International Journal of Electrical and Computer Engineering (IJECE)
Vol. 6, No. 4, August 2016, pp. 1647~1653
ISSN: 2088-8708, DOI: 10.11591/ijece.v6i4.10553

A Zone Based Approach for Classification and Recognition of Telugu Handwritten Characters

N. Shoba Rani, Sanjay Kumar Verma, Anitta Joseph
Department of Computer Science, Amrita School of Arts and Sciences, Mysuru Campus,
Amrita Vishwa Vidyapeetham, Amrita University, Karnataka, India

Article Info

Article history:

Received Mar 18, 2016
Revised May 21, 2016
Accepted Jun 6, 2016

Keyword:

K-nearest neighboring
Statistical features
Zoning

ABSTRACT

Realization of high accuracies and efficiencies in South Indian character recognition systems is one of the principle goals to be attempted time after time so as to promote the usage of optical character recognition (OCR) for South Indian languages like Telugu. The process of character recognition comprises pre-processing, segmentation, feature extraction, classification and recognition. The feature extraction stage is meant for uniquely recognizing each character image for the purpose of classifying it. The selection of a feature extraction algorithm is very critical and important for any image processing application and mostly of the times it is directly proportional to the type of the image objects that we have to identify. For optical technologies like South Indian OCR, the feature extraction technique plays a very vital role in accuracy of recognition due to the huge character sets. In this work we mainly focus on evaluating the performance of various feature extraction techniques with respect to Telugu character recognition systems and analyze its efficiencies and accuracies in recognition of Telugu character set.

Corresponding Author:

N. Shoba Rani,
Department of Computer Science,
Amrita VishwaVidyapeetham University, Mysuru Campus,
#114, 7th cross Bogadi 2nd stage, Mysuru- 570026.
Email: n.shobha1985@gmail.com

1. INTRODUCTION

The evolution of any technology relies on the reliability and efficiency of the outcomes generated by it. These factors can be constituted into optical technologies [1] only when the internal procedures defined to perform the processing are consistent with the type of input data. Especially for technologies like South Indian OCR packages it is very dominant factor that highly influences the accuracy of the system. The South Indian languages like Telugu has very wide character set of around 436 distinct characters which includes vowels, consonants, single and multi-conjunct vowel consonant clusters, however the dataset excludes the non-frequently occurring characters [2]. The identification and selection of unique features for recognition of each character of wider character set increases the chances of erroneous outcomes and that may lead to non-reliability of the OCR.

Recognition of handwritten optical character is very difficult due to different writing style of the different person. Due to large number of character and presence of half character and some confusing characters makes the recognition process even more complex. In this we take data from many users and found that writing style of every user is different. So, the recognition of the character is very difficult. In this work objective is to recognize character in Telugu by using some feature extraction techniques.

2. RELATED WORK

There are various feature extraction techniques that are used for extraction of features that are unique to a particular character. The broad categorization of technique includes directional and zone wise features. The feature extraction technique based on directional features [3] comprises the identification of features like starting point and intersection point locations, distinguish individual line segments, labeling line segment information and line type normalization. The techniques like zoning [4] involve computation of directional features with respect to every zone from which a fixed size feature vector is derived and sent as input to the classifier. The features like line segment length, line segment direction and intersection points are computed from feature vector consisting of all zones information. The other kinds of techniques used for character recognition includes the transition features [5] that are computed with respect to the locations of the images where the transition is happened from back ground to foreground pixels. Lakshmi et. al. [6] had provided some novel ideas of extracting features using K-means, with results similar to auto-encoding techniques and also employed SVM classifier. Mallikarjun Hangarge et.al [7] had proposed an algorithm using diagonal feature extraction scheme for recognizing off-line handwritten characters, every character image of size 90x 60 pixels is divided into 54 equal zones, each of size 10x10 pixels and features are extracted from each zone pixels by moving along the diagonals of its respective 10X10 pixels. Pai et.al [8], had proposed a technique for recognition using KarhunenLoeve transformation, and the topographic feature maps obtained through weight sharing in the system. Sarkar et.al [9]: In this paper, they present a system, which automatically separates the scripts of handwritten words from a document, which is written in Bangla or Devanagri mixed with Roman scripts. In this script separation technique they are extracting the text lines and words from document pages using a script independent Neighboring Component Analysis technique. Then for the script separation they have designed a Multi Layer Perception (MLP) based classifier, trained with 8 different word level holistic features. For the System evaluation they prepared two equal sized datasets, one with Bangla and Roman scripts and the other with Devanagri and Roman scripts. KandulaVenkata Reddy [10]: In this paper there two techniques for identify handwritten character those are active character detection (ACR) and contour algorithms. These two techniques can be implemented by using fuzzy logic. Pattern detection and artificial neural network and fuzzy logic. The unknown character to be tested for identification is also converted to an image and compare with standard image and there by recognized by using the fuzzy logic generators.

3. PROPOSED METHODOLOGY

The proposed methodology for feature extraction and classification is accomplished in two stages. The stage one involves the computation of Hu-Moments features, statistical features and classification of features to various classes is performed in stage two. The Block Diagram of proposed system is depicted in Figure 1.

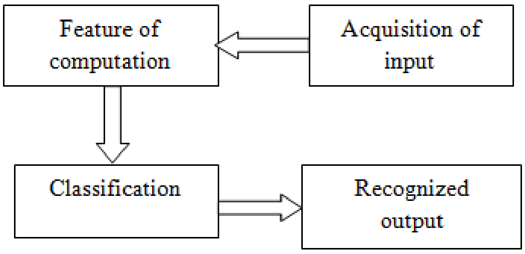

Figure 1. Block diagram of Character recognition

Initially the proposed algorithm assumes an input of the segmented characters from the document image. The present work has employed the handwritten character samples that are synthetically generated from various users. Figure 2 shows the instances of few of the character samples that are considered for experimentation.

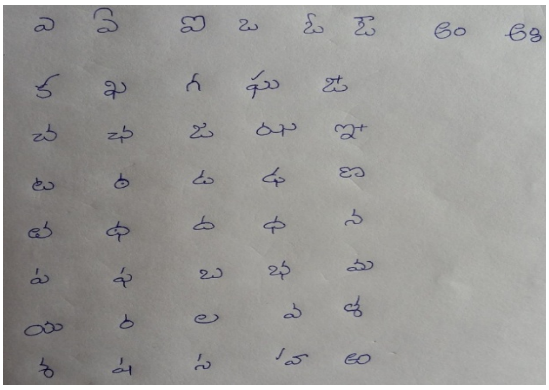

Figure 2. samples of Handwritten Characters

The feature computation and classification of characters are discussed in the subsections A and B.

3.1. Feature Computation

The feature computation is performed on the pre-processed input samples. Each input sample or character is initially divided into '9' Zones. The segmentation of character image into zones is as presented in Figure 3.

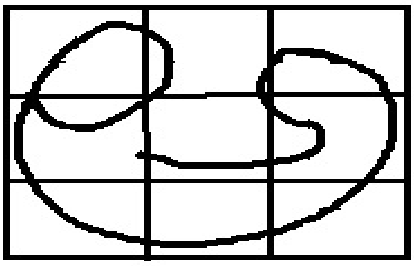

Figure 3. Division of character in Zones

Each Zone is further subjected to the process of feature computation. In the proposed work Hu-moments and statistical features are employed as features. If Z1, Z2, Z3, Z4…Z9 represents all the nine Zones, then the Hu features computed is given by equation (1).

$$Hu= \{Hu\_f_1(z_1), Hu\_f(Z_2)\ldots Hu\_f(Z_9)\} \quad (1)$$

where each Hu-feature (Zi) and i=1, 2, 3, 4...9 is further composed of seven features and given by equation (2).

$$Hu\_f(Z_9)= \{m_1Zi, m_2Zi, m_3Zi, m_4Zi \ldots m_7Zi\} \quad (2)$$

Thus we have nine zones and seven features from each zone leading to a total of 9*7 features of Hu-moments. Similarly the statistical features like centroid of each zone, entropy of each zone is given by equation (3).

$$S_{t=}h.S_{t\_}f(Z_1), S_{t\_}f(Z_2)\ldots S_{t\_}f(Z_9) \quad (3)$$

where each $S_{t\_}f(Z_i)=\{cZ_i,eZ_i\}$

Where cZ_i, eZ_irepresents the centroid and entropy features of each zone. Indicating statistical feature $S_{t.}$ The feature compution is as depicted in Figure 4. Thus for each character a total of 81 features are obtained. These features are forwarded for classification stage for the recognition of character.

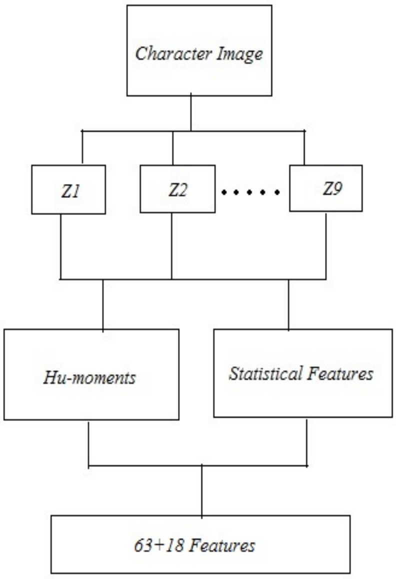

Figure 4. Feature computation

3.2. Classification and Recognition

The classification in the present work is performed using KNN and SVM Classification. If C1, C2, C3…C_n represents the classes and classifier is applied on the each feature set. The classification of features to various classes is as given in Figure 5.

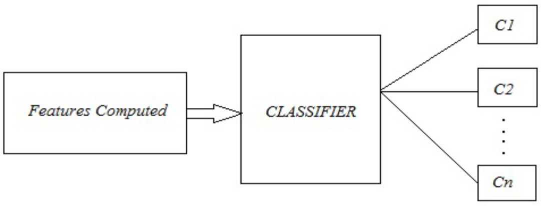

Figure 5. Classification of feature computed

The proposed work considers the Telugu handwritten vowels have the various classes. There are totally 16 vowels and which are shown in Figure 6.

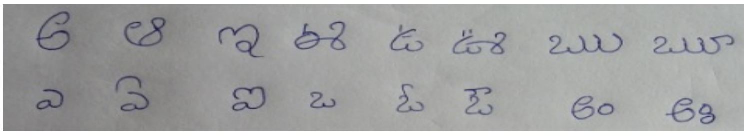

Figure 6. Vowel set

The entire dataset for classification considered is of 100 users in which 70% is used for training and 30% is used for testing. The training set is composed of 70 user reference of 10 vowels; the training matrix representation of features is as shown in Figure 7.

$$\begin{array}{cc} & \begin{array}{cccccc} f_1 & f_2 & f_3 & f_4 & \dots & f_{81} \end{array} \\ \begin{array}{c} V_1 \\ V_2 \\ V_3 \\ V_4 \\ . \\ . \\ . \\ V_{70} \end{array} & \left(\begin{array}{cccccc} i_{11} & i_{12} & i_{13} & i_{14} & \dots & i_{18} \\ i_{21} & i_{22} & i_{23} & i_{24} & \dots & i_{28} \\ i_{31} & i_{32} & i_{33} & i_{34} & \dots & i_{38} \\ . & . & . & . & & . \\ . & . & . & . & & . \\ . & . & . & . & & . \\ I_{701} & i_{702} & i_{703} & i_{704} & \dots & i_{708} \end{array}\right) \end{array}$$

Figure 7. Training matrix representation

Here V1, V2, V3… V70 represents the character references considered for training set and on the same way a test set of size 30*81 is computed for each vowel consisting of three references each. The class labels are of 16 and a matrix of dimension 70*1is created. Each row instance of training matrix. The matrix representation of labels is as shown in the Figure 8.

$$\left(\begin{array}{c} \text{label 1} \\ \text{label 1} \\ \text{label 1} \\ \text{label 1} \\ \text{label 2} \\ \text{label 2} \\ \text{label 2} \\ \text{label 2} \\ . \\ . \\ . \\ \text{label 16} \\ \text{label 16} \\ \text{label 16} \\ \text{label 16} \\ . \\ . \\ . \end{array}\right)$$

Figure 8. Class label matrix representation

The KNN classifier and SVM classifier are employed on training set for recognition of class labels. The outcome of recognition is represented in Figure 9 and Figure 10 with respect to both KNN and SVM Classifier

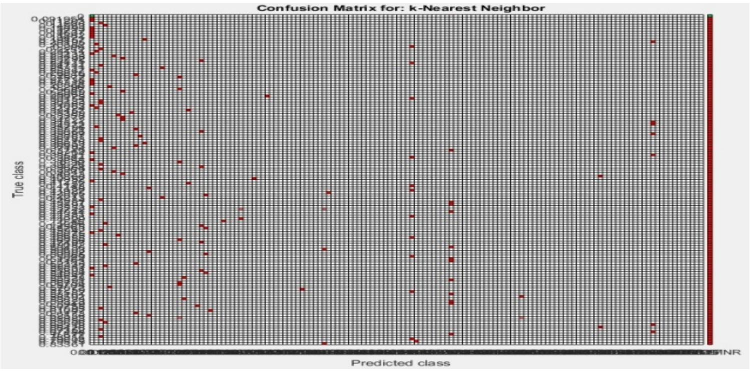

Figure 9. KNN classifier

View percentage per true class including true positive rates (TPR) and false negative rates (FNR).The accuracy (AC) is that the proportion of the full range of predictions that were correct. It’s determined using the equation (4)

$$\text{Accuracy} = \frac{a+d}{a+b+c+d} \quad (4)$$

The recall or true positive rate (TP) is that the proportion of positive cases that were properly known, as calculated using the equation (5)

$$\text{True Positive Rate} = \frac{d}{c+d} \quad (5)$$

The false negative rate (FN) is that the proportion of positives cases that were incorrectly classified as negative, as calculated using the equation (6)

$$\text{False Negative Rate} = \frac{c}{c+d} \quad (6)$$

Here matrix confusion for KNN is achieving accuracy of 76.47% and overall error is showing in red color. When we are performing same feature extraction on train data as well as testdata (new character) then it directly proportional and represented in Figure 10.

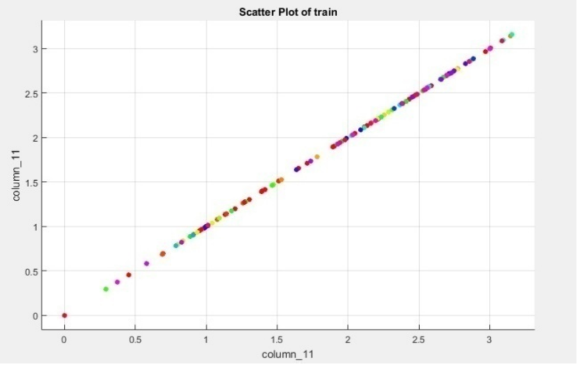

Figure 10. KNN Scatter plot of train data

SVM classifier analysis new character testset data with train data and try to find the nearest match. It is showing how two different handwritten characters are similar with nearest match is represented in Figure 11.

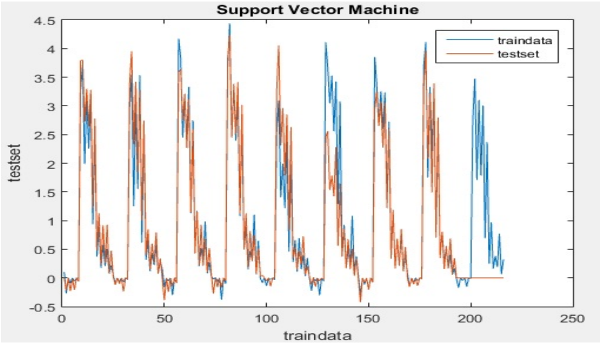

Figure 11. SVM classifier

4. EXPRIMENTAL RESULT

We collected characters from several handwritten documents of Telugu.The number of characters in the testing set is 51.All the characters square measure collected in an exceedingly systematic manner from handwritten pages. We have performed Feature Extraction techniques on each zone of character image and we computed each zone in feature vector. When we get new character image then we are trying to identify the character with respect to KNN and SVM represented in Figure 12.

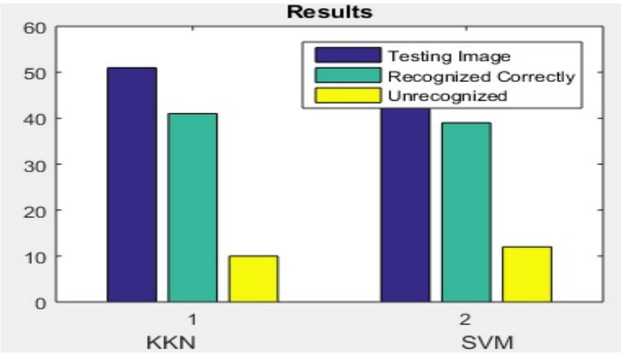

Figure 12. KNN and SVM classifier

We use here two classifier KNN & SVM and it represent different Accuracy. We have tested with 51 characters from both KNN and SVM. Accuracy is represented in Figure 12.

Classifier	Total Datasets	Total Testing Image	Recognized Correctly	Unrecognized
KNN	212	51	41	10
SVM	212	51	39	12

5. CONCLUSIONS

The proposed algorithm of feature extraction techniques for increasing accuracy and performance, particularly once it involves Telugu language. It depends upon the Algorithms that are wont to Extract and classify the character. The present Systems use varied totally techniques for Extraction and Classification that differs in accuracy. In this project we are considering scanned images as input in order to perform Zoning,. Statistical features (mean, entropy, standard division, K-nearest neighbor and Support Vector Machine classifier to get best result of character recognition. The future enhancements that we want to implement multiple Classifications.

REFERENCES

[1] A. Negi, *et al.*, "An OCR system for Telugu," *Document Analysis and Recognition, 2001. Proceedings. Sixth International Conference on.* IEEE, 2001.

[2] H. Swethalakshmi, *et al.*, "Online handwritten character recognition of Devanagari and Telugu Characters using support vector machines," *Tenth International workshop on Frontiers in handwriting recognition,* Suvisoft, 2006.

[3] M. Blumenstein, *et al.*, "A novel feature extraction technique for the recognition of segmented handwritten characters," *Document Analysis and Recognition, Proceedings. Seventh International Conference on.* IEEE, 2003.

[4] O. D. Trier, *et al.*, "Feature extraction methods for character recognition-a survey," *Pattern recognition,* vol/issue: 29(4), pp. 641-662, 1996.

[5] P. D. Gader, *et al.*, "Handwritten word recognition with character and inter-character neural networks," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on,* vol/issue: 27(1), pp. 158-164, 1997.

[6] C. V. Lakshmi and C. Patvardhan, "An optical character recognition system for printed Telugu text," *Pattern analysis and applications,* vol/issue: 7(2), pp. 190-204, 2004.

[7] M. Hangarge, *et al.*, "Statistical texture features based handwritten and printed text classification in south indian documents," *arXiv preprint arXiv: 1303.3087,* 2013.

[8] N. R. Pai and S. K. Vijaykumar, "Design and implementation of optical character recognition using template matching for multi fonts/size," *International Journal of Research in Engineering and Technology,* vol/issue: 4(2), 2015.

[9] R. Sarkar, *et al.*, "Word level script identification from Bangla and Devanagri handwritten texts mixed with Roman script," *arXiv preprint arXiv:1002.4007,* 2010.

[10] K. V. Reddy, *et al.*, "Hand Written Character Detection by Using Fuzzy Logic Techniques," *International Journal of Emerging Technology and Advanced Engineering,* vol/issue: 3(3), 2013.