Research Article

IJCRR
Section: Technology
Sci. Journal
Impact Factor
4.016
ICV: 71.54

Multi-stage Strategy to Classify Handwritten Characters of Telugu

T. R. Vijaya Lakshmi

Department of Electronics and Communications Engineering, MGIT, Gandipet, Hyderabad, India.

ABSTRACT

The aim of this work is to recognize handwritten characters of Indian language, Telugu. Single stage of classifying similar Telugu characters leads to low recognition rate. However similar characters of Telugu (Indian language) are recognized in two stages in the current work. Various preprocessing steps are carried out first to extract characters from the handwritten documents. The preprocessed characters are then utilized to extract features from them. These features are further used in the proposed two-stage classification. The misclassified characters from the first stage of classification are fed to the second classifier in the proposed method. The recognition rates obtained with the two stage system are better compared to the single stage classification system.

Key Words: Handwritten characters, Two-stage strategy, Instance-based classifier, Support vector machine, Histogram profile

INTRODUCTION

Exhaustive work has been contributed on printed text and relatively very less amount of research has been reported on handwritten text [1,2,3,4,5,6,7,8,9,10]. A comprehensive survey on handwritten character recognition were reported in [11,12,13,14]. Relatively very less amount of research was found on South Indian languages like Tamil, Kannada and Telugu etc. [14,15,16,17,18,19,20,21].

There are several benchmark datasets available for Latin numerals such as MNIST, CEDAR, NIST and CENPARMI [22]. The standard dataset available for English alphabets is UNIPEN. A few Chinese standard/benchmark handwritten databases are as follows:

- HCL 2000 for Chinese alphabetical characters.
- ETL8B and ETL9B datasets comprises 956 and 3036 character classes, respectively.
- SCUT-IRAC for Chinese numerals.
- CASIA-HWDB 1.1 for Chinese alphabets, numerals and punctuation marks.

A few Devanagari (Indian script) standard small datasets available are V2DMCHAR and ISIDCHAR. As such no standard database available for other Indian scripts to conduct tests. This is the major problem to do research on Indian scripts [23]. All the earlier studies have been reported on collection of small datasets from laboratory environment.

It is evident from the literature survey [23] 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 [23]. Therefore in the present work the first stage of research is to develop and build a handwritten Telugu character dataset.

Most of the Telugu characters are similar and recognizing such characters is highly challenging task. The number of vowels in the script is 16 and the number of consonants is 36. Identifying such similar characters is a very difficult task.

This paper deals with the handwritten character recognition for Telugu script written on paper documents. It includes the methodology used for handwritten Telugu database, the various preprocessing steps, feature extraction methods and various classifiers involved in the current work.

To develop the dataset, various scribers of different age groups are used to scribe on paper documents. These documents are scanned at 300 dpi and stored in the hard disk of the computer system. In the next step, preprocessing operations are performed to extract characters. The various feature extraction algorithms such as 'cell-wise pixel count' and 'Histogram profile' are employed to extract features from the preprocessed character images. In the proposed two-stage classification system, the classifiers employed are k-NN (k

Corresponding Author:
T. R.Vijaya Lakshmi, Department of Electronics and Communications Engineering, MGIT, Gandipet, Hyderabad, India.
E-mail: trvl.ece.mgit.phd@gmail.com

ISSN: 2231-2196 (Print) **ISSN:** 0975-5241 (Online) **DOI:** 10.7324/IJCRR.2017.9209
Received: 01.09.2017 **Revised:** 20.09.2017 **Accepted:** 07.10.2017

Nearest Neighbor) and SVM (Support Vector Machines) to classify the characters.

Data collection and preprocessing

Due to lack of standard data set to conduct experiments on handwritten Telugu characters [23], 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 preprocessed character samples are divided into folds. Each fold contains characters written by equal number of scribers. To test t_{th} fold, the remaining (V-1) folds are used as training. The average classification rates obtained from all these folds is considered as the classification rate/recognition accuracy of the model.

The number of characters considered for simulation is 18,000 containing 50 different classes, written by 360 different scribers. Thereby the number of samples per class is 360. All the characters are cross validated, by dividing them into 8 folds. Each fold contains characters written by 45 different scribers. The number of samples considered in each fold is 2,250 i.e., 50 × 45 (where 50 is number of classes and 45 is the number of scribers). To test a fold of characters, the remaining 15,750 characters are used as training. These preprocessed character images are used in the proposed step by step algorithm as discussed below.

RESEARCH METHODOLOGY

The flowchart of the proposed two-stage classification strategy for handwritten Telugu characters is shown in Figure 1. The raw preprocessed character image after noise removal and character extraction phases is first transformed into useful features. Each character image is represented in the form of a vector after the feature extraction stage.

Each fold of characters is tested in two stages. In the first stage of classification, 'Classifier-1' is trained with the training set to classify the characters under test fold. If the predicted class of the test character is same as that of its actual class then it is said to be recognized. The Recognition Accuracy (RA) of T_{th} testing fold is computed from the confusion matrix generated and is depicted in Equation (1).

$$RA = \frac{CR_1}{Total\ no,of\ characters\ tested} \times 100\% \quad (1)$$

where CR_1 is the number of characters correctly classified in stage-1.

Based on the confusion matrix generated by 'Classifier-1' in the first stage, the most confusing Telugu characters are found out and are classified in the second stage of classification. The unrecognized characters of the T_{th} test fold from the first stage are stored in a bin and are tested in the second stage of classification. To improve the character recognition rate, the unrecognized characters from the first stage are once again classified using another classifier i.e., 'Classifier-2'. To classify the unrecognized characters in the second stage, 'Classifier-2' is trained with the same training set, as indicated in Figure 3. The overall recognition accuracy (ORA) of the two-stage classification system for T_{th} testing fold is computed as depicted in Equation (2).

$$ORA = \frac{CR_1+CR_2}{Total\ no,of\ characters\ tested} \times 100\% \quad (2)$$

where CR_2 is the number of characters classified in stage-2.

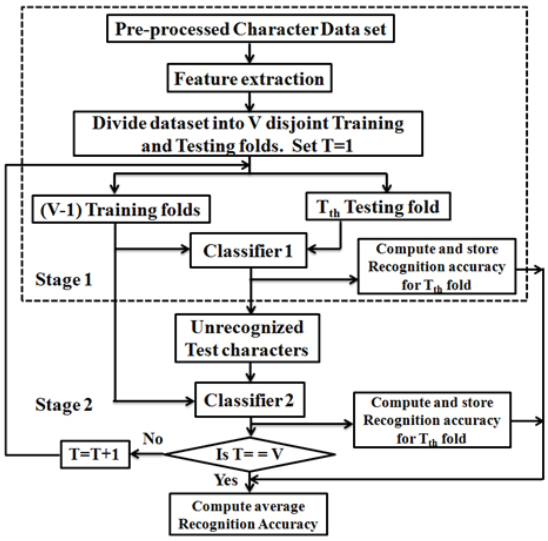

Figure 1: Flowchart of proposed Multi-stage classification strategy.

The overall classification rate is improved with this two-stage classification strategy. The procedure is repeated for all the 'V' folds and the average recognition accuracy from all these folds is considered as the recognition accuracy of the model. The various features extracted in the two-stage classification strategy are as follows.

FEATURE EXTRACTION

a. Cell-wise pixel count: The image, I, is divided into cells. The number of cells obtained from a M × M character pattern is M^2/n^2, where $n \times n$ is the size of the cell. The number of object pixels is counted in each cell/zone i.e., the pixels distributed in various cells are considered as features to classify the handwritten Telugu characters.

For each cell/zone i.e., say z_{11} the number of foreground pixels are summed up and is considered as a feature. This procedure is repeated for other cells. The features computed from these cells/zones of the character image are concatenated to form a feature vector, represented by $C_f = [z_{11}, z_{12}, z_{13}, ... z_{54}, z_{55}]$. Hence for an image, I, in the proposed work 25 features are extracted (for M=50). In this way for all the database images feature set consisting of 25 features for each image are extracted.

b. Histogram profile: The flowchart of histogram profile is shown in Figure 2. The histograms of the character image are computed along four directions. This is described in the flowchart. All these profiles are appended to form a feature vector of size 298 for a normalized character image of size 50 × 50.

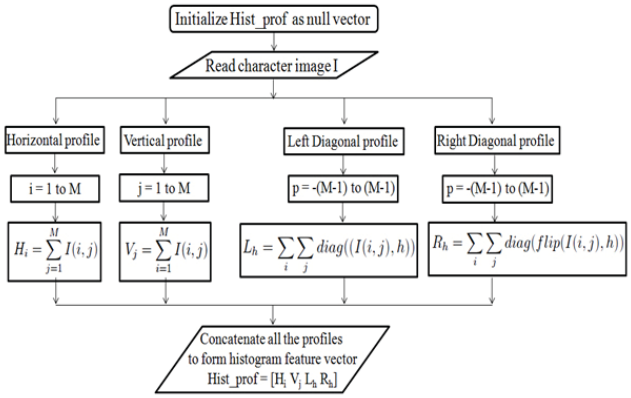

Figure 2: Flowchart of histogram profile.

RESULTS

The two-stage classification model is developed on a system having i5 processor of 2.2 GHz CPU clock speed with 4 GB RAM and 64 bit operating system running with Windows 8.1 using MATLAB 2014a. The number of characters considered for simulation is 18,000 from 50 different classes, written by 360 different scribers. The number of samples per class is 360. All the characters are cross validated, by dividing them into training and testing sets. Each fold contains characters written by 45 different scribers. The number of samples considered in each fold is 2,250 i.e., 50 × 45 (where 50 is number of classes and 45 is the number of scribers). To test a fold of characters, the remaining 15,750 characters are used as training. With 8-fold cross validation all the characters are tested once, provided the training and testing sets are disjoint. The average classification rates obtained from all these folds is considered as the classification rate/recognition accuracy (RA) of the model.

In the first stage, once the characters undergo tests using k-NN (k-nearest neighbor) classifier, the unrecognized character images from this stage are forwarded to undergo classification in the second stage. In the second stage, SVM (Support Vector Machine) classifier is trained with the same training set to classify only the unrecognized characters from the test set. This is done to improve the recognition accuracy and to reduce the misclassification rate. The recognition accuracies obtained with the two stage classification system are tabulated in Table 1.

Table 1: Two stage results obtained using k-NN and SVM classifiers

Feature extraction	% Recognition accuracy		
	k-NN	SVM	k-NN+SVM
Cell-wise pixel count	82.3	88.4	90.8
Histogram profile	73.5	77.9	79.3

DISCUSSIONS

It is evident from Table 1 that with the framework of two stage classification, there is a significant improvement in recognition rates. With the two stage classification framework, the cell-based approach gave a quantum improvement in recognizing the handwritten Telugu characters, compared to the Histogram profile-based approach. An improvement of 4-5% in recognition accuracy is obtained using the two-stage framework with both the feature extraction approaches.

CONCLUSIONS

There is no standard dataset for Indian scripts to conduct experiments for handwritten character recognition. Hence, in this work a dataset containing 18,000 handwritten Telugu isolated basic characters is developed. The various feature extraction algorithms employed for character recognition are cell-wise pixel count and histogram profile. The performance of these feature sets are tested with the proposed two-stage classification system.

An improvement of 3-5% in recognition rate is achieved with the proposed two-stage classification system when compared to single-stage classification for the feature extraction approaches considered. The best recognition accuracy obtained using the proposed two stage classification framework is 90.8% with 'cell-wise pixel count' feature set.

ACKNOWLEDGEMENT

Authors acknowledge the immense help received from the scholars whose articles are cited and included in references of this manuscript. The authors are also grateful to authors / editors / publishers of all those articles, journals and books from where the literature for this article has been reviewed and discussed.

REFERENCES

1. M. Cheriet, M. E. Yacoubi, H. Fujisawa, and D. Lopresti, "Handwritten recognition research: twenty years of achievement... and beyond," *Pattern Recognition*, vol. 42, no. 12, pp. 3131 – 3135, 2009.
2. A. Amin, "Off line Arabic character recognition: a survey," *Proceedings of the Fourth International Conference on Document Analysis and Recognition*, vol. 2, pp. 596–599, 1997.
3. M. S. Khorsheed, "Off-line Arabic character recognition–a review," *Pattern analysis & applications*, vol. 5, no. 1, pp. 31–45, 2002.
4. T.-H. Su, T.-W. Zhang, D.-J. Guan, and H.-J. Huang, "Off-line recognition of realistic Chinese handwriting using segmentation-free strategy," *Pattern Recognition*, vol. 42, no. 1, pp. 167–182, 2009.
5. P.-K. Wong and C. Chan, "Off-line handwritten Chinese character recognition as a compound Bayes decision problem," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 9, pp. 1016–1023, 1998.
6. 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.
7. B. Chaudhuri, U. Pal, and M. Mitra, "Automatic recognition of printed Oriya script," *Sadhana*, vol. 27, no. 1, pp. 23–34, 2002.
8. S. Antani and L. Agnihotri, "Gujarati character recognition," *Proceedings of the Fifth International Conference on Document Analysis and Recognition (ICDAR)*, pp. 418–421, 1999.
9. P. P. Kumar, C. Bhagvati, A. Negi, A. Agarwal, and B. L. Deekshatulu, "Towards improving the accuracy of Telugu OCR systems," *International Conference on Document Analysis and Recognition (ICDAR)*, pp. 910–914, 2011.
10. B. Chaudhuri and U. Pal, "A complete printed Bangla OCR system," *Pattern recognition*, vol. 31, no. 5, pp. 531–549, 1998.
11. L. M. Lorigo and V. Govindaraju, "Offline Arabic handwriting recognition: a survey," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 28, no. 5, pp. 712–724, 2006.
12. N. Arica and F. T. Yarman-Vural, "An overview of character recognition focused on off-line handwriting," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 31, no. 2, pp. 216–333, 2001.
13. G. Nagy, "Chinese character recognition: a twenty-five-year retrospective," *9th International Conference on Pattern Recognition*, pp. 163–167, 1988.
14. U. Pal and BB. Chaudhuri, "Indian script character recognition: a survey," *Pattern Recognition*, vol. 37, no. 9, pp. 1887–1899, 2004.
15. T.R.Vijaya Lakshmi, P.N. Sastry and T.V.Rajinikanth, "A novel 3D approach to recognize Telugu palm leaf text," *International Journal of Engg. Science and Technology*, vol. 20, no.1, pp. 143–150, 2017.
16. P.N. Sastry, T.R. Vijaya Lakshmi, K. Rama Krishnan and N. V. K. Rao, "Modeling of palm leaf character recognition system using transform based techniques," *Pattern Recognition Letters*, vol. 84, pp. 29–34, 2016.
17. P. N. Sastry, T.R. Vijaya Lakshmi, N.V. Koteswara Rao, Krishnan Rama Krishnan, 2017, "A 3D Approach for Palm Leaf Character Recognition Using Histogram Computation and Distance Profile Features," *Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications, Advances in Intelligent Systems and Computing*, Odisha, pp. 387-395, Sept. 16-17, 2017.
18. T.R. Vijaya Lakshmi, P.N. Sastry, and T.V. Rajinikanth, "Hybrid approach for Telugu handwritten character recognition using k-NN and SVM classifiers," *International Review on Computers and Software*, vol. 10, no. 9, pp. 923–929, 2015.
19. T.R.Vijaya Lakshmi, P.Narahari Sastry, and T.V.Rajinikanth, "Feature optimization to recognize Telugu handwritten characters by implementing DE and PSO techniques," *Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications, Advances in Intelligent Systems and Computing*, Odisha, pp. 397-405, Sept. 16-17, 2017.
20. 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, pp. 1-4, 2014.
21. 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.
22. Ramappa, Mamatha Hosahalli, Sucharitha Srirangaprasad, and Srikantamurthy Krishnamurthy, "An approach based on feature fusion for the recognition of isolated handwritten Kannada numerals,"
23. *International Conference on Circuits Power and Computing Technologies*, pp. 1496-1502, 2014.
24. 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.