ANCIENT TAMIL CHARACTER RECOGNITION FROM TEMPLE WALL INSCRIPTIONS

Rajakumar S.[1], **Subbiah Bharathi V.**[2]

[1]Research Scholar, Sathyabama University, Chennai, India
[2]DMI College of Engineering, Chennai, India
Email: [1]rajkumarmeae70@gmail.com

Abstract

Recognition of any ancient Tamil characters with respect to any language is complicated, since the ancient Tamil characters differ in written format, intensity, scale, style, and orientation, from person to person. Researchers for the recognition of ancient Tamil languages and scripts are comparatively less with other languages, this is a result of the lack of utilities such as Tamil text databases, dictionaries etc. The problem of ancient Tamil character recognition is the technical challenge than other languages in respects to the similarity and complexity of characters that are composed of circles, holes, loops and curves. Hence ancient Tamil recognition requires more research to reach the ultimate goal of machine simulation of human reading. In this paper, we have made an attempt to recognize ancient Tamil characters by using SIFT features and presented a new and efficient approach based on bag-of key points representation. Collection of SIFT features are first extracted from local patches on the pre-processed images, and they are then quantized by K-means algorithm to form the bag-of-key points representation of the original images. These fixed-length feature vectors are used to classify the characters. A recognition system consists of the activities, namely, digitization, pre-processing, feature extraction and classification. This system achieves a maximum recognition accuracy of 84% using SIFT features.

Key words: Temple wall inscriptions, SIFT features, SVM, Character recognition, K-means algorithm

I. INTRODUCTION

Ancient Tamil character recognition has been one of the most fascinating and challenging research areas in the field of image processing and pattern recognition in recent years. It contributes immensely to the advancement of automation process and can improve the interface between man and machine in numerous applications. Several research works have been focusing on new techniques and methods that would reduce the processing time while providing higher recognition accuracy. In general, character recognition is classified into two types as off-line and on-line recognition methods. In the off-line recognition, the input image is usually captured by a high resolution Digital camera. But, in the on-line recognition, the two dimensional coordinates of successive points are represented as a function of time and the order of strokes made by the writer are also available. The Offline ancient Tamil text recognition is the process of finding letters and words that are present in the Ancient stone inscriptions. However, in the off-line systems, the neural networks have been successfully used to yield comparably high recognition accuracy levels.

The off-line ancient Tamil Character recognition continues to be an active area of research towards exploring the newer techniques that would improve recognition accuracy. The first important step in any ancient Tamil recognition system is pre-processing followed by segmentation and feature extraction. Pre-processing includes the steps that are required to shape the input image into a form suitable for segmentation. In the segmentation, the input image is segmented into individual characters and then, each character is resized into $m \times n$ pixels towards training the network. The selection of appropriate feature extraction method is probably the single most important factor in achieving high recognition performance. Several feature extraction methods are available to recognize ancient Tamil characters. The widely used feature extraction methods are Template matching, Image transforms, Projection Histograms, Contour profiles, Zoning, Geometric moment invariants, Fourier descriptors, Gradient feature and Gabor features. An artificial neural network as the backend is used for performing classification and recognition tasks. In the offline ancient Tamil character recognition system, the neural networks have emerged as the fast and reliable tools for classification towards achieving high recognition accuracy.

II. FEATURE EXTRACTION

A. Algorithm

The SIFT algorithm takes a character image and transforms it into a set of local features, each of which describes a local part around a key point. Each of these feature vectors is supposed to be distinctive and invariant to any scaling, rotation or translation of the image. The SIFT descriptor builds a representation for each key point based on a patch of pixels in its local neighbourhood. For each sample character, the dimensional SIFT features are calculated. All the SIFT features of a particular character are concatenated to make bigger feature space. In the feature extraction process, resized individual character of size 120×120 pixels is further divided into 54 equal zones, each of size 20×20 pixels. The features are extracted from the pixels of each zone by moving along their diagonals. This procedure is repeated for all the zones leading to extraction of 3 features for each character. These extracted features are used to train a feed forward back propagation neural network employed for performing classification and recognition tasks. Extensive simulation studies show that the recognition system using diagonal features provides good recognition accuracy while requiring less time for training. Fig. 1 shows the block diagram of Ancient tamil Character Recognition (ATCR) system.

B. Block Diagram

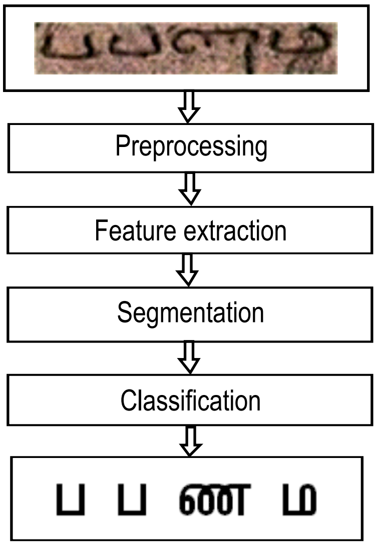

Fig. 1. Block Diagram of ATCR system

III. PROPOSED SYSTEM

The main steps of our method are:

- Perform various pre-processing operations to enhance the quality of ancient Tamil character stone inscription images.
- Detection of interest points in SIFT descriptors.
- Constructing visual codebooks by means of clustering techniques (K-means). The codebook is the set of centres of the learnt clusters.
- Constructing a bags of key points, which counts the number of patches assigned to each cluster.
- Applying a SVM classifier, treating the bag of key points as the feature vector, and thus determine which character to assign to the image.
- Selection of a codebook and classifier giving the best overall classification accuracy.

The overall architecture of the Tamil character recognition system developed in this study is shown in Fig. 2 below.

vowels	அ ஆ இ ஈ உ ஊ எ ஏ ஐ ஒ
consonants	க் ங் ச் ஞ் ட் ண் த் ந் ப் ம் ய்
Grantha	க்ஷ் ஷ்
Aytam	ஃ
others	ஸ்ரீ

Fig. 2. Modern Tamil characters

Pre-processing

Pre-processing covers all those functions carried out prior to feature extraction to produce a cleaned up version of the original image so that it can be used directly and efficiently by the feature extraction components of the Tamil character recognition. The steps in pre-processing involves

- Size normalization: Bi-cubic interpolation is used for standard sized image.
- Binarization: The process of converting a gray scale image into binary image by thresholding.

- Smoothing: The erosion and dilation smooth the boundaries of objects without significantly changing their area.
- Edge detection: Morphological gradient operators are used in edge detection because they enhance intensity of edges of characters.
- Segmentation: Horizontal histogram profile, vertical histogram profile and connected component analysis are able to handle the character segmentation problem.

IV. CLASSIFIERS

Seven different classifiers like projection distance, subspace method, linear discriminant function, modified quadratic discriminant function, Euclidean distance, modified projection distance, Mirror Image Learning and Linear Discriminant Function are used for comparative study. Both parametric and non-parametric classifiers are used for our experiment. Detail descriptions of these classifiers can be obtained in the literature and hence we are not giving their details here. However, some of the classifiers are briefly discussed as follows.

Euclidian Distance : The Euclidean distance between the input pattern and the mean vector is defined by

$$g_1^2(X) = \|X - M_l\|^2 \qquad (1)$$

where X is the input feature vector of size (dimensionality) n, Ml is the mean vector of class l. The input vector is classified to such class $l*$ that minimizes the Euclidean distance. Hereafter the subscript l denoting the class is omitted for the sake of simplicity.

Projection Distance : The projection distance is defined by

$$g_{pd}^2(X) = \|X - M\|^2 - \sum_{i=1}^{k} \{\phi_i^T (X - M)\}^2 \qquad (2)$$

and gives the distance from the input pattern X to the minimum mean square error hyper plane that approximates the distribution of the sample, where Öi denotes the ith eigenvector of the covariance matrix, and k is the dimensionality of the hyper plane as well as the number of the dominant eigen vectors ($k < n$). When $k = 0$ the projection distance reduces to the Euclidean distance.

Subspace method : For a bipolar distribution on a spherical surface with $\|X\| = 1$ the mean vector M is a zero vector ($M = 0$) because the distribution is symmetric in respect to the origin. Then the projection distance for the distribution is given by

$$g^2(X) = 1 - \Sigma_{i=1}^{k} \{\phi_i^T X\}^2 \qquad (3)$$

where ϕi is the ith eigen vector of the autocorrelation matrix. The second term of the above expression is used as the similarity measure of CLAFIC (Class Featuring Information Compression) and the subspace method.

Modified Quadratic Discriminant Function :

Modified quadratic discriminant function is defined as follows

$$g(X) = (N + N_0 + n - 1) \ln\left[1 + \frac{1}{N_0 \sigma^2}\left[\|X - M\|^2 \right.\right. \qquad (4)$$

$$\left.\left. - \sum_{i=1}^{k} \frac{\lambda_i}{\lambda_i + \frac{N_0}{N}\sigma^2} \{\phi_i^T (X - M)\}\right]\right] + \Sigma_{i=1}^{k} \ln\left(\lambda_i + \frac{N_0}{N}\sigma^2\right)$$

where X is the feature vector of an input character; M is a mean vector of samples; Ti is the ith eigen vector of the sample covariance matrix; $\lambda_i i$ is the i^{th} eigen value of the sample covariance matrix; k is the number of eigen values considered here, n is the feature size; σ^2 is the initial estimation of a variance; N is the number of learning samples.

Modified Projection Distance: The modified projection distance is defined by

$$g^2(X) = \|X - M\|^2 - \sum_{i=l}^{k} \frac{(1-\alpha)\lambda_i}{(1-\alpha)\lambda_1 + a\sigma^2} \{\phi_i^T (X - M)^2\} \qquad (5)$$

where α is a parameter which takes [0, 1]. When $\alpha = 0$, this classifier gives the same value as that of Projection Distance. When $\alpha = 1$, this gives the same value as that of Euclidian Distance. The value of α we used here is decided by preliminary experiment.

Linear Discriminant Function: Linear discriminant function is defined by

$$g(X) = W^T X + W_0 \qquad (6)$$

$$W = S_W^{-1} M$$

$$W_0 = -\frac{1}{2} M^T S_w^{-1} M \qquad (7)$$

where s_w is within-class covariance matrix.

Mirror Image Learning: Mirror Image Learning is a corrective learning algorithm proposed to improve the learning of class conditional distributions. The MIL generates a mirror image of a pattern which belongs to one of a pair of confusing classes to increases the size of the learning sample of the other class.

V. RESULTS AND DISCUSSION

This paper presents bag-of-key points approach to offline ancient Tamil character recognition. We have presented a simple but novel approach to character recognition using SIFT feature vector descriptors constructed from character image patches. This approach has been evaluated on a twenty different ancient Tamil characters image database. The bag-of-key points approach takes advantage of using histograms of number of occurrences of particular image patterns in an image. This approach is also invariant to transformation, rotation, and intra-class variations. The experimental results have demonstrated that the approach is able to improve the classification accuracy without imposing high computational cost. The testing results indicate that the proposed method is to be extended by considering the entire character classes as well as using a large database of ancient Tamil character samples. The approach can be extended to recognize other language scripts as well. The below fig. (3). Shows the Recognition of Ancient Tamil Characters from stone inscriptions.

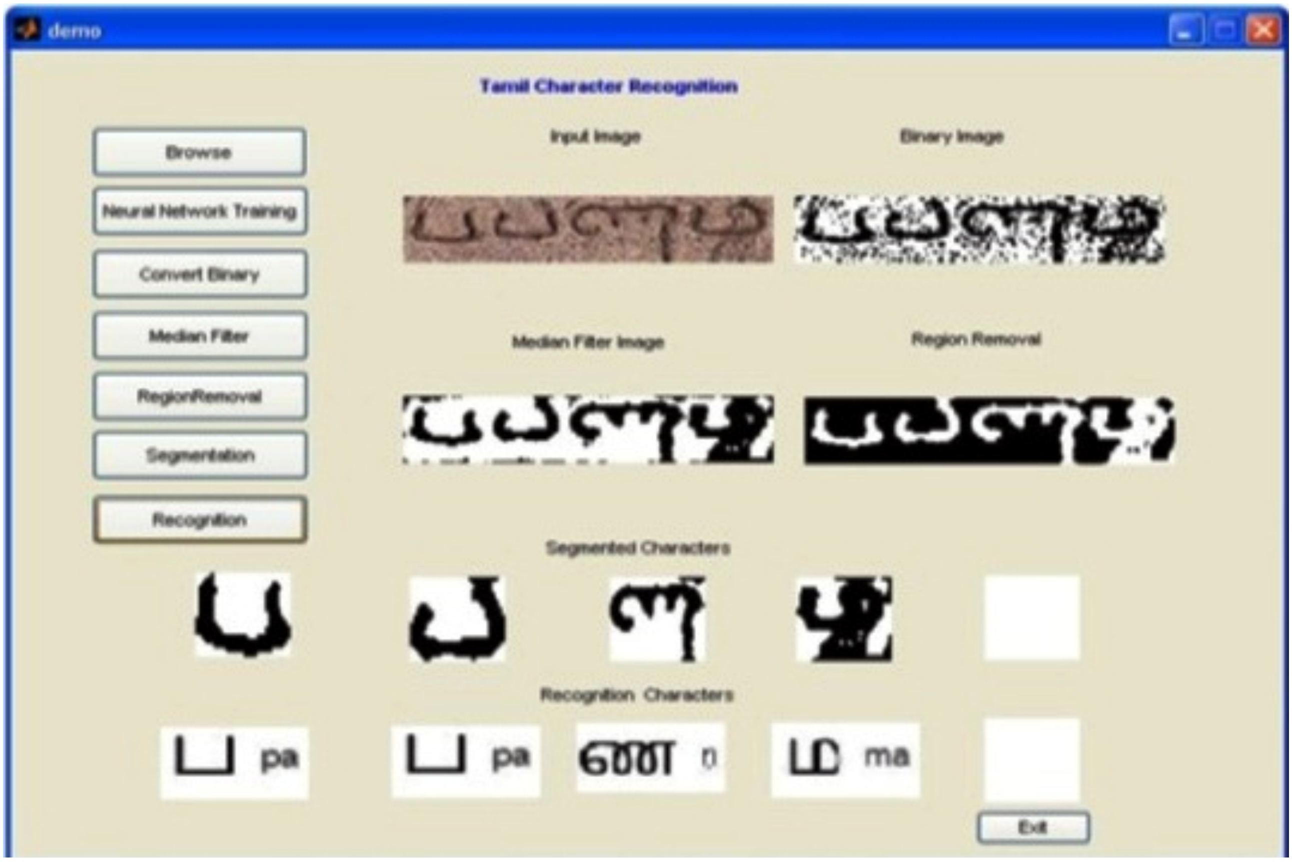

Fig. 3. Recognized Image

VI. CONCLUSION

The Recognized Ancient Tamil Characters are taken from 11th century stone inscriptions. To extract the characters or components containing strokes in vertical, horizontal, right diagonal and left diagonal directions, we have performed the erosion operation on the input binary image with the line structuring element. The length of the structuring element is thresholded to 70% of distance between mean line and base line of the word image obtained by horizontal projection profile. The Recognized Ancient Tamil Character image is illustrated in Fig. 3. The proposed algorithm is inspired by a simple observation that every script or language defines a finite set of text patterns, each having a distinct visual appearance, and hence every character could be identified based on its discriminating features.

REFERENCES

[1] R. Plamondon and S.N. Srihari, "On-Line and off-line handwritten recognition: A comprehensive survey", IEEE Trans on PAMI, Vol.22, pp. 62-84, 2000.

[2] U. Pal and B.B. Chaudhuri, "Indian script character recognition: A Survey", Pattern Recognition, Vol. 37, pp. 1887-1899, 2004.

[3] C.L. Liu and C.Y. Suen, "A new benchmark on the recognition of handwritten Bangla and Farshi numeral characters", In Proc. 11th ICFHR, 2008.

[4] I.K. Sethi and B. Chatterjee, "Machine Recognition of constrained Hand printed Devnagari", Pattern Recognition, Vol. 9, pp. 69-75, 1977.

[5] U. Pal, T. Wakabayashi, N. Sharma and F. Kimura, "Handwritten Numeral Recognition of Six Popular Indian Scripts", In Proc. 9th ICDAR, pp. 749-753, 2007.

[6] M. Hanmandlu and O.V.R. Murthy, "Fuzzy Model Based Recognition of Handwritten Hindi Numerals", In Proc. Intl. Conf. on Cognition and Recognition, pp. 490-496, 2005.

[7] S. Kumar and C. Singh, "A Study of Zernike Moments and its use in Devnagari Handwritten Character Recognition", In Proc. Intl. Conf. on Cognition and Recognition, pp. 514-520,2005.

[8] N. Sharma, U. Pal, F. Kimura and S. Pal, "Recognition of Offline Handwritten Devnagari Characters using Quadratic Classifier", In Proc. Indian Conference on Computer Vision Graphics and Image Processing, pp-805-816, 2006.

[9] U. Pal, N. Sharma, T. Wakabayashi and F. Kimura, "Off- Line Handwritten Character Recognition of Devnagari Script", In Proc. 9th ICDAR, pp. 496-500, 2007.

[10] N. Otsu, "A Threshold selection method from grey level histogram", IEEE Trans on SMC, Vol.9, pp. 62-66, 1979.

[11] F. Kimura, T. Wakabayashi, S. Tsuruoka and Y. Miyake, "Improvement of handwritten Japanese character recognition using weighted direction code histogram", Pattern Recognition, Vol.30, No.8, pp.1329-1337, 1997.

[12] M. Shi, Y. Fujisawa, T. Wakabayashi, and F. Kimura, "Handwritten numeral recognition using gradient and curvature of gray scale images", Pattern Recognition, Vol. 35, pp.2051-2059, 2000.

[13] F. Kimura, K. Takashina, S. Tsuruoka and Y. Miyake, "Modified quadratic discriminant function and the application to Chinese character recognition", IEEE Trans. on PAMI, Vol. 9, pp 149-153, 1987.

[14] T. Wakabayashi, M. Shi, W. Ohyama, and F. Kimura: "A Comparative Study on Mirror Image Learning and ALSM":In Proc. 8th IWFHR, pp. 151-156, 2002

[15] Tong, T., and Koller, D., "Support vector machine active learning with applications to text classification", In proceedings of the Seventeenth International Conference on Machine Learning (ICML), 2000