Asian Research Consortium

Asian Journal of Research in Social Sciences and Humanities
Vol. 6, No. 5, May 2016, pp. 604-615.

ISSN 2249-7315
A Journal Indexed in Indian Citation Index
DOI NUMBER:10.5958/2249-7315.2016.00139.8

Asian Journal of Research in Social Sciences and Humanities

www.aijsh.com

Recognition of Ancient Stone Inscription Characters using Normalized Positional Distance Metric Features

G. Bhuvaneswari*; Dr. V.Subbiah Bharathi**

*Research Scholar,

Department of CSE,

Anna University,

Chennai, India.

**Professor,

Easwari Engineering College

Chennai, India.

Abstract

Recognition of inscribed characters on stone is an essential part of our work to reveal information's of our ancestors. The concept of Optical Character Recognition is not adopted as the rock surface becomes texture of an image posing challenges in recovering the text. In this paper, we proposed a new feature called Positional Distance Metric which was independent to any language script to address the various issues occurred on stone inscribed images. Then Normalized Positional Distance Metric (NPDM) feature is computed and combined with structural and regional features to yield a better recognition rate. We repeated this procedure for all the training images and finally populated in to the list. Nearest Neighbor classifier is used for subsequent classification and recognition of characters. Experiments are performed on 350 characters and showed significant recognition rate.

Keywords: feature extraction, character recognition, Classifiers, thinning, Image processing, Indian Script.

I. Introduction

Ancient Character Recognition (ACR) System is very helpful to Archeological Department that translates ancient letters into current modern characters. The technical challenges in ancient character recognition arise from two sources. First is an image defect: imperfections in stone image

due to natural climates like wind, rain, lighting and thunder. Second is deformation: cracks and dents on rock surface treated as part of character and time dependent distortion. In this paperan Ancient Tamil Stone Inscription Characters have been considered for experimental analysis. Tamil is one of the oldest southern languages of India. The evolution of Tamil started from 3rd century BC. The beginning of evolution of Tamil comprised the period between the 3rd century BC and the 6th century AD, Medieval Tamil existed between the 6th century AD and the 12th century AD, and Modern Tamil, from the 12th century down to the present day. The ancient Tamil characters i.e. the early and medieval Tamil can be mostly found in stone inscriptions and palm leaves. Only epigraphists can read those stone inscriptions. To extend the readability and to preserve the ancient historical values, we need a good recognition system that can convert the ancient text to modern text.

While most work has been published for printed and handwritten Tamil text, very little is reported for ancient stone inscription script. One of the first attempts for printed characters has been by Siromoney et al. [1] which described a method of encoded character string extracted by row and column-wise scanning of character matrix and compared with the strings in the dictionary. Chinnuswamy et al. [2] proposed an approach that uses topological matching procedure to compute the correlation coefficients and then maximizes the correlation coefficient. Suresh et al. [3] described the fuzzy concept on handwritten Tamil characters to classify characters using a feature called distance from the frame and a suitable membership function. This algorithm obtained success rate varies from 76% to 94%.

Hewavitharana, S, and H.C. Fernando [4] described a system that uses both structural and statistical techniques to recognize handwritten Tamil characters using a two-stage classification approach. In [5] they described a recognition system for offline handwritten Tamil characters where pixel densities are calculated for different zones of the image and these values are used as the features of a character. These features are used to train and test the support vector machine. Shivsubramani et al. [6] showed method for recognizing printed Tamil characters exploring the interclass relationship between them, which should be accomplished using Multiclass Hierarchical Support Vector Machines. Szedmak et al.,[7] used a Multiclass Hierarchical SVM algorithm that provided the accuracy 96.85% compared with many commonly used classifiers.

In [8] they proposed non-text block classification method for obtaining 94% recognition rate when the text block having a few touching characters. In [9] they developed elastic matching scheme for writer dependent on-line handwriting recognition of isolated Tamil characters. In [10] the system extracted structural features such as character height, width, number of horizontal lines from image glyphs and mapped onto Unicode for recognition.

In [11] a generalized framework was proposed for Indic script character recognition. Unique strokes in the script were identified which were then compared with the database using the proposed flexible string-matching algorithm producing 86.1% performance.

In [12, 13] a subspace-based method using Principal Component Analysis (PCA) was applied for Tamil character recognition. The paper [13] analyzed the performance of DTW and PCA on the following three cases such as writer independent, writer dependent and writer adaptive.DTW worked better than PCA in all above three cases. In [14], features like Angle features, Fourier

coefficients and Wavelet features were used for recognition using a Neural Network classifier. In [15] Data-driven HMM-based online handwritten word recognition system was proposed for Tamil. An ancient Tamil character recognition algorithm based on artificial immune was proposed to improve the rate of character recognition than the classification done by neural network[16]. In paper [17], template matching method was proposed for recognition of Stone Inscripted Kannada Characters especially Hoysala and Ganga timeframes.

However, these feature extraction methods were not given desirable results in stone inscription image since rock surface becomes texture of an image and rock surface is not having uniform plane. It includes various dents and cracks on it that may lead misclassification.

This paper describes a system for recognizing an ancient characters and convert into modern characters. The processing steps of our ACR system are represented in Fig. 1. The stone inscribed images are first subjected to various preprocessing techniques like filtering, binarization, normalization and thinning that are discussed in chapter 2. Then the feature extraction module extracts the various features like structural, regional and new proposed feature called Normalized Positional Distance Metric (NPDM). It is described in chapter 3. The final classification is achieved using Nearest Neighbor algorithm that is given in chapter 4 and the experimental results are discussed in chapter 5. The applications of this system are many as in archeological department and unconstrained ancient script recognition. This paper is concluded in chapter 6.

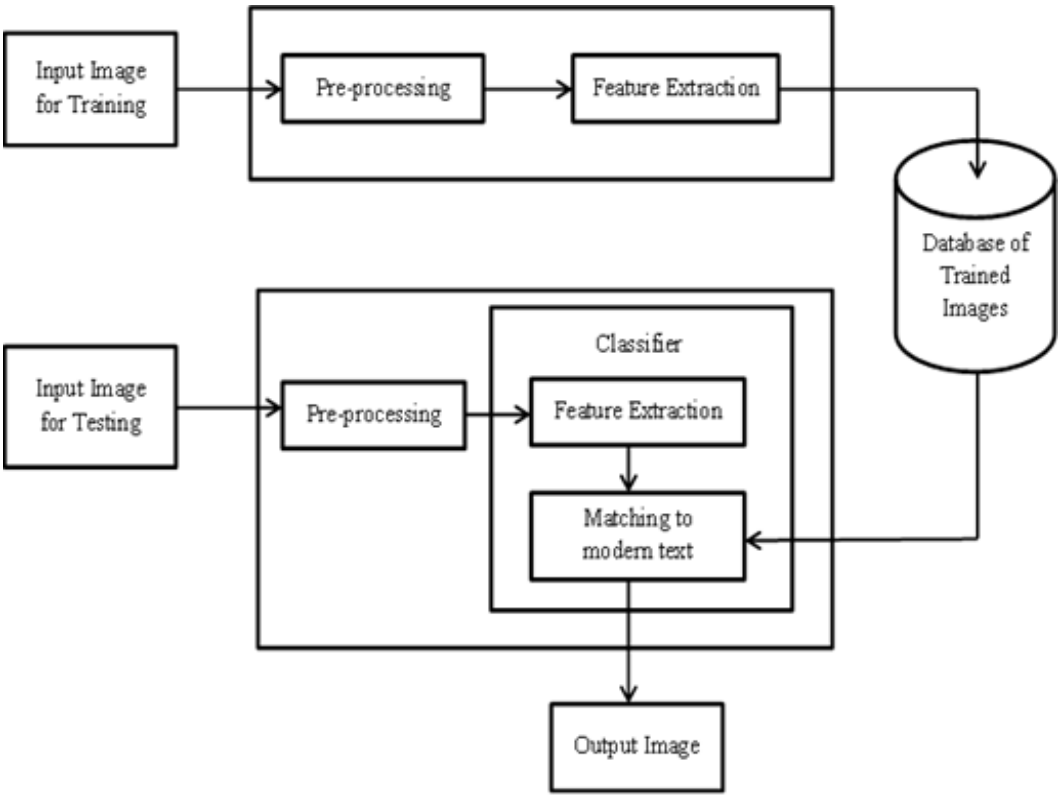


Figure: 1 Ancient Character Recognition (ACR) system

II. Preprocessing

Once the stone inscription image has been acquired, the input image has to undergo various preprocessing steps. This step is required to increase the recognition rate by eliminating abnormalities in the image. The image used for testing was acquired directly from Thanjavur, Brihadeeswarar Temple using ordinary digital camera. This temple inscriptions are belonging to 11^{th} century built by Emperor Rajendra Chola I. One of the sample inscription is shown in figure 2. It was undergone into various preprocessing steps.

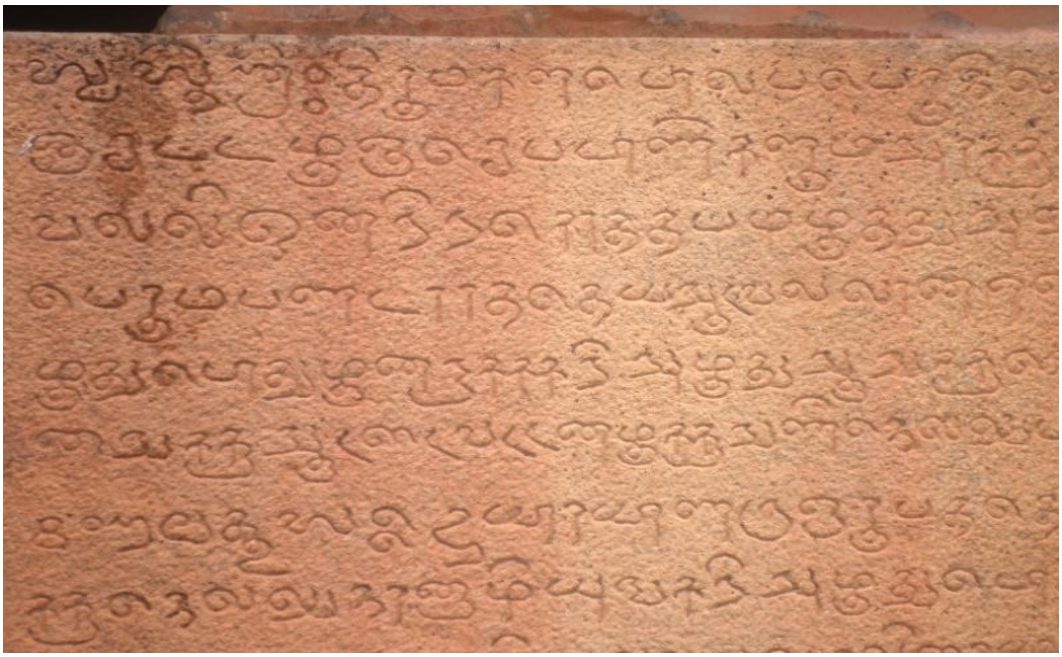


Figure 2. Sample Stone Inscription

The preprocessing steps are explained as follows.

2.1 Noise Removal

Noise can be removed from the image using the median filter of size 5X5. The median filter replaces the pixel value with the median of neighboring pixel values. The median is calculated by first sorting all the pixel values from the surrounding neighborhood into numerical order and then replacing the pixel being considered with the middle pixel value.

2.2 Binarization

It converts the input image into binary image containing only 0's and 1's to represent the image details. If the input image is RGB image, then it converts into grayscale image that in turn convert into binary image by thresholding [18]. The output image replaces all pixels in the input image with luminance which is greater than the specified level value by the value 1 (white) and replaces all other pixels with the value 0 (black).

$$b(x,y) = \begin{cases} 1, \text{if } s(x,y) > th \\ 0, \text{otherwise} \end{cases}$$

$0 \leq x < R,\ 0 \leq y < C$

whereb(x,y) - binary image

s(x,y) – Input image

th - threshold (luminance) value

R,C - dimension of an image

2.3 Normalization

In size normalization, the images of different sizes are normalized to constant dimensions by keeping the aspect ratio of the images. The height and width ratio of original pattern is retained using a bilinear interpolation algorithm.

2.4 Image Thinning

It is the process of reducing thickness of the image into one pixel width images. The morphological thinning [19] is one of the thinning algorithms that is used in this paper. Then the features of thinned image are extracted. The output of each step is shown in fig 3.

Figure 3. a) Original Image b) Filtered Image c) Binarized Image d) Thinned Image

III Feature Extraction

Feature is a unique characteristics used to identify the character efficiently. This is the main module where the feature vectors are created for recognition. A feature called Positional Distance Metric is proposed and is based on the spatial properties of a character. The important aspect of recognition system is selection of good feature set which is invariant with respect to shape variations due to various styles and pressure applied while carving the stone. The major strength of this approach is its robustness to small variation and provides high recognition rate. In this section, we explain a new feature that will helpful in good classification and recognition of the character.

The position of first pixel in the binary thinned image is computed and the average distance between that pixel and all the remaining pixels in the image that comprises the structure of the

characters are calculated. This feature is extracted from the selected character in two ways such as Image based and Zone based. We have repeated this procedure for some set of training characters and collected as trained feature set. For classification and recognition, nearest neighbor classifier is used.

Proposed Algorithm 1: Image based NPDM

Input: Thinned Image

Output: Feature set

% Algorithm for INPDM %

Start

Find position of first pixel

Store x value in $left$ and y value in $upper$

Repeat

Calculate Euclidean Distance between current pixel and a first pixelp(left, upper)

Sum them as sum_Dist

Until end of pixel

Find the average distance Avg_Dist for n X m character

Store Avg_Dist

Stop

Proposed Algorithm 2: Zone based NPDM

Input: Thinned Image

Output: Feature set

% Algorithm for ZNPDM %

Start

Split character into four equal zones

If split is not possible

Pad with trailing zeroes and bottom zeroes

End if

On each zone, do the following

{

Find position of first pixel

Store x value in $left$ and y value in $upper$

Repeat

Calculate Euclidean Distance between current pixel and a first pixelp(left, upper) of zone

Sum them as sum_Dist

Until end of pixel of zone

Find the average distance Avg_Dist for n X m zone

Store Avg_Dist in vector

}

Stop

These two new features can also be worked with structural features and regional features in order to bring more accuracy. The structural features used here are number of horizontal lines, the number of vertical lines and the number of intersection points and regional features are orientation, convex area, eccentricity and extent. These features will form prototype vector that will be used in classification.

IV Classification & Recognition

In this paper, the one among the supervised statistical pattern recognition called nearest neighbor classifier is considered for classification. It is based on matching that represents each class label by a prototype pattern vector. It computes the Euclidean distance between the unknown and each of the prototype vectors. It selects the smallest distance to make a decision. The Euclidean distance between training set and testing set can be computed as follows.

$$\mathrm{d}(\mathrm{xtrain}, \mathrm{xtest}) = \sqrt{(\mathrm{xtrain} - \mathrm{xtest})^2}$$

The training phase of the algorithm consists of collecting the feature vectors of the training images and class labels of the training samples. In the actual classification phase, the same features as before are computed for the test samples. Distances from the new vector to all stored vectors are computed. Then classification and recognition is achieved on the basis of similarity measurement.

V Experimentation and Results

Due to lack of data source, we have collected 35 samples for each 10 characters.60% of the database is used for feature extraction. Feature vectors of 3 prototypes (INPDM, ZNPDM, Structural + Regional) of each character are stored. The remaining database is used for validation. The recognition rates of these 3 prototypes for 10 characters are shown in Table I. As observed that both INPDM and ZNPDM produced better accuracy rate than structural + regional features.

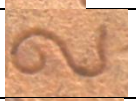
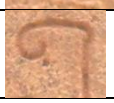
Table 1 Recognition Rates of existing and Proposed Features

Ancient Stone Inscribed Characters	Recognized Modern Characters	Nearest Neighbor Classifier		
		Existing Features (Structural+Regional)	Proposed Features	
			INPDM	ZNPDM
	க (Ka)	74.28	77.14	80.0
	ர (Ra)	77.14	82.86	82.86
	ம (Ma)	80.0	85.71	88.57
	ய (Ya)	77.14	77.14	80.0
	வ (Va)	74.28	80.0	77.14
	ண (Na)	71.43	74.28	80.0
	த (Tha)	68.57	71.43	77.14
	ல (La)	77.14	82.86	80.0
	ப (Pa)	80.0	85.71	88.57
	எ (Ea)	71.43	77.14	82.86
	Average Recognition	75.14	79.4	81.7

The result of Algorithm Fusion is given in Table II. It showed that the average recognition rate of about 84.8 % was achieved when the fusion was taken place between ZNPDM and structural + regional features. The output of ACR system is shown in figure 4 for character Na (2 – loops). The time complexity of the proposed method is O (n^2). This system rejects 1.5 % of the characters that

are broken or do not match with any of the stored features of characters. Errors are occurred when the character is not accurately cropped. High rate of error is observed in some characters that are written in structurally different ways. Additional models have to be considered to recognize them.

Table 2 Recognition Rates of Algorithm Fusion

Ancient Stone Inscribed Characters	**Recognized Modern Characters**	**Nearest Neighbor Classifier**	
		INPDM + (Structural+Regional)	**ZNPDM + (Structural+Regional)**
	க (Ka)	82.86	85.71
	ர (Ra)	80.0	82.86
	ம (Ma)	82.86	88.57
	ய (Ya)	82.86	85.71
	வ (Va)	80.0	82.86
	ண (Na)	77.14	80.0
	த (Tha)	82.86	80.0
	ல (La)	80.0	82.86
	ப (Pa)	82.86	91.43
	எ (Ea)	85.71	88.57
	Average Recognition	81.7	84.8

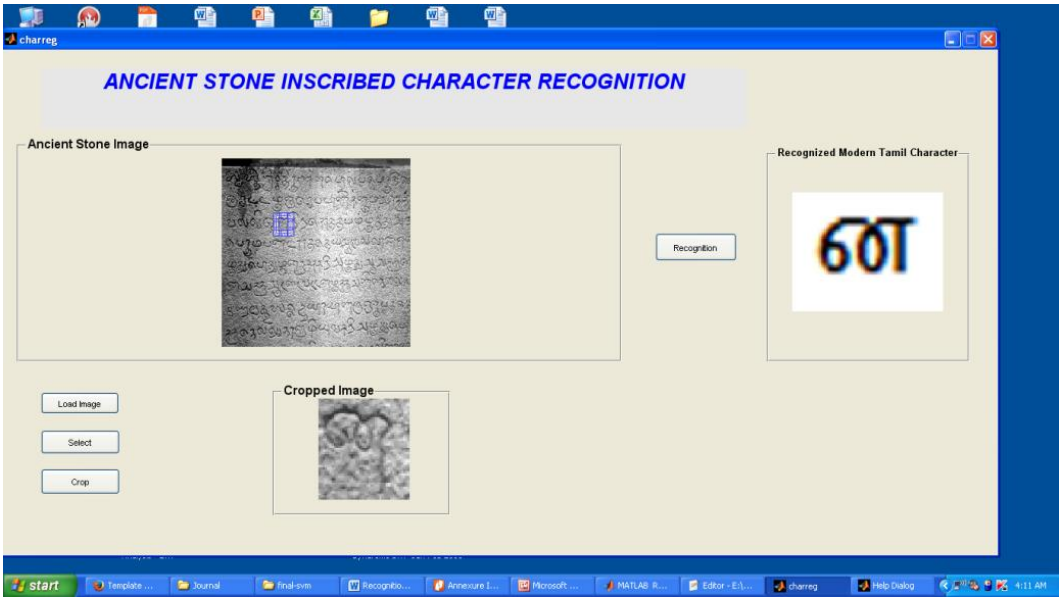


Figure 4.Recognition of Ancient Tamil Character Na (2-loops)

VI Conclusion

The proposed work presents new feature for recognition of ancient stone inscribed character that is independent to any regional languages. Database has been collected from various writers / sources of same century. The proposed ZNPDM algorithm though simple, works efficiently for stone inscribed letters when combined with structural and regional properties of characters. The overall recognition accuracy obtained is 84.8%. This work is a step towards unconstrained ancient script recognition. Further it can be extended by introducing the process called contextual recognition and developing an Android application to recognize the characters on-site. This Ancient Character Recognition system can be used in the field of epigraphy.

References

Siromoney et al., "Computer Recognition of Printed Tamil Character", Pattern Recognition, 1978, pg no: 243-247.

Chinnuswamy, P., and S.G. Krishnamoorthy, "Recognition of Hand printed Tamil Characters", 1980, Pattern Recognition, 12: 141-152.

Suresh et al., "Recognition of Hand printed Tamil Characters Using Classification Approach",1999, ICAPRDT' 99, pp: 63-84.

Hewavitharana, S, and H.C. Fernando, "A Two-Stage Classification Approach to Tamil Handwriting Recognition", pp: 118-124, Tamil Internet 2002, California, USA.

N. Shanthi and K. Duraiswamy, "Performance Comparison of Different Image Sizes for Recognizing Unconstrained Handwritten Tamil Characters using SVM",Journal of Computer Science,Vol 3, Issue 9, Pages 760-764, 2007.

Shivsubramani K, Loganathan R, Srinivasan CJ, Ajay V, Soman KP, "Multiclass Hierarchical SVM for Recognition of Printed Tamil Characters",In:Proc.of IJCAI, 2007.

Szedmak ,SandorSzedmak, John Shawe- Taylor, "Learning Hierarchies at Two-class Complexity", Kernel Methods and Structureddomains,NIPS 2005.

K.H.Aparna, SumanthJaganathan, P.Krishnan, V.S.Chakravarthy, "Document Image Analysis: with specific Application to Tamil Newsprint",International Conference on Universal Knowledge and Language (ICUKL), Goa, India, Nov. 2002.

N. Joshi, G. Sita, A. G. Ramakrishnan, and S. Madhvanath," Comparison of Elastic MatchingAlgorithms for Online Tamil Handwritten Character Recognition" Proceedings of the 9th International Workshop on Frontiers in Handwriting Recognition, 2004.

Seethalakshmi R., Sreeranjani T.R., Balachandar T., AbnikantSingh,Markandey Singh, RitwajRatan, Sarvesh Kumar, "Optical Character Recognition for printedTamil text using Unicode", Journal of Zhejiang University SCIENCE, Vol. 6A No. 11, 2005.

H. Aparna, V. Subramanian, Kasirajan, V. Prakash, V. Chakravarthy, and S. Madhvanath, "Online Handwriting Recognition for Tamil", Proceedings of the 9th International Workshop onFrontiers in Handwriting Recognition, 2004.

V. Deepu and S. Madhvanath,"Principal Component Analysis for Online Handwritten Character Recognition", Proceedings of the 17th International Conference on Pattern Recognition, 2004.

N. Joshi, G. Sita, A. G. Ramakrishnan, and S. Madhvanath,"Tamil Handwriting RecognitionUsing Subspace and DTW Based Classifiers", Proceedings of the 11th International Conferenceon Neural Information Processing, 2004.

C. S. Sundaresan and S. S. Keerthi, "A Study of Representations for Pen based HandwritingRecognition of Tamil Characters", Proceedings of the 5th International Conference onDocument Analysis and Recognition, 1999.

Bharath A, SriganeshMadhvanath, "Hidden Markov Models for Online Handwritten Tamil Word Recognition." HP Laboratories India, HPL-2007-108, July 6, 2007

RajaKumar S, SubbiahBharathi V, "Eighth century tamil consonants recognition from stone inscriptions" Proceedings of the International conference on Recent Trends In Information Technology,2012.

Rajithkumar B K and H.S. Mohana, "Template Matching Method for Recognition of Stone Inscripted Kannada

Characters of Different Time Frames Based on Correlation Analysis", International Journal of Electrical and Computer Engineering, Vol. 4, No. 5, October 2014, pp. 719-729.

N.Otsu, "A threshold selection method from gray level histograms ", IEEE Transactions on systems, Man and Cybernetics , 9(1), pp.62-66, 1979.

L.Lam, S.Lee, and C.Suen, "Thinning Methodologies – A Comprehensive Survey", IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 17, no. 9, pp. 914-919, 1995.