International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 7, Issue 8, August 2017)

Recognition of Telugu Characters using Correlation Concept

T. R. Vijaya Lakshmi

Dept. of ECE, MGIT, Gandipet, Hyderabad, India

***Abstract*— This paper deals with the handwritten character recognition for Telugu script. The similar characters of Telugu are classified into subgroups based on the similarity between the characters. The similarity between the characters is measured by computing the correlation coefficient between them. The recognition results obtained with each group of characters are reported in this paper.**

***Keywords*— Correlation coefficient, Handwritten Telugu characters, grouping of characters, erosion and dilation.**

I. Introduction

Irregular handwriting aggravates ambiguities and makes it harder to group symbols and to distinguish relations among them. A cause of this is due to inexperienced users, because they normally take excessive freedom with the location and alignment of handwritten symbols. Other kinds of irregular writing arise during the correction, deletion, and insertion of symbols.

The handwritten recognition systems are categorized into Online and Offline. The dynamic characteristics of the writings are captured in case of Online recognition system, whereas the documents are generally scanned in case of Offline recognition system. In 1950s, with the innovated data acquisition devices such as electronic tablets, researchers were motivated to work on recognizing On-line handwritten characters. The X-Y coordinates of an Online character were captured by this device. This captured data was further used to recognize the handwritten characters. But Offline handwritten CR system is still a challenging task for researchers.

Later in 1990s, with more powerful electronic gadgets such as cameras, scanners, tablets and with advanced artificial intelligence techniques, the researchers were motivated to contribute work on recognizing Offline handwritten characters [1].

Off-line handwriting recognition is the task of recognizing the image of a handwritten text, in contrast to on-line recognition where the dynamic characteristics of the writing are available as well (as is the case with the pen input of handheld devices). They do not carry temporal or dynamic information such as the order of pen-on and pen-off movements.

II. Related Work

Angadi et al. [2] presented zonal statistical approach for identifying images on display boards. They carried out the work on Kannada characters in scene images from display boards. The image was divided into zones and summation of the pixel intensities in the zone was considered as a feature. The number of samples considered for experiments was 1043. For character classification nearest neighbor classifier was used.

Deepak Kumar et al. [3] worked on recognizing camera captured images using the trial version of Nuance Omni page OCR. The captured images were binarized by Mid-line Analysis and Propagation of Segmentation (MAPS). They reported the benchmark recognition accuracies for five standard datasets.

Comparison of three different feature vectors using Gabor, structural and 2D Discrete Cosine Transform (DCT) on isolated Gurumukhi and English words were reported in [4] using different classifiers like k-Nearest Neighbor, Parzen Probabilistic Neural Network and Support Vector Machines. To extract Gabor features the word image was divided into zones and further each zone was divided into sub zones. To extract 2D DCT features the word image was split into two parts and then 2D DCT was applied for the two parts separately. They concluded that Gabor features gave better results for Gurumukhi and English words compared to structural and 2D DCT features.

Kavita Bhardwaj et al. [5] explored to identify various Indian document scripts by employing Empirical Mode Decomposition (EMD). For all the script images Radon transform was used to find the orientation angles. Further they were decomposed into a finite set using empirical mode decomposition technique. The maximum energy computed from this technique helps to select an unique feature vector of various individual scripts for identification. The experiments were conducted on 300 documents for each script.

Pavan Kumar et al. [6] explored the performance evaluation of Indic scripts. Analysis of accuracy and error rates were evaluated using edit distance measure proposed by them. They concluded that the accuracy and error rates were uncorrelated except for few special cases [6].

Siva Reddy et al. worked on recognizing On-line Assamese numerals [7]. The dataset was developed using Hidden Markov Model (HMM). For feature extraction, the first and second derivatives were used. The Assamese numerals 5 and 6 are similar and the authors focused on improving the performance of these numerals using a novel distance measure.

Pradeepta and Ravula Kollu [8] presented dimensionality reduction approach using row-wise decimal conversion to recognize handwritten Odia numerals. Experimented on 1500-isolated Odia numerals using recurrent neural network and reported recognition rate of 92.4%.

Aradhya et al. [9] worked on Telugu and Kannada handwritten numerals. The structural features, independent of the size of the symbol, such as largest profile distance and estimation of directional density were extracted from the handwritten numeral images. Employing probabilistic neural network they classified the 10-class problem. The average classification rates reported for Telugu and Kannada were 99.6% and 99.4%, respectively.

The weaknesses found in all the existing methods are 1)Ignoring the semantic information 2)Many assumptions and parameters of the algorithms are set at the initial phase. Hence, even for minor changes out of the lab environment the assumptions are not valid [1]. Hence no complete recognition system was found for handwritten text.

III. METHODOLOGY

It is evident from the literature survey [10] that no standard dataset of Indian languages is readily available for the research activity. Hence, there is a need to develop the dataset in the laboratory environment for any Indian language [10]. Therefore in the present work the first stage of research is to develop and build a handwritten Telugu character dataset.

Due to lack of standard data set to conduct experiments on handwritten Telugu characters [10-17], the data is collected from various scribers from different age groups in the laboratory environment. The characters written on high quality papers in an isolated manner, from 360 individuals are collected to develop the handwritten Telugu character set. The number of basic handwritten Telugu characters considered in this work is 50, this account to 18,000 samples in total (50 × 360). All the documents collected from various scribers are scanned at 300 dpi and stored as images. The sample scanned document collected from a scriber containing 50 different classes is shown in Fig. 1(a).

In the next step, pre-processing operations are performed to extract characters. The documents are binarized first. The distortions introduced during scanning are filtered from the binarized document images using morphological tools such as erosion and dilation. Using 8-connectivity neighbourhood all the connected objects having pixels less than or equal to ‘c’ (empirical value of c is set to 30 by trial and error method) are eliminated, to remove noise from the binarized document images. The effect after noise removal is shown in Fig. 1(b).

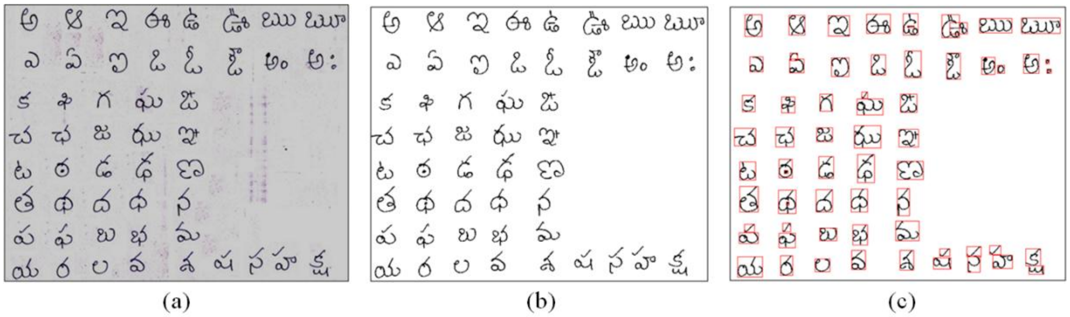

Fig. 1 Preprocessing result a) Original document b) After noise removal c) Character extraction

The connected objects are labeled after removing noise from the documents. For every labeled object in the document, the regional properties are measured to extract the characters using the minimum boundary rectangle method as shown in Fig. 1(c).

The characters are then grouped based on the similarity measure. This is done by finding the correlation coefficient between the characters. The correlation coefficient (r) between any two character images X and Y is given by

$$r = \frac{\sum_m \sum_n (X_{mn} - \bar{X})(Y_{mn} - \bar{Y})}{\sqrt{\left(\sum_m \sum_n (X_{mn} - \bar{X})^2\right)\left(\sum_m \sum_n (Y_{mn} - \bar{Y})^2\right)}} \quad (1)$$

where $\overline{\mathrm{X}}$ and $\overline{\mathrm{Y}}$ are their mean values.

If the correlation coefficient between characters is more than 0.75 then they are grouped into one category. Further the correlation coefficient is used to classify the characters among them. The maximum the value of r, the more the characters are similar. The classification results are discussed in the next section.

IV. RESULTS AND DISCUSSIONS

As discussed in the methodology section, the number of character samples considered in the current work is 18000 written by 360 scribers. These characters are divided into 10 folds. Each fold contains character samples written by 36 writers. Each fold of character samples are tested by training the remaining 9 folds of character samples. Further these character samples are divided into groups based on their correlation coefficient.

The complexity lying with the recognition of Telugu characters is carried out in two steps in the current work. In the first step of the recognition model, the correlation matrix is generated by measuring the similarity between each test character and all the training characters. If the correlation coefficient is more than 0.75 then they are grouped to a subgroup of characters.

The characters grouped into 6 subgroups with the proposed methodology are shown in Fig. 2. It is observed from Fig. 2 that the similarity among the characters in a subgroup is very high. It is a very challenging task to classify such similar characters.

In the second step the characters within the group are classified based on the same similarity measure (correlation coefficient). The classification results (group-wise) obtained with this approach are tabulated in Table 1. The bar graph comparison of the results is shown in Fig. 3.

Table 1:
Group-wise recognition results obtained using correlation coefficient

Group number	No. of characters in the group	Recognition Accuracy (%)
1	9	84.43
2	14	83.03
3	2	98.75
4	8	78.28
5	9	90.4
6	5	97.5
Average		88.71

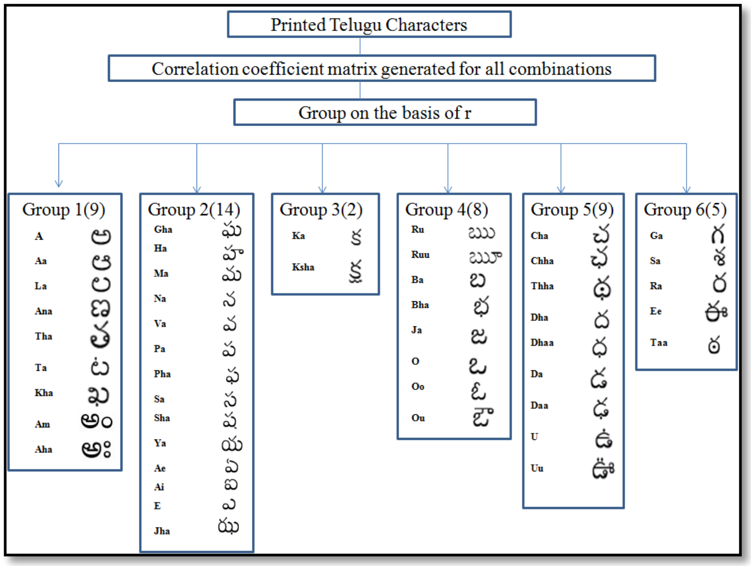

Fig. 2 Grouping of characters based on similarity measure

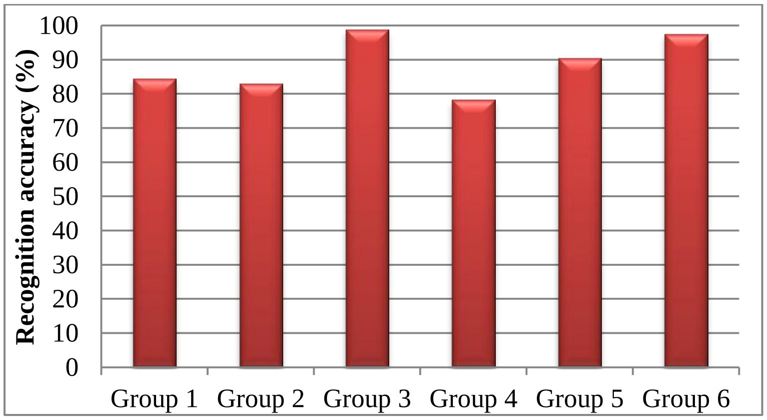

Fig. 3 Group-wise recognition results

V. CONCLUSION

In this work the confusing characters of Telugu are categorized into six groups by computing correlation coefficient between the characters. The similarity is very high among the Telugu characters. This problem is addressed by dividing them into groups. Further the computed correlation coefficient is utilized to recognize the characters within a group. The group-wise recognition rates obtained are reported in this paper. The average recognition rate obtained is 88.71%. In future these characters can be recognized by extracting the most relevant features and with better classifiers.

REFERENCES

[1] Arica, Nafiz and Yarman-Vural, Fatos T, “An overview of character recognition focused on off-line handwriting,” IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, vol. 31, 2001, pp.216-233.

[2] Angadi, S.A and Kodabagi, M.M. and Jerabandi, M.V., “Character recognition of Kannada text in low resolution display board images using zone wise statistical features,” World Congress on Information and Communication Technologies (WICT), 2012, pp. 61-66.

[3] Kumar, Deepak and Ramakrishnan, A. G., “Recognition of Kannada Characters Extracted from Scene Images,” Proceeding of the Workshop on Document Analysis and Recognition, Mumbai, India, 2012, pp.15-21.

[4] Rani, Rajneesh and Dhir, Renu and Lehal, Gurpreet Singh, “Performance Analysis of Feature Extractors and Classifiers for Script Recognition of English and Gurmukhi Words,” Proceeding of the Workshop on Document Analysis and Recognition, Mumbai, India, 2012, pp.30-36.

[5] Bhardwaj, Kavita and Chaudhury, Santanu and Roy, Sumantra Dutta, “An Empirical Intrinsic Mode Based Characterization of Indian Scripts,” Proceeding of the Workshop on Document Analysis and Recognition, Mumbai, India, 2012, pp.120-123.

[6] Kumar, P. Pavan and Bhagvati, Chakravarthy and Agarwal, Arun, “On Performance Analysis of End-to-end OCR Systems of Indic Scripts,” Proceeding of the Workshop on Document Analysis and Recognition, Mumbai, India, 2012, pp.132-138.

[7] Reddy, G. Siva and Sarma, Bandita and Naik, R. Krishna and Prasanna, S. R. M. and Mahanta, Chitralekha, “Assamese Online Handwritten Digit Recognition System Using Hidden Markov Models,” Proceeding of the Workshop on Document Analysis and Recognition, Mumbai, India, 2012, pp.108-113.

[8] Sarangi P.K. and Ravulakollu K.K., “Feature extraction and dimensionality reduction in pattern recognition using handwritten Odia numerals,” Journal of Theoretical and Applied Information Technology, vol.65, 2014, pp.770-775.

[9] Manjunath Aradhya, VN and Hemantha Kumar, G and Noushath, S, “Multilingual OCR system for South Indian scripts and English documents: An approach based on Fourier transform and principal component analysis,” Engineering Applications of Artificial Intelligence, vol.21, n0.4, 2008, pp.658-668.

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

[11] T.R.Vijaya Lakshmi.; Sastry, P.N.; Krishnan, R.; Rao, N.V.K.; Rajinikanth, T.V., “Analysis of Telugu Palm Leaf Character Recognition Using 3D Feature,” International Conference on Computational Intelligence and Networks (CINE), pp.36,41, Jan. 2015.

[12] P. N. Sastry, T.R.Vijaya Lakshmi, R. Krishnan, and N. Rao, “Analysis of Telugu palm leaf characters using multi-level recognition approach,” J. Applied Engn. Sci, vol. 10, no. 20, pp. 9258–9264, 2015.

[13] P. N. Sastry, T.R. Vijaya Lakshmi, N. V. K. Rao, T.V. Rajinikanth and A. Wahab, ”Telugu Handwritten Character Recognition Using Zoning Features,” International Conference on IT Convergence and Security (ICITCS), Beijing, 2014, pp. 1-4.

[14] T.R. Vijaya Lakshmi, P.N. Sastry, and T.V. Rajinikanth, “Hybrid approach for Telugu handwritten character recognition using k-NN and SVM classifiers,” Inter- national Review on Computers and Software, vol. 10, no. 9, pp. 923–929,2015.

[15] T.R.Vijaya Lakshmi, P.Narahari Sastry, T.V.Rajinikanth, “Feature optimization to recognize Telugu handwritten characters by implementing DE and PSO techniques,” International conference on Frontiers in Intelligent Computing Theory and Applications, 2016, pp. 397-405.

[16] T.R.Vijaya Lakshmi, P.Narahari Sastry, T.V.Rajinikanth, “A novel 3D approach to recognize Telugu palm leaf text,” International Journal on Engg. Science and Technology, Vol.20, No.1, 2017, pp. 143-150.

[17] P.Narahari Sastry, T.R.Vijaya Lakshmi, R.K.Krishnan, N.V.Koteswara Rao, “Modeling of palm leaf character recognition using transform based techniques,” Pattern Recognition Letters, Vol.84, 2016, pp. 29-34.