International Journal of Advance Research in Science and Engineering
Vol. No.6, Issue No. 08, August 2017
www.ijarse.com
IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

A SURVEY ON OFFLINE RECOGNITION OF SOUTH INDIAN SCRIPTS

Krupashankari S Sandyal[1], Dr. Y C Kiran[2]

[1] *Department of Information Science & Engg. Dayananda Sagar College of Engg. Bangalore (India)*

[2] *Department of Computer Science & Engg. BNM Institute of Technolog, Bangalore (India)*

ABSTRACT

Handwritten character recognition is always a frontier area of research in the field of pattern recognition. Even though, sufficient studies have performed in foreign scripts like Arabic, Chinese and Japanese, only a very few work can be traced for handwritten character recognition mainly for the south Indian scripts. Multiple combinations of vowels and consonants along with its modifiers led to generation of huge number of classes with respect to character recognition systems. The feature extraction and classification of characters from such huge number of classes in south Indian language Optical Character Recognitions remains as a non-trivial problem.OCR system development for Indian script has many application areas like preserving manuscripts and ancient literatures written in different Indian scripts and making digital libraries for the documents. Feature extraction and classification are essential steps of character recognition process affecting the overall accuracy of the recognition system. This paper presents a brief overview of digital image processing techniques such as Feature Extraction and Image classification.

Keywords— Feature extraction, Classification, Neural network, Support vector machine, Handwritten character recognition;

I. INTRODUCTION

Handwritten character recognition has been divided into offline and online character recognition. Offline documents are scanned images of handwritten text, generally on a sheet of paper. In online recognition a special input device, e.g. an electronic pen, tracks the movement of the pen during the writing process. 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. System recognition of offline handwritten documents has been an active research area in the field of pattern recognition.

Over the last few years, the numbers of laboratories all over the world are involved in research on handwriting recognition. The recognition of cursive handwriting is very difficult due to large number of variations in shapes and overlapping of characters. In the offline handwriting recognition system, the pre-written document is converted into a digital image through optical scanner. In the handwritten script, there are variations in writing style of different age groups. Handwritten Character Recognition (HCR) system will enable the computer to process the handwritten documents, which are currently processed manually. One can find these handwritten documents at various places such as post offices, banks, insurance offices and colleges etc. for processing data.

India is a multi-lingual, multi-script country with 22 scheduled languages. Most of the Indian scripts are originated from ancient Brahmi script. The South Indian languages are derived from Kadamba and Grantha

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

scripts of ancient Brahmi. Here detailed study on the offline recognition of south Indian scripts Malayalam, Tamil, Telugu and Kannada is been presented.

Brief History on Offline Recognition of South Indian Scripts

Many methods have been proposed for character recognition they are often subjected to substantial constraints due to unexpected difficulties.

II. STUDIES ON TAMIL CHARACTER RECOGNITION

Tamil handwritten OCR is more complicated as the letters consists of more angles and modifiers. Additionally, Tamil script contains large number of character sets. A total of 247 characters; consisting of 216 compound characters, 18 consonants, 12 vowels and one special character. Challenges that researches face during recognition process are due to the curves in the characters, number of strokes and holes, sliding characters, differing writing styles so on.

Researchers have come up with many approaches for the character recognition, however, some of them have surveyed in the paper.

Tiji M Jose et al [1] illustrated the wavelet decomposition technique for the extraction of the features from the Tamil characters. The feed forward back propagation network classifier is used for the intention of classification. The recognition rate achieved in this paper was about 89%.

Indra Gandhi et al proposed a new approach of using Kohonen SOM (Self Organizing Map) for recognizing the online Tamil character [2]. The vectors of the binary image are created. When the segmentation of the character is over, then the images are scaled to unique height and weight. Some unwanted portions are included, but it can be removed by sobel edge detection. The median filter is used to increase the efficiency. The SOM is not applicable to the cursive characters which are used in this paper. The median filter is not suited for the offline Tamil characters. So the wavelets are used for feature extraction.

Jagadeesh Kannan et al [3] used Octal Graph method for the recognition of the Tamil Handwritten characters. Here, the character return on the octal graph's pixel is converted into the node of the graph. Each node has eight fields, that's why called as octal graph. Each node is connected to the other node based on the threshold value. The image is converted to the octal graph by the steps such as normalization, conversion, Identification of weighing factors and feature extraction. If the character is tedious and if it contains many curves, then octal graph method is not suitable.

A. Feature Extraction

a) Structural Technique:

b) Scale Invariant Feature Transform (SIFE) [4] is used to transfer the character image into a set of local features. Using this approach, 128 dimensions of SIFE features (Interesting points) are identified from the character image.

c) Statistical Technique:

In the Zone based method, the normalized characters are divided into non interleaving zones. Pixel density is calculated for each zone. Images are divided into 9 non overlapping blocks of equal size using the Gabor channel method. This gives 24 responses for blocks passing through each channel. Mean and standard deviations are calculated and used as features.

d) Hybrid Technique:

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

Hough Transform has been used to detect the horizontal and vertical lines. Bilinear interpolation has been used to extract the features such as slant and strip. A feature extraction technique Zone based hybrid approach, which has been used to extract the zone centroid and Image centroid based distance metric features.

B. Classification

K-Nearest Neighbour approach used as classifier to recognize the character sets and better accuracy. The result of the recognition is obtained using the Histogram Equalization. In the Multi Layer Perception (MLP) for 2 hidden layer approach, a Neural Network (NN) is designed with all weights mapped to a random number between 1 and -1. 2 hidden layers used for better performance. If the hidden layers are increased then the performance will be reduced [5][6]. In some works, Hidden Markov Model (HMM) has been used to support handwritten characters recognition model [7]. It provides the maximum probable result.

III. STUDIES ON KANNADA CHARACTER RECOGNITION

The Kannada alphabet is classified into 2 main categories: vowels and consonants [8]. There are 16 vowels and 35 consonants. Words in Kannada are composed of aksharas, which are analogous to characters in an English word, while vowels and consonants are aksharas, the vast majority of aksharas are composed of combinations of these in a manner similar to most other Indian scripts.

Rajashekar Aradhya and Manjunath Aradhya et al. described new method for feature extraction based on vertical projection distance metric and zoning [8]. In this method, the image is divided into 25 equal parts. Hence for each grid, there will be 10 columns and 10 rows. For each grid column they compute the pixel distance that is vertical projection distance metric. If there is more than 1 pixel then compute average pixel distance to get feature vector of that grid column. Repeat this for rest of the columns in that grid.

In [9], Mamatha H.R. et al. used the Run Length Count (RLC) method for feature extraction. First of all image is divided into zones and for each zone apply RLC to get features. In RLC whenever there is change in pixel value that is from 0 to 1 or vice versa count is taken for each horizontal and vertical columns.

A. Feature Extraction

Feature extraction is problem dependent. Good features are those whose values are similar for objects belonging to the same category and distinct for objects in different categories. The better approach for recognition is to segment characters into basic symbol and recognize each symbol subsequently. The most important aspect of handwriting recognition scheme is the selection of good feature set, which is reasonably invariant with respect to shape variations caused by various writing styles.

Various feature extraction methods employed for recognizing the normalized segmented characters:

a) Hu's Invariant Moments:

Moment invariants have been frequently used as features for image processing, remote sensing, shape recognition and classification. Moments can provide characteristics of an object that uniquely represent its shape. It was Hu (Hu, 1962), that first set out the mathematical foundation for two-dimensional moment invariants and demonstrated their applications to shape recognition .The seven moment invariants are defined as Hu's seven moment invariants have been widely used in pattern recognition, and their performance has been evaluated under various deformation situations.

b) Zernike Moments:

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

Zernike moments [10] are used for extracting the features of printed digits in grayscale images. Zernike moments are pure statistical measure of pixel distribution around center of gravity of characters and allow capturing global character shapes information. They are designed to capture both global and geometric information about the image. Moment-based invariants explore information across an entire image rather than providing information just at single boundary point, they can capture some of the global properties missing from the pure boundary-based representations like the overall image orientation. In Hu's moment invariants, the whole concept is based on the central moments which have integrated the translation and scale normalization in the definitions. The Zernike moments, however, are only invariant to image rotation for themselves. To achieve translation and scale invariance, extra normalization processes are required. The translation normalization is achieved by moving the image center to the image centroid.

c) Zoning:

In this method the image is split into different zones and simple features are extracted from each of the zones. In this method, the segmented character is first area normalized so that the numbers of ON pixels in all the normalized characters are equal. The normalized character is divided into smaller zones. Various regional features such as minor axis length, major axis length, centroid, eccentricity, convex area are calculated. Along with the regional features, structural features such as geometric moments, variance are calculated and used as feature vector.

d) Fourier- Wavelet Coefficient:

Fourier transform is a powerful tool for pattern recognition. Fourier transform is translation and rotation invariant, but the frequency information of Fourier transform is global and so a local variation of the shape will affect the Fourier coefficients. Wavelet transform has multi-resolution ability but is translation variant. A small shift of the original signal will lend totally different wavelet coefficients. Therefore Fourier and Wavelet transforms are combined to obtain a feature vector which is not only invariant to translation and rotation, but also has multi-resolution ability.

B. Classification

The extracted *features* are given as the input to the classification process. A bag of key points extracted from the feature extraction approaches are used for classification.

There are some approaches that are used to classify the character features in the existing systems such as K-Nearest Neighbour approach (KNN), Fuzzy system, Neural Network (NN), discriminate classifier, unsupervised classifier and so on.

IV. STUDIES ON TELUGU CHARACTER RECOGNITION

Telugu script is other Indian script which is generally written in non-cursive style, unlike English handwriting, which is normally written in cursive style rendering recognition difficulty. Telugu language is mostly spoken in South India; it’s the script which is complex of all Indian scripts because of 2 reasons: it has the largest number of vowels and consonants and it has Complex Composition Rules. The Telugu Script Consists of 14 Vowels, 36 Consonants and 3 Special Characters.

The first reported work on OCR of Telugu Character is by Rajasekharan et al [11]. This work identifies 50 primitive features and proposes a 2-stage syntax-aided character recognition system. Primitives are joined and superimposed appropriately to define individual characters.

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

In Sukhaswamy [12] a Neural Network based system was proposed to recognize Telugu script. An extensive study is undertaken to identify the structural characteristics of Telugu script and the distinct symbols of the Telugu language are categorized based on their relative size.

A. Feature Extraction

A feature-based approach, rather than a template-based one, is perhaps more appropriate for handwritten character recognition, considering the extent of distortion that handwritten characters undergo. In an earlier work by V.S. Chakravarthy et al [13], certain general features have been identified – known as the shape features of handwritten characters, which are less susceptible to distortion introduced by writing.

a) Candidate Search (Zoning):

For a candidate searching measure of density of pixel distribution is used in different zones of the input glyph as a feature vector. First the input glyph is broken into zones by super-imposing a grid and then the percentage of the number of foreground pixels is calculated. A codebook of this feature vector is pre-computed from the training set. The feature vector of the input glyph is computed and searched in the codebook to obtain k nearest neighbors. The distance measure is Euclidean Distance between the feature vectors.

b) Cavity Based Structural Analysis:

Many Telugu characters have cavities (holes) in them. Cavities are used as structural features in the recognition process. The existence and position of these cavities is a structurally distinguishing feature. Cavities are used since they provide discrimination between glyphs which are could be very confusing for recognition.

B. Classification

A HMM is a collection of finite states connected by transitions. Each state is characterized by two sets of probabilities: a transition probability and either a discrete output probability distribution or continuous output probability density function. This gives the condition probability of emitting each output symbol from a finite alphabet or a continuous random vector. SVM is generalized as kernel machines and are maximum margin methods for classification. SVM is based on finding maximum separability or margin between the different classes with the help of training data features. The main idea during training phase is to find the parameters for the hyper plane, which maximally separates the different classes involved in the problem.

V. STUDIES ON MALAYALAM CHARACTER RECOGNITION

Malayalam is a Dravidian language with about 35 million speakers. It is spoken mainly in the south western India, particularly in Kerala. Until the 16th century Malayalam was written in the vattezhuthu script. Modern Malayalam script is derived from the Grantha script, a descendant of the ancient Brahami Script. There are 53 letters, called akaras [14]. As a result of the difficulties of printing Malayalam, a simplified or reformed version of the script was introduced during the 1970s and 1980s. The main change involved writing consonants and diacritics separately rather than as complex characters. These changes have not been consistently so the modern script is often a mixture of traditional and simplified letters [15]. The character set consists of 13 vowels, 2 left vowel signs, 7 right vowel signs, some appear on both sides of the conjunct/consonant, 30 commonly used conjuncts.

The first work in Malayalam OCR was reported by Lajish [16]. It used fuzzy-zoning and normalized vector distance measures for the recognition of segmented characters. The size normalized image was divided into 3x3

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

uniform sized zones. The normalized vector distance for each zone is computed and fuzzification is performed on these. The 9 features, thus obtained were classified using class modular neural network. The major motivation of using modular neural networks instead of monolithic neural network is to improve the capacity of the neural networks. The system had attained an overall accuracy of 78.87% for the 44 Malayalam handwritten characters.

Wavelets were applied by G. Raju [17] for the recognition of isolated Malayalam characters. Author used Daubechies 4 wavelet (db4), a member of the Daubechie wavelet family with order 4, for decomposition into ten sub-images. The count of zero-crossing in each of the ten sub bands were used for classification. From the analysis of zero crossings, 25 consonant characters that were taken in the data set could be classified into 11 sets. No preprocessing steps were done in this method.

A. Feature Extraction

Feature extraction is the process of extracting relevant features of the characters to form feature vectors which are used by classifiers for the recognition process. The feature extraction methods for handwritten character recognition can be classified into Statistical, Structural and Hybrid techniques [18]. Statistical approaches use quantitative methods for extracting the features such as geometrical moments, projection histograms, direction histograms, crossing points. Structural approaches use qualitative measurements for feature extraction. These features are based on topological and geometrical properties of the character, like strokes, loops, end points, intersection points, etc. Hybrid approaches combines the features of these two techniques.

B. Classification

Classification is the final phase of character recognition, which is done by assigning labels to character images based on the features extracted. Bayesian classifier, Binary tree classifier, Nearest Neighbor Classifier[19], Neural Networks, Modified Quadratic Discriminant Function [20] and Support Vector Machines are some of the classifiers that are used for this purpose.

An Artificial Neural Network [21] is composed by a collection of artificial neurons interconnected among them to form a neuronal system able to learn and to understand the mechanisms. Each artificial neural network is characterized by its specific architecture; this architecture is denoted by the number of neurons of the input layer, the number of hidden layers, the number of neurons in each hidden layer and the neurons number in the output layer. The operating principle of artificial neural networks is similar to the human brain; first, it must necessarily pass on the learning phase to record knowledge in the memory of the artificial neural network. The storage of knowledge is the principle of reputation and compensation to a collection of data that forms the basis of learning.

VI. CONCLUSION

A study on the different handwritten character recognition works so far developed for four South Indian languages - Kannada, Tamil, Telugu and Malayalam is presented.This problem demands more attention as a complete Optical character recognition (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. This survey aid researchers working in the handwritten character recognition domain of South Indian languages.

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

REFERENCES

[1.] Tiji M Jose and Amitabh Wahi, "Recognition of Tamil Handwritten Characters using Daubechies Wavelet Transforms and Feed-Forward Backpropagation Network" International Journal of Computer Applications 64(8):26-29, February 2013.

[2.] Indra Gandhi R and Iyakutti K, "An attempt to Recognize Handwritten Tamil Character using Kohonen SOM", Int. J. of Advance d Networking and Applications, Volume: 01 Issue: 03 Pages: 188-192, 2009.

[3.] R. Jagadeesh Kannan and R. Prabhakar, "An Improved Handwritten Tamil Character Recognition System using Octal Graph", Journal of Computer Science 4 (7): 509-516, 2008.

[4.] Subashini A and Kodikara N.D, "A Novel SIFE-based Codebook Generation for Handwritten Tamil character Recognition" , 6th IEEE Int. Conf. on Industrial and Information Systems (ICIIS), Page(s): 261 – 264, 2011.

[5.] Stuti Asthana, Farha Haneef and Rakesh K Bhujade, "Handwritten Multiscript Numeral Recognition using Artificial Neural Networks", Int. J. of Soft Computing and Engineering ISSN: 2231-2307, Volume-1, Issue-1, March 2011.

[6.] Sutha J and RamaRaj N, "Neural network based offline Tamil handwritten character recognition System", International Conference on Conference on Computational Intelligence and Multimedia Vol : 2, page(s): 446 – 450, 2007.

[7.] El-Yacoubi, M. Gilloux,R. Sabourin, Member, IEEE, and C.Y. Suen, Fellow, IEEE , "An HMM-Based Approach for OffLine Unconstrained Handwritten Word Modeling and Recognition", IEEE Transactions On Pattern Analysis And MachineIntelligence, Vol. 21, No. 8, Auguest 1999.

[8.] Rajashekar aradhya S. V., Vanaja Ranjan P., Manjunath Aradhya V. N., "Isolated Handwritten Kannada and Tamil Numeral Recognition: A Novel Approach", First International Conference on Emerging Trends in Engineering and Technology- ICETET, pp.1192-1195,16-18 July 2008.

[9.] Mamatha H. R., Karthik S., Srikanta Murthy K., "Feature Based Recognition of Handwritten Kannada Numerals – A Comparative Study", International Conference onComputing, Communication and Applications (ICCCA),22-24 Feb, 2012.

[10.] Toharia P., Robles O.D., Rodríguez Á., Pastor L, "A Study of Zernike Invariants for Content-Based Image Retrieval", Advances in Image and Video Technology. PSIVT 2007.

[11.] S. N. S. Rajasekharan and B. L. Deekshatulu. Generation and Recognition of printed Telugu characters, Computer graphics and image processing 6, 335 – 360, 1977.

[12.] M. B. Sukhaswami, P. Seetharamulu, Arun K Pujari. Recognition of Telugu characters using neural networks. Int. J. of Neural Systems, 6(3):317-357, 1995.

[13.] V.S. Chakravarthy and B. Kompella, "The Shape of Handwritten Character," National Conference on Document Analysis and Character Recognition – NCDAR2001, Mandya, India. July 12-13, 2001.

[14.] Malayalam Script Features [Online]. Available: http://scriptsource.org/scr/Mlym

[15.] Malayalam [Online]. Available: http://www.omniglot.com/writing/malayalam.htm

[16.] 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

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

[17.] G. Raju, "Recognition of unconstrained handwritten Malayalam characters using zero-crossing of wavelet coefficients", Proc. of 14th International conference on Advanced Computing and Communications, 2006, pp 217-221.

[18.] Trier.O.D, Jain.A.K and Taxt.J, "Feature extraction methods for character recognition - A survey", Pattern Recognition, vol.29, no.4, pp.641-662, 1996.

[19.] Z.M.V. Kovacs and R. Guerrieri, "A generalization technique for nearest-neighbor classifiers", IEEE International Joint Conference on Neural Networks,1991.

[20.] Bindu S Moni and G Raju, "Modified Quadratic Classifier and Directional Features for Handwritten Malayalam Character Recognition", IJCA Special Issue on "Computational Science - New Dimensions & Perspectives NCCSE, 2011.

[21.] Le Hoang Thai, Tran Son Hai and Nguyen Thanh Thuy, "Image Classification using Support Vector Machine and Artificial Neural Network", I.J. Information Technology and Computer Science, 2012.