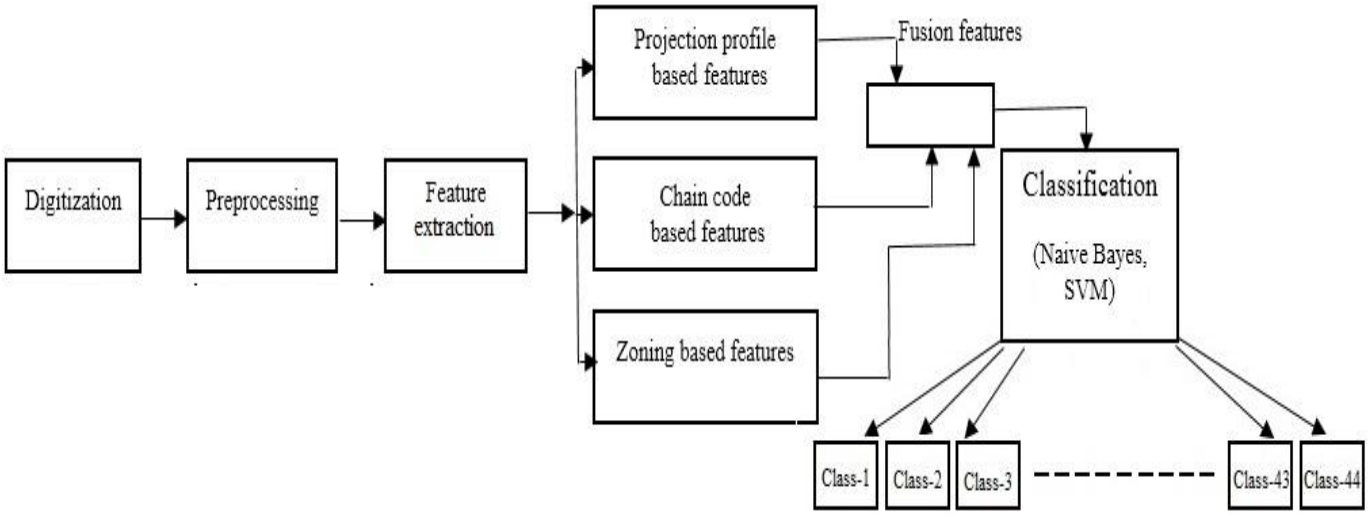

Fig. 6. Gujarati handwritten character recognition process

TABLE II
FEATURE SETS

Combinations	Name of Features	Acronym	Number of Features
Feature Set 1	Chain Code	CC	100
Feature Set 2	Projection Profile based Features	PP	94
Feature Set 3	Zone based Features (64 zones)	ZF	64
Feature Set 4 - Fusion Features 1	CC + PP	CCPP	194
Feature Set 5 - Fusion Features 2	CC + ZF	CCZF	164
Feature Set 6 - Fusion Features 3	PP + ZF	PPZF	158
Feature Set 7 - Fusion Features 4	CC + PP + ZF	CCPPZF	258

TABLE III
PERFORMANCE OF PREDICTION MODELS ON GUJARATI CHARACTER DATASET

Representation	Prediction Models								
	Linear SVM			Polynomial SVM			Naive Bayes		
	Accuracy (%)	F-measure (%)	BVDP	Accuracy (%)	F-measure (%)	BVDP	Accuracy (%)	F-measure (%)	BVDP
CC	98.98	98.98	c=0.1	99.47	99.47	c=1, degree=2	90.76	89.88	--
PP	85.25	85.23	c=0.01	91.27	91.27	c=1, degree=2	76.03	75.97	--
ZF	86.69	86.68	c=0.1	92.78	92.78	c=10, degree=2	69.82	69.90	--
CCPP	99.49	99.49	c=0.1	99.73	99.73	c=0.01, degree=3	96.80	96.75	--
CCZF	99.56	99.56	c=0.1	99.80	99.80	c=1, degree=2	96.50	96.47	--
PPZF	87.44	87.43	c=0.01	92.22	92.23	c=1, degree=2	76.64	76.70	--
CCPPZF	99.63	99.63	c=0.01	99.73	99.73	c=1, degree=2	96.43	96.43	--

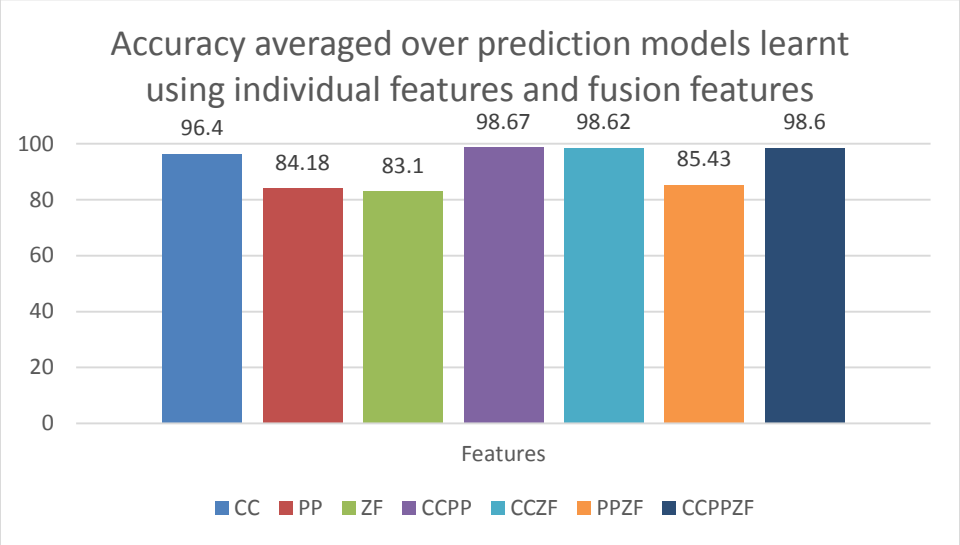

Fig. 7. Accuracy averaged over prediction models learnt using seven different features

Dataset was divided into training set and testing set, in ration of 4:1. In order to decide the Best Values of Design Parameters (BVDP) 5-fold cross validation of training set was implemented. Prediction models designed with these BVDP were learnt through the entire training set. Performance of linear SVM, polynomial SVM and naive Bayes on Gujarati character dataset is shown in Table III. For SVM, BVDPs are also shown in the table.

Results indicate that CC feature provides best accuracy in case of all prediction models compared to other individual feature sets. Highest accuracy of 99.47% was achieved using CC with polynomial SVM. Fusion of CC and ZF (CCZF features) provides highest accuracy of 99.80% with polynomial SVM. It is clear that fusion of features further enhances recognition accuracy, in case of all prediction models. Accuracy values averaged over prediction models are shown in Fig. 7. These results indicate the usefulness of CC and fusion features for character recognition.

IX. Conclusions and Future Work

In this paper, the problem of handwritten Gujarati character recognition is addressed. Chain code based, zone based and projection profile based features are used for generating feature vector. Fusion of these features is also proposed for improving the recognition accuracy. Machine learning techniques based on Bayes classifier and Support Vector Machines are utilized. Success of experiment results indicate that the proposal can be considered as the noteworthy contribution to the research. One important direction for the future work may be to extend this work by incorporating modifiers along with handwritten Gujarati characters. Increase in number of classes can make the problem difficult and fusion of some more significant features may emerge as potential solution.

Acknowledgment

The authors are thankful to Institute of Technology, Nirma University for their support to carry out this research.

References

[1] S. Antani, L. Agnihotri, “Gujarati character recognition”, in: Document Analysis and Recognition, 1999. ICDAR’99. Proceedings of the Fifth International Conference on, IEEE, pp. 418 – 421.

[2] Prof S K Shah, A Sharma, “Design and implementation of optical character recognition system to recognize gujarati script using template matching",IE(I) Journal ET, Vol 86 , (January 2006).

[3] Jignesh Dholakia, Archit Yajnik, Atul Negi, “Wavelet feature based confusion character sets for Gujarati script",International Conference on Computational Intelligence and Multimedia Applications”, IEEE,(2007), 0-7695-3050-8/07.

[4] Archit Yajnik , and Dharmendra Singh, “Feature Extraction (Image Compression) of Printed Gujarati and Amharic Letters Using Discrete Wavelet Transform”, 2010 S-JPSET, Vol. 1, Issue 1.

[5] Mandar Chaudhary,Gitam Shikkenawis, Suman K. Mitra Mukesh Goswami, “Similar looking gujarati printed character recognition using locality preserving projection and artificial neural networks",Third International Conference on Emerging Applications of Information Technology (EAIT),IEEE,(2012), 978-1-4673-1827-3/12.

[6] Mukesh Goswami and Suman Mitra, “Structural feature based classification of printed gujarati characters",Springer-Verlag Berlin Heidelberg,(2013)LNCS 8251, P. Maji et al. (Eds.): PReMI 2013, LNCS 8251, pp. 82–87, 2013.

[7] Desai, Apurva A. "Gujarati handwritten numeral optical character reorganization through neural network." Pattern recognition Vol 43.7, pp. 2582-2589, 2010.

[8] Chhaya Patel, Apurva Desai, “Gujarati handwritten character recognition using hybrid method based on binary tree-classifier and k-nearest neighbour", International Journal of Engineering Research Technology (IJERT),(volume 2 issue 6, June - 2013) ISSN: 2278-0181.

[9] Lipi Shah, Ripal Patel, Shreyal Patel, Jay Maniar, “Handwritten Character Recognition using Radial Histogram",International Journal of Research in Advent Technology,(2014) Vol.2, No.4, E-ISSN: 2321-9637.

[10] Hetal R. Thaker , C. K. Kumbharana, “Structural feature extraction to recognize some of the offine isolated handwritten Gujarati characters using decision tree classifier", International Journal of Computer Applications(0975 – 8887),(August 2014) Volume 99 – No.15.

[11] Desai, Apurva A. "Support vector machine for identification of handwritten Gujarati alphabets using hybrid feature space." CSI Transactions on ICT, pp. 1-7, 2015.

[12] J. Dholakia, A. Negi, S. R. Mohan, "Zone identification in the printed gujarati text", in: Document Analysis and Recognition, 2005. Proceedings. Eighth International Conference on, IEEE, pp. 272–276.
[13] E. Hassan, S. Chaudhury, M. Gopal, Feature combination for binary pattern classification, International Journal on Document Analysis and Recognition (IJDAR) 17 (2014), pp. 375–392.
[14] M. Maloo, K. Kale, Support vector machine based Gujarati numeral recognition, International Journal on Computer Science and Engineering 3 (2011) 2595–2600.
[15] M. Baheti, K. Kale, M. Jadhav, Comparison of classifiers for gujarati numeral recognition, International Journal of Machine Intelligence 3 (2011).
[16] M. Hanmandlu, O. R. Murthy, Fuzzy model based recognition of handwritten numerals, Pattern Recognition 40 (2007) 1840–1854.
[17] Y. Wen, Y. Lu, P. Shi, Handwritten Bangla numeral recognition system and its application to postal automation, Pattern recognition 40 (2007), pp. 99–107.
[18] U. Pal, N. Sharma, T. Wakabayashi, F. Kimura, "Handwritten numeral recognition of six popular indian scripts", in: Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on, volume 2, IEEE, pp. 749–753.
[19] U. Bhattacharya, B. B. Chaudhuri, Handwritten numeral databases of indian scripts and multistage recognition of mixed numerals, Pattern Analysis and Machine Intelligence, IEEE Transactions on 31 (2009), pp. 444–457.
[20] N. Das, J. M. Reddy, R. Sarkar, S. Basu, M. Kundu, M. Nasipuri, D. K. Basu, A statistical–topological feature combination for recognition of handwritten numerals, Applied Soft Computing 12 (2012) 2486–2495.
[21] R. K. Mohapatra, B. Majhi, S. K. Jena, Classification of handwritten odia basic character using stockwell transform, International Journal of Applied Pattern Recognition 2 (2015), pp. 235–254.
[22] S. Abirami, V. Essakiammal, R. Baskaran, Statistical features based character recognition for offline handwritten tamil document images using hmm, International Journal of Computational Vision and Robotics 5 (2015), pp. 422–440.
[23] N. Otsu, A threshold selection method from gray-level histograms, Automatica 11 (1975), pp. 23–27.
[24] D. Impedovo, G. Pirlo, Zoning methods for handwritten character recognition: A survey, Pattern Recognition 47 (2014), pp. 969–981.
[25] Freeman, Herbert. "On the encoding of arbitrary geometric configurations." IRE Transactions on Electronic Computers 2 (1961): 260-268.

Mr. Ankit Sharma received his M.E. degree from Thapar University, Patiala, India in 2010. Currently, he is working as an assistant professor in Instrumentation & Control Engineering department, Institute of Technology, Nirma University, Gujarat, India. His research interests include Machine Learning, Soft Computing, Biomedical Instrumentation and Image Processing.

Dr. Priyank Thakkar is an associate professor at Computer Engineering Department of Institute of Technology, Nirma University. He received his BE and ME degrees from South Gujarat University and Sardar Patel University in 2000 and 2008, respectively, and his PhD degree from Nirma University in 2015. He is author of almost 15 journal papers. His current research interests include data and web mining, machine learning and soft computing and image processing.

Dr. Dipak Adhyaru has done his Ph.D. from IIT Delhi, India and at present he is working as Head, Instrumentation & Control Engineering department, Institute of Technology, Nirma University, Gujarat, India. His research interest includes Soft computing applications in control and automation.

Dr. Tanish Zaveri received his B.E. degree in Electronics Engineering from Sardar Vallabhbhai Regional College of Engineering, Surat in 1998. He obtained M.Tech. in Biomedical Engineering from Indian Institute of Technology, Mumbai in 2005 and Ph.D. from SVNIT, Surat in 2010. He is presently working as a professor at Institute of Technology, Nirma University, Ahmedabad, India. His research interest includes speech, image and video processing.