DOI: http://dx.doi.org/10.26483/ijarcs.v8i7.4280

ISSN No. 0976-5697

Volume 8, No. 7, July – August 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

PERFORMANCE OF FITNESS MEASURES TO RECOGNIZE TELUGU CHARACTERS

T.R.Vijaya Lakshmi
Dept. of ECE,MGIT
Hyderabad, India

Abstract: The research in character recognition is a good old application in the area of pattern recognition and has attracted many researchers during the last few decades. There are various applications of character recognition such as in banks, post offices, defence organizations, reading aid for the blind, library automation, language processing and multimedia design. The handwritten Telugu characters which are complex in nature to recognize are considered in the current work by extracting relevant features using genetic algorithm. The performance of the system is compared by testing the algorithm, with two different fitness measures.

Keywords: Genetic algorithm; cross correlation coefficient; Hamming distance; handwritten Telugu characters; pixel distribution

I. INTRODUCTION

Handwritten character recognition involves three basic steps namely development of database, feature extraction and classification of characters. One of the major aspects in HCR is digitization of documents to develop the database. There are several devices available for digitization of documents such as scanners, cameras, mobile-cameras, etc.

India is a diverse nation and is rich in literature. As of today there are 33 languages and 2000 dialects, of which 22 are recognized under the constitution. The most popular South Indian languages are Telugu, Tamil, Kannada, Malayalam, Tulu, etc., [1, 2]. The alphabets of these languages have large number of basic and compound characters.

The primary challenge of any character recognition system is to design a framework to handle the text layout in the image, character fonts, sizes and variability in imaging conditions with uneven lighting, reflection and shadowing. Even though all these problems can be taken care using sophisticated equipment, there are many more issues. The following are some of the issues in handwritten character recognition:

i. Several distortions are introduced while scanning the documents [3].
ii. For old and historical (palm-leaf) documents, which are poor in quality, these distortions are inevitable.
iii. Recognition of handwritten characters is a tedious task as the style of writing varies from individual to individual.
iv. Variation in font style, size, orientation, alignment and complex background makes the character recognition phase a challenging task.

There are 18 vowels and 36 consonants in Telugu language. The 18 vowels and 36 consonants are shown in Fig. 1(a) and (b) respectively. A few similar groups of characters are shown in Fig. 1(c). This clearly indicates that recognizing such similar groups of characters is very difficult. Moreover there are no standard databases available for Indian languages and hence it becomes very difficult for the development of handwritten character database [4].

This paper describes the process of recognizing the basic characters written on paper documents. Section II describes existing methods for recognizing handwritten characters. The procedure used to recognize the handwritten characters is explained in section III. The experimental results are discussed in section IV. The paper is finally concluded in section V.

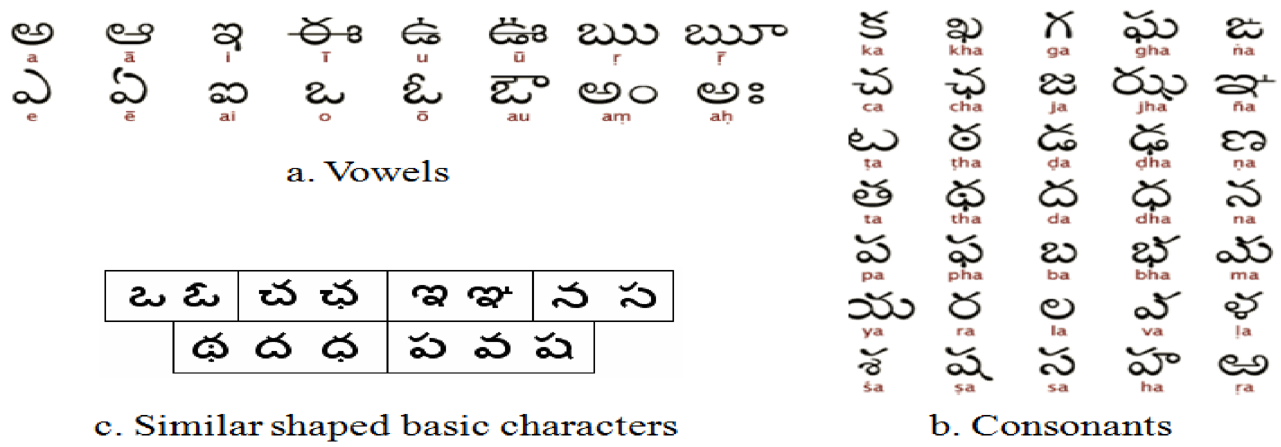

Fig. 1: Telugu basic characters a) Vowels b) Consonants c) Similar shaped characters

II. RELATED WORK

Nan-Xi Li and Lian-WenJin [5] contributed work on handwritten Chinese character recognition. The Modified Quadratic Discriminative Function (MQDF) classifier was trained with several Chinese databases containing isolated characters, digits and punctuation marks using Linear Discriminant Analysis (LDA) based strategy. They tested 353 classes of Chinese characters and reported character recognition accuracy in the range 77.36% to 91.89%.

Liang Xu et al. [6] presented over-segmentation approach for Chinese handwritten touching characters by designing a supervised learning filter. They conducted tests on 36,727 Chinese characters and extracted nine different geometric features from them. They reported the recall and precision rates of 74.8% and 69.6% respectively using linear Support Vector Machine (SVM).

Yunxue Shao et al. [7] have worked on handwritten Chinese character image restoration. The gradient features were extracted from the characters in eight different directions. The candidate classes' derived using Euclidean distance was reordered by MQDF classifier. On collection of 3,755 Chinese character dataset from 300 individuals, they conducted tests to categorize characters written by 60 individuals while those of remaining 240 were used as training. By varying the size of the Eigen vectors in MQDF classifier, they reported classification rates from 88.48% to 93.5%.

Impedovo and Pirlo [8] reviewed various types of zoning methods to recognize handwritten characters. Broadly zoning topologies are categorized into static and adaptive. The static topologies like slice-based, hierarchical based, uniform and non-uniform strategies were discussed in [8]. Adaptive topologies discussed by them were discrimination-based, perception-oriented, template-based and Voronoi-based. Combination of zoning topologies and parameter-based membership functions were used to recognize handwritten numerals from CEDAR database. In [9] the authors reported a recognition rate of 92% for CEDAR dataset, using neural network classifier and fuzzy zoning technique.

Radtke et al. [10] worked on NIST handwritten digit dataset (Latin numerals), consisting of 50,000 training and 10,000 testing patterns. Set of features extracted from these patterns were contour-based information, concavities and pixel distribution. The best recognition rate reported was 95% using nearest neighbor as a classifier.

III. METHODOLOGY

The handwritten characters written on A4 size paper documents are scanned at 300 dpi. In the current study the number of classes or characters for which the database developed is 50. Each character is written on a paper in a rectangular box in different sizes and styles by 85 individual writers. These scanned documents are preprocessed to extract the characters from the documents to conduct experiments. The document images are binarized first using Otsu's algorithm. The noise introduced due to scanning is removed using morphological tools such as dilation and erosion. The characters are then extracted by applying minimum boundary rectangle concept. Therefore a total number of 3,750 samples (50 characters x 75 samples) were developed for training whereas 500 samples (50 characters x 10 samples) were developed for testing purpose. The size of all the character images is then normalized to MxM (M=50 in the current work) without changing the aspect ratio of the character.

A. Feature Extraction

The motivation behind feature extraction is to estimate the attributes that are most appropriate to represent a character [11]. Its primary objective is to maximize recognition rate with minimum elements. The normalized raw character images are then used to extract the useful features. Due to the nature of penmanship (style of handwriting) with its high level of variability and imprecision, extracting such elements is a troublesome assignment [12]. In this work, the features are extracted from the character images by dividing them into smaller zones of size 10×10. Hence each character contains a total of 25 zones for a normalized image of size 50×50. From each zone the 4- and 8-directional pixel distributions are computed by superimposing 3×3 masks. The 4 and 8 directions considered are shown in Fig. 2 (a) and (b) respectively. The number of features extracted for each character image is 100 and 200 along 4- and 8-directions respectively.

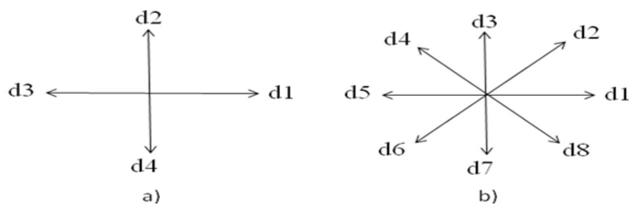

Fig.2: Pixel distribution along (a) 4 directions (b) 8 directions

The most challenging task is to identify the relevant features that help to distinguish any pair of characters [13, 14, 15]. The relevant or optimum features are then found using one of the fast search methods namely genetic algorithm. The relevant features extracted from the search mechanism are used for classifying the handwritten Telugu characters using k-NN classifier.

B. Genetic algotirhm

Genetic algorithm (GA) is a search method to find useful solutions over generations. Initially a population H of size D (desired number of features) is generated randomly. Each solution is represented as a binary string. These random solutions are evaluated with the two proposed fitness measures. In 'g' generations, the entire population is evaluated. The best fit solutions are selected as parents for the next generation. These are used to reproduce the new population by performing crossover and mutation steps. The primary steps involved in GA to generate a new solution are described as follows:

Selection: Based on the fitness measure computed the fit solutions survive in the selection step. The unfit solutions make space for the new solutions in the next iteration/generation.

Crossover: The new solution is produced by crossing two fit solutions. The bit-wise logical AND is performed between the solutions to generate the new one.

Mutation: One of the bits in the binary string obtained from the crossover step is muted to produce a new offspring.

The algorithm is terminated if maximum number of generations is reached. In every generation the worst solution is discarded to provide room for the offspring produced in the next generation. The newly generated population is again evaluated using the fitness function. The best solution containing subset of features from this algorithm is used to classify the handwritten Telugu characters using k-NN classifier.

The steps involved in genetic algorithm are as follows:

1. Initialize the population with random solutions.
2. Evaluate the population using fitness function.
3. Terminate if maximum number of generations reached else go to step 4.
4. Reproduce new population by crossover and mutation.
5. Go to step 2.

C. Proposed fitness measures

The solutions in the population survive in the next generation based on the fitness computed. Generally genetic algorithm is suited for maximization problem however minimization problems can also be solved by performing suitable transformation. The control parameters of genetic algorithm set in the proposed work are tabulated in Table 1.

Table 1: Control parameters of Genetic algorithm

Parameter	**Value**
Population size	40
Crossover rate	0.5
Mutation rate	0.004
No. of generations	50

The fitness measures presented in the current work to evaluate the population in every generation are Hamming distance and Cross correlation coefficient. These are described as follows:

Hamming distance: The number of positions at which the bits in two binary strings differ is the hamming distance. This distance is computed by taking XOR of the two strings/solutions of genetic algorithm x_i and y_i and is depicted in Equation (1).

$$HD(x) = \sum_{i=1}^{n} x_i \oplus y_i \quad (1)$$

The minimum the hamming distance the better is the solution in the whole population. For maximization problem of genetic algorithm the function *HD(x)* is modified as

$$HD'(x) = \frac{1}{1 + HD(x)} \quad (2)$$

Cross correlation coefficient: The number of positions at which the bits in two strings are same is the degree of similarity between them. It is given by

$$r(x) = \sum_{i=1}^{n} x_i \wedge y_i \quad (3)$$

The cross correlation coefficient is given by

$$C(x, y) = \frac{r(x).r(y)}{n_x n_y} \quad (4)$$

Where n_x and n_y are the number of cells occupied by the solutions x and y respectively. The larger the value in the entire population is preferred (maximization problem).

IV. RESULTS AND DISCUSSIONS

The Genetic algorithm is started with an initial population, whose members are uniformly distributed in the range $[L_b, U_b]$. The lower bound (L_b) is set to 1 and the upper bound (U_b) is set to n_f (total number of features). For the first generation, the features are selected randomly. The positions of the features selected are set to logic 1 and the features that are not selected are set to logic 0. From second generation onwards the solutions in the population are modified based on the steps involved in the technique as discussed in the previous section.

As the absolute subset size can't be anticipated, the desired number of features/dimension size (D) is allowed to vary in the simulations to identify the optimum subset size. The maximum subset size set as $n_f/2$, (half the size of the original feature vector) in the simulations to achieve at least 50% optimization.

The k-NN classifier is trained only with the optimum features derived from the optimization technique to classify the handwritten Telugu characters. The optimization results obtained for 4- and 8-directional pixel distributed feature sets are shown in Figs.3 and 4, respectively.

The best results obtained with the two fitness measures are tabulated in Table 2. The optimum number of features needed to better recognize the handwritten Telugu characters are also tabulated.

The results obtained with the cross correlation fitness measure are better compared to hamming distance measure. Hence the memory needed to save the features is reduced with the search algorithm GA and proposed fitness measures to recognize the handwritten Telugu characters.

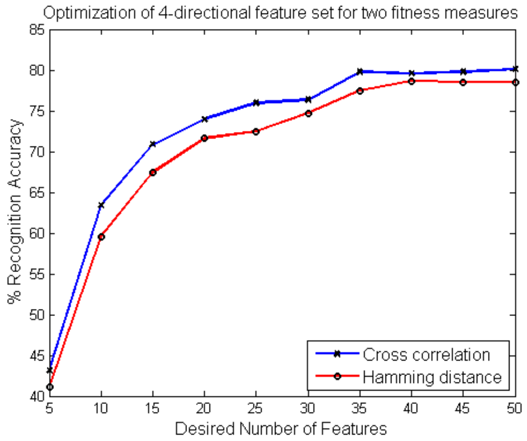

Fig.3: Average recognition accuracies obtained across 30 runs for 4-directional feature set

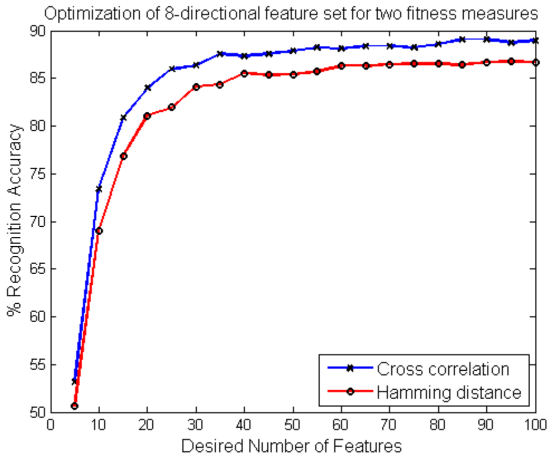

Fig.4: Average recognition accuracies obtained across 30 runs for 8-directional feature set

Table 2: Character recognition rates for various fitness measures using GA

Feature extraction method	**Fitness measure**	**Recognition Accuracy**	**No. of optimum features**
4-directional features	Cross-correlation	80.1%	50
	Hamming distance	78.63%	40
8-directional features	Cross-correlation	**89.08%**	**85**
	Hamming distance	86.7%	90

V. Conclusion

In this work, investigations were carried out to recognize handwritten Telugu characters by extracting 4- and 8-directional features. With two fitness measures namely cross-correlation and hamming distance were used to evaluate the population in the Genetic Algorithm. The optimum solution (subset of features) obtained from GA is used to recognize the handwritten Telugu characters. The best recognition rate obtained is 89.08% with an optimum subset of 85 features. In future the work can be extended by employing other feature extraction and search methods.

VI. References

[1] U. Bhattacharya and B. B. Chaudhuri, “Handwritten Numeral Databases of Indian Scripts and Multistage Recognition of Mixed Numerals,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 3, pp. 444–457, 2009.

[2] S. Bag, G. Harit, and P. Bhowmick, “Recognition of Bangla compound characters using structural decomposition,” Pattern Recognition, vol. 47, no. 3, pp. 1187–1201, 2014.

[3] D. Fernandez-Mota, J. Llados, and A. Fornes, “A graph-based approach for segmenting touching lines in historical handwritten documents,” International Journal on Document Analysis and Recognition (IJDAR), vol. 17, no. 3, pp. 293–312, 2014.

[4] U. Pal, R. Jayadevan, and N. Sharma, “Handwriting recognition in Indian regional scripts: A survey of offline techniques,” ACM Transactions on Asian Language Information Processing (TALIP), vol. 11, no. 1, pp. 1–19, 2012.

[5] N.-X. Li and L.-W. Jin, “A Bayesian-based method of unconstrained handwritten offline Chinese text line recognition,” International Journal on Document Analysis and Recognition (IJDAR), vol. 16, no. 1, pp. 17–31, 2013.

[6] L. Xu, F. Yin, Q.-F. Wang, and C.-L. Liu, “An over-segmentation method for single-touching Chinese handwriting with learning-based filtering,” International Journal on Document Analysis and Recognition (IJDAR), vol. 17, no. 1, pp. 91–104, 2014.

[7] Y. Shao, C. Wang, and B. Xiao, “A character image restoration method for unconstrained handwritten Chinese character recognition,” International Journal on Document Analysis and Recognition (IJDAR), vol. 18, no. 1, pp. 73–86, 2015.

[8] D. Impedovo and G. Pirlo, “Zoning methods for handwritten character recognition: A survey,” Pattern Recognition, vol. 47, no. 3, pp. 969–981, 2014.

[9] S. Impedovo and G. Pirlo, “Tuning between exponential functions and zones for membership functions selection in voronoi-based zoning for handwritten character recognition,” International Conference on Document Analysis and Recognition (ICDAR), pp. 997–1001, 2011.

[10] Paulo V. Radtke, Luiz S. Oliveira, R. Sabourin, and T. Wong, “Intelligent zoning design using multi-objective evolutionary algorithms,” International conference on Document Analysis and Recognition, pp. 824, 2003.

[11] R. Jayadevan, S. R. Kolhe, P. M. Patil, and U. Pal, “Offline recognition of Devanagari script: A survey,” IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, vol. 41, no. 6, pp. 782–796, 2011.

[12] M. Abaynarh, H. El Fadili, and L. Zenkouar, “Enhanced feature extraction of handwritten characters and recognition using artificial neural networks,” Journal of Theoretical and Applied Information Technology, vol. 72, no. 3, pp. 355–365, 2015.

[13] P.A. Estevez, M. Tesmer, C.A. Perez, and J.M. Zurada, “Normalized mutual information feature selection,” IEEE Transactions on Neural Networks, vol. 20, no. 2, pp. 189–201, Feb 2009.

[14] A. Al-Ani and M. Deriche, “Feature selection using a mutual information based measure,” in Proceedings International Conference on Pattern Recognition, vol. 4, pp. 82–85, 2002.

[15] H. Peng, F. Long, and C. Ding, “Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 8, pp. 1226–1238, Aug 2005.