ISSN:0975-9646

Offline Handwritten Character Recognition in South Indian Scripts: A Broad Visualization

Sunitha Anne M. O. Chacko*, Ansu Joseph*, Jeena Joji Anchanattu*, Sreelakshmi .S * , Veena A. Kumar**

**Student* ***Assistant Professor*
Department of Computer Science & Engineering, Saintgits College of Engineering, Pathamuttom.Kottayam ,686532

***Abstract* – Offline character recognition is one of the most challenging research areas in the pattern recognition domain. Even though many promising research results were reported in the area of handwritten character recognition for languages like English, Chinese, Japanese and Arabic, the problem is much more complicated for the Indian scenario. This paper presents an overview of the important advances in the offline handwritten character recognition domain of four major South Indian languages- Kannada, Tamil, Telugu and Malayalam.**

Keywords – Character Recognition, Classification, Feature Extraction, South Indian Languages.

I. INTRODUCTION

Character recognition is a special branch of pattern recognition which refers to the translation of handwritten or printed text into machine readable text. Offline handwriting recognition system has versatile range of applications including processing of bank cheques, mail addresses, white board reading, recognition of handwritten manuscripts etc.

Handwritten character recognition is divided into: online and offline recognition. The difference originates from the type of input data that is available for recognition. In online recognition a special input device, e.g. an electronic pen, tracks the movement of the pen during the writing process. No time information is available in offline handwriting recognition. Only an image of the handwriting is processed. Because less information is available, offline recognition is usually considered more difficult and challenging than online recognition.

Even though many promising research results were reported in the area of handwritten character recognition for languages like English, Chinese, Japanese and Arabic, the scenario is not so good for the Indian languages. Most of the works in the Indian character recognition domain deal with Devanagari and Bangla, the two most popular scripts in India.

India is a multi-lingual, multi-script country with twenty two scheduled languages. Most of the Indian scripts are originated from ancient Brahmi script. The South Indian languages are derived from Kadamba and Grantha scripts of ancient Brahmi. Here we present a detailed study on the on offline recognition of south Indian scripts- Malayalam, Tamil, Telugu and Kannada.

The paper is structured as follows: Section I presents an introduction of the paper. Detailed studies on Tamil, Telugu, Kannada and Malayalam character recognition systems are presented in section II, III, IV and V respectively. Section VI concludes the paper.

II. STUDIES ON TAMIL CHARACTER RECOGNITION

Tamil is the official language of the Indian state of Tamil Nadu and the union territories of Pondicherry and the Andaman and Nicobar Islands. The Tamil script has 10 numerals, 12 vowels 18 consonants and five grantha letters. The complete script, consists of 31 letters in their independent form, and an additional 216 combining letters representing every possible combination of a vowel and a consonant.

Bhattacharya et al. [1] proposed a two-stage recognition scheme for handwritten Tamil characters. In the first stage, an unsupervised clustering method is applied to create a smaller number of groups of handwritten Tamil character classes. Here an input character is grouped with one of few smaller groups of characters using use K-means clustering (KMC) technique. In this stage, component transition is used as feature vector. In the second stage, a supervised classification is considered in each of these smaller groups for final recognition. Chain code histogram features computed from the contour of the input character along with a distinct MLP classifier for each group to recognize the characters is used in the 2nd stage. The recognition accuracies of 92.77% and 89.66% were obtained respectively on the training and test sets.

Sutha and Ramaraj [2] proposed an approach to recognize handwritten Tamil characters using fourier descriptor based features. After boundary is obained, fourier descriptors are obtained. The 16 invariant fourier descriptors form input to the neural network. A multi-layer perceptron (MLP) with one hidden layer using back-propagation algorithm was used as classifier. A recognition accuracy of 97% was obtained for this system.

Gandhi and Iyakutti [3] used Kohonen neural network based Self Organizing Map (SOM) for handwritten Tamil character recognition. Initially during the pre-classification phase, the similar characters are grouped together. Here, the characters were categorized into two groups: crux characters group where the characters of the first group lie in the two baselines and Exhaustive group where the character that cross the base line. This exhaustive group was further divided into two sub groups : Ascending exhaustive characters and Descending exhaustive characters. Members of pre-classified groups are then analyzed using a SOM for final recognition. A

recognition rate of around 79.9% was achieved considering only the first choice and more than 98.5% was obtained for the top three choices.

Shanthi and Duraiswamy [4] proposed a scheme, in which pixel densities are calculated for 64 different zones of the character image are used for recognizing tamil characters. 64X64 sized image is divided into non-overlapping and overlapping zones of 8X8 which produced 64 feature vector for non-overlapping zones and 225 feature vectors for overlapping zones. These features were classified using an support vector machine (SVM). The system has achieved a recognition rate of 87.4%.

III. STUDIES ON TELUGU CHARACTER RECOGNITION

Telugu is the official language of Andhra Pradesh and also spoken by neighboring states. The script contains 10 numerals, 18 vowels, 36consonants and three dual symbols.

In the scheme proposed by Sitamahalakshmi et al. [5] for the recognition of handwritten Telugu numerals characters, the probability of identifying the given input character was obtained using five distance measurement methods: Similarity Based Methods, Hamming Method, Linear-Correlation Method, Cross-Correlation Method and Nearest Neighbor Method. The results obtained are then combined using the Dempster-Shafer theory (DST).

Sastry et al .[6] proposed a method for classification of Telugu characters extracted from the palm leaves, using Decision Tree approach which was developed using SEE5 Algorithm. Using a decision tree classifier, an overall accuracy of 93.10% was achieved.

Sastry and Krishnan [7] proposed a novel approach for telugu palm leaf character recognition using Radon Transform and Nearest Neighboorhood Classifier. After selecting any basic Telugu character from the given palm leaf, some of its pixel points along the character contour is identified. The Radon Transform of each image is found out and its absolute value is stored as the feature vector. The nearest neighborhood classifier is used as the classifier.

IV. STUDIES ON KANNADA CHARACTER RECOGNITION

Kannada is one of the oldest Dravidian languages and is the official language of Karnataka. Kannada alphabet consists of 49 characters- 13 Vowels, 2 part vowel, part consonants and 34 Consonants.The Kannada alphabets have evolved from the Kadamba and Calukya scripts.

Sangame [8] proposed a system for unconstrained handwritten Kannada vowels recognition based upon invariant moments. The invariant moments are evaluated using central moments of the image function up to third order. A Euclidian distance criterion and K-NN classifier is used for classification of the Kannada vowels. The system had overall accuracy of 85.53%.

Ragha and Sasikumar [9] proposed a method for extraction of moment features from Gabor directional images of handwritten Kannada characters. They find 4 directional binary images using Gabor wavelets from the dynamically preprocessed original images and then extract moments features from them.

Classification was performed using an MLP with back propagation. The overall performance of the system combined with these two features together is 92%.

Aradhya et al. [10] proposed a technique for kannada character recognition based on Fourier Transform and Principal Component Analysis (PCA). Fourier transform was combined with PCA to enhance the classification information and improve the recognition. For classification, Probabilistic Neural Network (PNN) was used.

Rajashekararadhya [11] proposed an offline handwritten numeral recognition technique for four south Indian languages like Kannada, Telugu, Tamil and Malayalam. They used a feature extraction technique, based on zone and image centroid. They used two different classifiers nearest neighbor and back propagation neural network to achieve 99% accuracy for Kannada and Telugu, 96% for Tamil and 95% for Malayalam.

V. STUDIES ON MALAYALAM CHARACTER RECOGNITION

Malayalam is the official language of Kerala. It is also spoken in the Union territories of Lakshadweep and Mahe. Malayalam script is derivedfrom the Grantha script which is an inheritor of the old Brahmi script. The complete character set includes 15 vowels, 36 consonants , 5 chillu , 3 consonant signs, 9 vowel signs, anuswaram, visargam, chandrakkala and 57 conjunct consonants. It also includes 9 numerals which are seldom used.

The first work in Malayalam OCR was reported by Lajish [12] which used fuzzy-zoning and normalized vector distance measures for the recognition of characters. Here, the size normalized image were divided into 3x3 uniform sized zones. The 9 fuzzy zones thus obtained are classified as corner regions, peripheral regions and the central region. The normalized vector distances were computed for each zone and fuzzification was performed on these. The 9 features, thus obtained from these zones were classified using class modular neural network. The system had attained an overall accuracy of 78.87% for the 44 Malayalam handwritten characters.

Chacko and Anto [13] proposed a method for producing smooth skeletons of Malayalam handwritten characters. Here skeleton pruning was done by contour portioning with discrete curve evolution. The DCE is used to obtain a hierarchical partitioning of character contour into subarcs that yields a hierarchical skeleton structure. The overall accuracy of the system was 90.18% for 33 character classes.

[14] presents a novel method for offline recognition of isolated basic Malayalam characters using a combination of Chain Code Histogram (CCH) and Differential Chain Code Histogram (DCCH) based features. An average accuracy of 92.75% was obtained for the proposed system using a neural network classifier.

Paper	Features	Classifier	Accuracy
Bhattacharya [2007]	Component Transition, CCH	KMC, MLP	89.66
Sutha and Ramraj [2007]	Fourier Descriptors	MLP	97
Shanthi and Duraiswami [2010]	Pixel Density	SVM	82.04
Sastry et al. [2010]	3D	Decision Tree	93.10
Sitamahalakshmi et al. [2010]	-	DST	87.3
Sastry and Krishnan [2012]	Radon Transform	Nearest Neighbor Classifier	93
Sangame [2012]	Invariant Moments	Euclidean distance, KNN	85.53
Ragha and Sasikumar [2010]	Moments, Gabor	MLP	92
Aradhya[2010]	Fourier transform, PCA	PNN	68.89
Lajish [2007]	Fuzzy zoning, NVD	CMNN	78.87
Anitha et.al [2014]	CCH,DCCH	Feedforward NN	92.75
Jomy John et al. [2011]	CCH, Image centroid	Neural Networks	72.1
Anitha et.al [2014]	Moment invariants, Projection, Gradient	Feedforward NN	96.16

The authors have proposed another novel method for the recognition of offline isolated Malayalam characters based on a combination of global and local features [15].The global features extracted are the moment invariants and projection features and the gradient features of the characters form the local feature. A two layer feedforward neural network is used as the classifier. The proposed method achieves a recognition accuracy of 96.16% using a five-fold cross validation technique for the 13 class recognition problem.

In [16], a novel method for Offline Malayalam Character Recognition is proposed using multiple classifier combination technique. From the preprocessed character images, they have extracted two features: Chain Code Histogram and Fourier Descriptors. These features were fed as input to two feedforward neural networks. The results of both neural networks were combined using a weighted majority technique. The proposed system achieves an accuracy of 92.84% and 96.24% respectively for the writer independant and writer dependant scheme considering top 3 choices.

In [17], Jomy John et al. proposed a method based on chain code and image centroid for the recognition of Malayalam vowels. From the chain code representation of the character, a chain code histogram and Normalized chain code histogram were constructed which were used for classification process. A two layer feed forward network with scaled conjugate gradient was used for classification. An average accuracy of 72.1% was obtained for the system.

VI. CONCLUSION

This paper presented a detailed study on the different handwritten character recognition works so far developed for the four South Indian languages- Kannada, Tamil, Telugu and Malayalam. This problem demands more attention as a complete OCR system for these languages has not yet been developed. One of the major challenges encountered in this field is the lack of a benchmark database. We believe that our survey aid researchers working in the handwritten character recognition domain of South Indian languages.

REFERENCES

[1]. U. Bhattacharya, S. K. Ghosh and S. K. Parui, “A Two Stage Recognition Scheme for Handwritten Tamil Characters”, Ninth International Conference on Document Analysis and Recognition" (ICDAR 2007)

[2] Sutha, J. and Ramaraj, N.,”Neural network based offline Tamil handwritten char- acter recognition system", In Proceedings of ICCIMA07,2007 ,pp 446-450.

[3] Gandhi, R. I. and Iyakutti, K, “An attempt to recognize handwritten Tamil character using Kohonen SOM", Int. J. Advanced Network. Appl. 1, 3,2009, pp-188-192.

[4]. N. Shanthi and K. Duraiswamy,”Performance Comparison of Different Image Sizes for Recognizing Unconstrained Handwritten Tamil Characters using SVM”, Journal of Computer Science 3 (9): 760-764, 2007 ISSN 1549-3636

[5] Sitamahalakshmi, T., Babu, V., and Jagadeesh, M, “Character recognition using Dempster- Shafer theory combining different distance measurement methods". Int.J. Engin. Sci. Technol. 2, 5,2010, pp 1177-1184

[6] Sastry, P. N., Krishnan, R., and Ram, B. V S, Classification and identification of Telugu handwritten characters extracted from palm leaves using decision tree approach", J. Applied Engn. Sci. 5, 3,2010,pp 22-32

[7] Panyam Narahari Sastry, Ramakrishnan Krishnan, “Isolated Telugu Palm Leaf Character Recognition Using Radon Transform-A Novel Approach", 2012 World Congress on Information and Commmunication Technologies, pp 796-802.

[8] Sangame S.K., Ramteke R.J., Rajkumar Benne, “Recognition of isolated handwritten Kannada vowels”Copyright 2009, Bioinfo Publications, Advances in Computational Research, ISSN: 09753273, Volume 1, Issue 2, 2009 pp-52-55

[9] L R Ragha, M Sasikumar, “Using Moments Features from Gabor Directional Images for Kannada Handwriting Character Recognition ,International Conference and Workshop on Emerging Trends in Technology (ICWET 2010) TCET, Mumbai, India.

[10] Aradhya, V. N. M, Niranjan,S. K., AND Kumar, G. H.,”Probabilistic neural network based approach for handwritten character recognition", Int. J. Comp. Comput. Technol. (Special Issue) 1,2010.

[11] Rajasekararadhya SV, Ranjan PV,”Efficient zone based feature extraction algorithm for handwritten numeral recognition of popular south Indian scripts", J Tech Appl Inform Technol 7(1):pp 1171-1180 (2009)

[12] Lajish V. L., “Handwritten character recognition using perpetual fuzzy zoning and class modular neural networks”, Proc. 4th Int. National conf. on Innovations in IT, 2007, pp 188-192.

[13] Binu P. Chacko, Babu Anto P, "Discrete Curve Evolution Based Skeleton Pruning for Character Recognition", Seventh International Conference on Advances in Pattern Recognition, 2009.
[14] Anitha Mary M.O. Chacko., Dhanya, P.M.:, "A differential chain code histogram based approach for offline Malayalam character recognition. In: International Conference on Communication and Computing (ICC-2014), pp. 134–139 (2014).
[15] Anitha Mary M.O. Chacko, Dhanya P.M., Offline Malayalam Character Recognition Using Global and Local Features, Proceedings of the Second International Conference on Emerging Research in Computing, Information, Communication and Applications, *ERCICA*, 2014, Elsevier Publications, p. 806-812.
[16] Anitha Mary M.O. Chacko, Dhanya P.M., "Combining Classifiers for Offline Malayalam Character Recognition *Emerging ICT for Bridging the Future – Volume 2*, Advances in Intelligent Systems and Computing 338, DOI: 10.1007/978-3-319-13731-5_3, Springer International Publishing Switzerland 2015.
[17] Jomy John, Pramod K. V. and Kannan Balakrishnan, "Offline Handwritten Malayalam Recognition Based on Chain Code Histogram", Proceedings of ICETECT, pp. 736–741, (2011).