ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 9, September 2014

An Efficient Period Prediction System for Tamil Epigraphical Scripts Using Transductive Support Vector Machine

S Venkata Krishna Kumar[1] Poornima T V[2]

Associate Professor, Department of Computer Science, PSG College of Arts and Science, Coimbatore, India[1]

Research Scholar, Department of Computer Science, PSG College of Arts and Science, Coimbatore, India[2]

Abstract: Tamil is one of the ancient languages in the world with rich in literature. The writers used various materials like stone, metal, pottery, wood, palm leaves, cloth, conch shell, mural paintings and copper plates to encrypt their writing. The information gathered from these inscriptions gives us knowledge about the astronomy, history, culture, religious, economic tax, administrative and educational conditions. These epigraphical inscriptions plays an important role in knowing the civilized past and classification of characters belonging to various periods. Therefore a system is proposed to read the ancient Tamil characters belonging to various periods by testing a small amount of characters referred to as examined characters in Tamil language. These examined characters are taken from the script automatically and coordinate with the characters belonging to different periods using machine intelligence. Hence the proposed system consists of various modules like image acquisition, binarization, preprocessing, feature extraction, segmentation, and at last classification and prediction of period using Transductive Support Vector Machine (TSVM).The experimental results shows higher accuracy when compared with Support Vector Machine (SVM)

Keywords: Epigraphical Scripts, Tamil Language, TSVM, Feature Extraction

I. INTRODUCTION

Tamil, the native language of a southern state in India has several million speakers across the world and is an official language in countries such as Srilanka, Malaysia & Singapore. Tamil has 12 vowels and 18 consonants. These are combined with each other to yield 216 composite characters and 1 special character (aayatha ezhuthu) counting to a total of (12+18+ 216+1) 247 characters. Vowels in Tamil are of two types such as short (kuril) and long (Nedil) and it is also called as UyirEzhuthu. Consonants are of three types such as Vallinam, Idaiyinam, and Mellinam.

1.1 TAMIL INSCRIPTIONS

There are inscriptions of many ages, carved in Tamil-Brahmi, Vattelettu, Tamil and Grantha scripts. The inscriptions in Tamil Nadu can be divided either chronologically or typologically into different groups. They can be broadly classified into three groups, known as (1) Tamil-Brahmi inscriptions (TBI), (2) Hero Stone Inscriptions (HSI) and (3) Temple Tamil Inscriptions (TTI). The vast corpus of Tamil inscriptions shows the gradual development of the Tamil epigraphic culture. The earliest of the Tamil inscriptions written in Tamil-Brahmi script and so named as Tamil-Brahmi inscriptions (TBI), date from 300 BCE up to 500 CE. These short inscriptions carved on the rocks and in the natural caves, on pot shreds and coins have been studied in detail by Mahadevan (2003). The HSI comes immediately after the (TBI) both chronologically and typologically. The HSI, short memorial or funerary texts, are limited in number and are distributed mainly in the northern part of Tamil Nadu. Finally, the TTI are the structurally developed diverse monumental inscriptions carved on the walls of the temples and span from 5th century up to 19th CE. The TTI inscriptions contain mostly descriptions of donations (of lands, ornaments, jewellery, cows, goats, statues and images) made to temples, village assemblies, their maintenance, administration. From the expansion of Chola (8th to 14th CE), epigraphy flourished everywhere, and inscriptions in Tamil are literally innumerable. So it's possible only for the epigraphists to read those various inscriptions because the characters used to write on ancient centuries are differ when compared to today's century. To extend the read-ability and to preserve the ancient historical values, we need a good recognition system. So, I have used image processing techniques to recognize the characters efficiently. In this research, an experiment is concerned for period prediction of Tamil Epigraphical scripts using TSVM technique. In the following discussion, section 2 gives Related Works, section 3 is about Proposed System, section 4 discuss about the Preprocessing techniques which are adopted in the system, section 5 deals with Segmentation of line and character, section 6 discuss about the Feature extraction, section 7 deals with the Classification section 8 and 9 the performance of the proposed system with experimental results and conclusions are presented respectively.

II RELATED WORKS

Bhattacharya et al. [1] proposed a two stage approach. In the first stage an unsupervised clustering method was applied to create a minimum amount of groups of hand written Tamil character classes.In the second stage a supervised classification technique was considered in each of these small groups for final recognition. The number of transistion and the chain code histogram are the features

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 9, September 2014

used in the first and second stages respectively. 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. Jagadeesh Kannan et al [26] 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. All these works mainly focus on recognizing the Tamil characters by using different classification.But the proposed work recognize the ancient Tamil characters and predict the period by using TSVM to get a higher accuracy.

III PROPOSED SYSTEM

The Proposed system is designed to ease the manual barrier by helping the computer to understand human handwritten characters through an automated system. The design mainly aims in implementation of character period prediction system in which computer will be able to understand a few simple commands and identify the century of these characters.

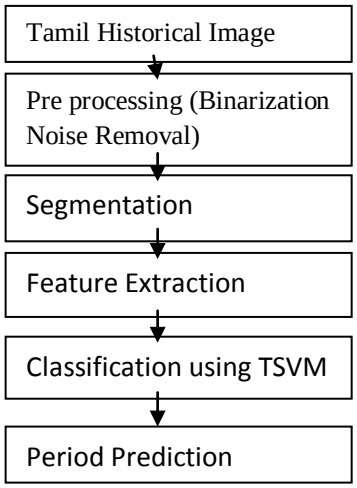

Fig 1 Overview of the Proposed System

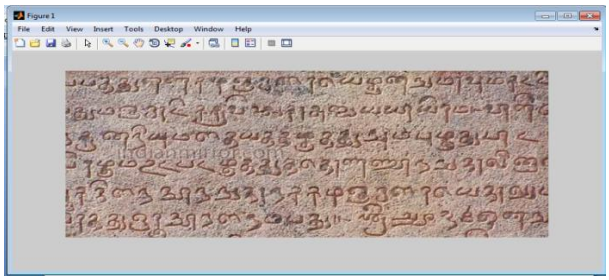

Fig 2 Tamil Historical Image

IV PREPROCESSING

There are numerous tasks to be completed before performing character recognition. A handwritten image must be scanned and converted into a suitable format for processing. Preprocessing consists of a few types of sub processes to dicepher the image and makes it appropriate to carry the recognition process accurately. The sub processes which get involved in pre-processing are binarization and noise removal.

4.1 Binarization

Binarization is the process of converting a gray scale image (0 to 255 pixel values) into binary image (0 and 1 pixel values) by selecting a global threshold that separates the foreground from background. Each pixel is compared with the threshold and if it is greater than the threshold it is made 1 or else 0. This can be done by using Otsu's method. The process is estimated using the following equation

$$\sigma_w^2(t) = \omega_1(t)\,\sigma_1^2(t) + \omega_2(t)\,\sigma_2^2(t)$$

Weights ω_i are the probabilities of the two classes separated by a threshold t and σ_i^2 variances of these classes.

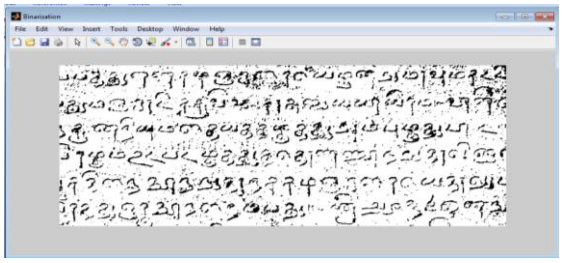

Fig 3 Binarized Image

4.2 Noise Removal

The input image of the Tamil historical inscriptions may be degraded due to the presence of the broken characters, erased characters, touching characters, distortion due to fossils settled, irrelevant symbols engraved by the scribes and so on. The non uniform spacing between the lines and characters of epigraphical images and the skew could complicate the process of deciphering the script. Hence, Median filter technique is adopted for removing noise from the scripted image. The process is estimated using the following eqation.

$$W_t = \begin{cases} 1 & if\ i = \frac{\Omega-1}{2} \\ 0 & otherwise \end{cases}$$

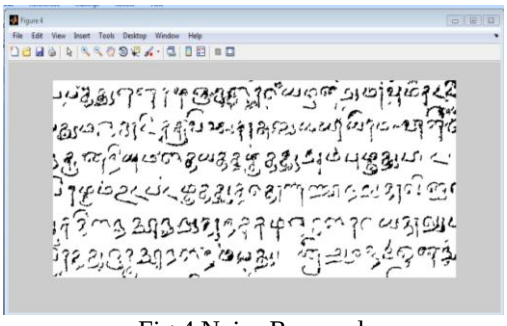

Fig 4 Noise Removal

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 9, September 2014

V SEGMENTATION

After pre-processing, the noise free image is passed to the segmentation phase, where the image is decomposed into individual characters.

Algorithm for segmentation:

(1) The binarized image is checked for inter line spaces using horizontal and vertical projection technique.

(2) If inter line spaces are detected then the image is segmented into sets of paragraphs across the interline gap.

(3) The lines in the paragraphs are scanned for horizontal space intersection with respect to the background. Histogram of the image is used to detect the width of the horizontal lines. Then the lines are scanned vertically for vertical space intersection. Here histograms are used to detect the width of the words. Then the words are decomposed into characters using character width computation.

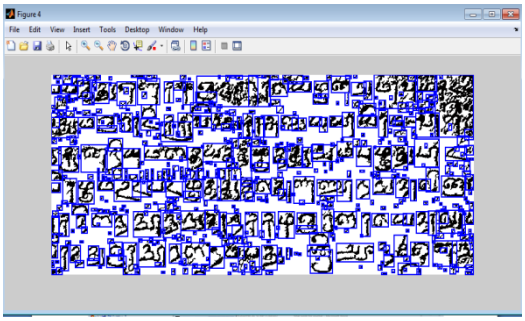

Fig 5 Segmented Image

VI FEATURE EXTRACTION

Feature extraction phase extracts the basic components of Tamil characters, such as Height, Width, Horizontal Projection, Vertical Projection, Horizontal Center, Vertical center, Horizontal Projection Skewness, vertical Projection Skewness, HCurves, VCurves, number of circles, number of slope lines and branching points. Each feature plays an important role in pattern recognition. So in order to extract the features from the segmented Tamil character images, first scale the image into common height and width by using bilinear interpolation technique. Hence each image is divided into equal number of horizontal and vertical strips. This linear interpolation technique can done first in x direction and then again in the y direction.Linear interpolation in the x- direction is calculated using the equation

$$F(R_1) \approx \frac{x_2 - x}{x_2 - x_1} f(Q_{11}) + \frac{x - x_1}{x_2 - x_1} f(Q_{21})$$

Linear interpolation in the *y*-direction is calculated using equ

$$F(P) \approx \frac{y_2 - y}{y_2 - y_1} f(R_1) + \frac{y - y_1}{y_2 - y_1} f(R_2).$$

Horizontal centerline: Horizontal centerline is calculated based on scanning the character form left to right of the whole character .Let Horizontal center is calculated based on the position of the horizontal centerline as follows

$$Hf1 = \sum_{i=2}^{k} |h(i) - h(i-1)|$$

Vertical center: For exploiting the information coming from the detection of the vertical centerline of the character, generates a new group of the features. To create the second group of the feature vector, the letter image is then scanned from top to bottom with a sliding window.The equation to calculate Vertical center as follows

$$Vf1 = \sum_{i=2}^{k} |v(i) - v(i-1)|$$

Thus the Feature extraction describes the relevant shape information contained in a pattern so that the task of classifying the pattern is made easy.

VII CLASSIFICATION

Classification is done by using the features extracted in the previous step, which corresponds to the each character attribute. Here the system uses a Transductive Support Vector Machine (TSVM) for classification. TSVM is a set of related semi supervised learning method used for classification. It is based on the hyper planes that maximize the separating margin between two classes using the available labeled samples. However, in many real-life applications, obtaining labeled patterns is expensive, while large unlabeled samples are readily available. Since the unlabeled patterns are significantly easier to obtain than labeled ones, TSVMs were proposed are iterative algorithms that gradually search the optimal separating hyper plane in the feature space with a transductive process that incorporates unlabeled samples in the training phase. TSVM uses maximizing separation between labeled and unlabeled data, the equation to solve

$$\min_{y_j, f \in \mathcal{F}} C_1 \sum_{i=1}^{n_l} L(y_i f(x_i)) + C_2 \sum_{j=n_l+1}^{n} L(y_j f(x_j)) + J(f),$$

Where f is a decision function in F, a candidate function class L (z) = $(1\text{-}z)_+$ is the hinge loss, and J(f) is the inverse of the geometric separation margin. In the linear case, $F(x) = w^T x + b$ and $J(f) = \frac{1}{2}||w||^2$ .In the nonlinear kernel case $f(x) = (K(x,x_1), K(x,x_n))w^T + b$,

Minimization of the condition with respect to f ∈ F is non convex, which can be solved through integer programming, and is known to be NP

VIII PERFORMANCE OF THE SYSTEM

Comparision Table to predict the period for SVM and TSVM

Parameters	Period Prediction Using SVM	Period Prediction Using TSVM
Accuracy	88.7640	94.6667
Precision	0.8919	0.9457
Recall	0.8848	0.9457

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 9, September 2014

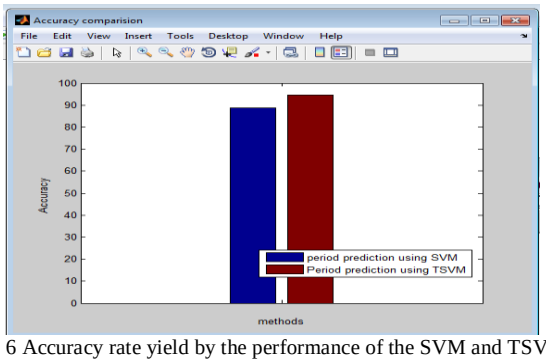

Fig 6 Accuracy rate yield by the performance of the SVM and TSVM

IX CONCLUSION

The period prediction of epigraphical script is the area of research where it is possible for the user to the computer and has it understand or recognize the character. A calculative approach is proposed here for predictiving the Tamil Scripts using Transductive Support Vector Machine (TSVM).To perform this, proposed methodology uses many tests like finding whether the character is on which century. This research works gives overall accuracy of 94.66% when compared with SVM.

REFERENCES

1. Bhattacharya U., Ghosh SK., Parui S.K., "A two stage recognition scheme for handwritten Tamil characters" In: Proceedings of the ninth international conference on document analysis and recognition (ICDAR 2007). IEEE Computer Society, Washington, DC, pp 511–515
2. Indra Gandhi R.,Iyakutti K.,"An Attempt to recognize Handwritten Tamil Character Using Kohonen SOM,"IJANA,Vol. 01 No.3,pp 188-192,Mar 2009.
3. Rajkumar S., Subbiah Bharathi V., "Ancient Tamil Character Recognition from Tamil Wall Inscriptions," IJCSE, Vol. 3 No.5, pp 673-677, Nov 2012.
4. Sowmya A., Hemanathan Kumar G.,"SVM Classifier for the Prediction of Era of an Epigraphical Script,"IJP2P,Vol.2 No.2, pp 12-22,April 2011.
5. Seethalakshimi R.,Sreeranjani BalachandranT.,"Optical Character Recognition for Printed Tamil Text Using Unicode,"Journal of Zhejiang University SCIENCE, Vol 6 A(11), pp 1297-1305.
6. Villingiriraj E., Balasubramanie P.,"Recognition of Ancient Tamil Handwritten Characters in Palm Manu Scripts using Genetic Algorithm,"IJSET,Vol.2 No.5,pp 342-346,May 2013.
7. Hossein Ziaei Nafchi et," Phase Based Binarization of Ancient Document Images:Model and Applications,"IEEE Transactions on Image Processing,Vol . 23,No.7,July 2014
8. Christianini N.,Shawe-Taylor.J," An Introduction to Support Vector Machines: and other Kernel-Based Learning Methods," Cambridge University Press,Cambridge (2000).