THE PROCESS OF EPIGRAPHY – A COMPARATIVE STUDY

[1]B. Priyadarshini, [2]Dr.J. Preethi
[1]Research Scholar, [2]Assistant Professor
[1]Department of Computer Science and Engineering,
[1]Anna University Regional Campus, Coimbatore, Tamil Nadu, India

Abstract: The study of Ancient Tamil land and its language are favored by the sources of literary, archaeological, epigraphic and numismatic arena. The inscriptions are of immense value. It is one of the major sources of historical facts. Tamil inscriptions form about 60% of all the inscriptions found in India. Tamil inscriptions are also found in faraway lands, which confirms ancient trade. It has been found in Italy, Egypt, Thailand, etc. Epigraphic as well as archaeological data has to be deciphered and digitized so that the history of information will not be lost any further. These have been projected as a major concern for all the epigraphists and archaeologists as well. This paper consists of an extensive survey on various research works based on the conversion of ancient Tamil characters to modern text. This survey paper highlights the various image processing techniques, methodologies and algorithms used.

Index Terms - **Image processing, Epigraphy, Archeology, Ancient Tamil characters, Deciphering and Digitization.**

I. INTRODUCTION

Epigraphy is the study of inscriptions. It is the art of categorizing graphemes, elucidating their significances, sorting their usages according to periods and racial situations and drawing inferences about the inscription. The archetypal discoveries on epigraphy in several regions of South Asia are obtained as the decoded messages of the language. The language of these inscriptions is a peculiar kind of Tamil and not actually the conventional Tamil of the Sangam verse. Both the modern Tamil script and the Vatteluthu lettering progressed from this parental lettering. The inscriptions can be found in Temples, Palm Leaf Manuscripts, Caves, Pottery and Copper Plates. The inscriptions talk about history, forgotten unrecoverable literature, trade etc. Many literary works have been lost due to various reasons. Without the inscriptions, study of literature would be incomplete. The inscriptions not only talk about the kings but also about common man. This helps us in understanding the socio economic status of the common man at that time. From the inscriptions, one can learn about the ancient justice system as well. The inscriptions also talks about the donations given out by the rulers to various eminent citizens praising their contributions. During the later periods of history when the ***Kalabhras*** became prominent, Tamil lost its importance. But the small rulers were able to safeguard the history. The inscriptions in ***'Veeran Nadukal'*** talks about that.

As every day passes, the ancient inscriptions in the caves and dilapidated structures are being destroyed unknowingly. Also some of the restorations have being carried out in temples either obscures them or damages them. Prehistoric murals are vanished as well. Since these inscriptions are a marker for the historical evidence of the language and culture, it is imperative that we save these relics. In order to understand the ancient Tamil culture, it is very important that those inscriptions are documented, archived and translated to modern Tamil. This will provide a great source of learning and research for the students and researchers of today. If the history and culture is lost, the entire civilization will lose its pride. In this regard it is very essential that these ancient inscriptions are saved and documented. Many efforts are being made to document. But it is also important that we understand what is written in the inscriptions. These are some of the compelling reasons to digitize and convert it to contemporary Tamil. The study might also establish the connections between the inscriptions, if they are digitized and made searchable.

The following are the limitations of these translation projects:

- The inscriptions may not be continuous. It could have been damaged or could be unreachable. This will leave holes in the text that are deciphered out of the inscriptions.
- When it comes to the translation of regional dialects, the context could cause the issues with respect to accurate translation.
- Some of the meanings of those words in the inscriptions would have changed over time and there may be some of the words that might not have been documented.

- These challenges are to be handled with the help of literary scholars.

II. LITERATURE REVIEW

Recent research publications exhibit different techniques to digitize and deal with epigraphic processes.

R. Angelin Jennifer et al. [1] discusses on the recognition of ancient Tamil characters and converting it into modern Tamil. Thus, a good system was introduced to achieve this methodology named *Image Glazing for Thinning of images*. Hence, it produces fine-tuned thin images for each and every character input.

E.K.Vellingiriraj et al. [2] develops the system that can recognize Tamil characters from the source and convert them into text format using the *Boolean Matrix and Breadth First Search (BFS)*.

S. Rajakumar et al. [3] finds a way to identify the centennial data of ancient Tamil characters and translating them into modern day Tamil characters using a method called *Contour-let Transform and Fuzzy Logic* plays a major role in training the data sets and comparing the data with the current century information to have a more accurate recognition.

N. Sridevi et al. [4] attempts to recognize ancient Tamil scripts such as Tamil Brahmi and Tiruvalangadu plates of Rajendra Chola by the method of segmentation. It divides the script documents in a linear fashion from line-by-line to word-by-word and then character-by-character using projection profile method and *PSO algorithm*. By this method, an efficient result on segmentation is arrived when compared to other methods of segmentation.

III. PREPROCESSING TECHNIQUES

The characters from the input image source of the stone inscriptions are very difficult to recognize and the recognition is very complicated since the ancient Tamil scripts differ from person to person, in intensity, style, scale and orientation and finally in written format too. In all these four papers, the immediate step to recognition is to process the acquired source image.

In R. Angelin Jennifer et al. technique [1], the first step to preprocessing is background subtraction which obtains the Region of Interest (ROI) using Otsu`s method of global thresholding. The next step would be the noise removal using the generalized method of thresholding, which cleans an infused image with noise. Following is the step for varnishing the image that glosses the borderlines or a structure of an image by applying linear Gaussian filter and using iterative thinning. Thinning is one such morphological process that peels the concentrated tiers of the edges from an image producing a one-pixel thickness image representation. By this, it is rest assured that this technique of thinning the image lowers the storage requirements of the features being extracted, and also takes to the next step, which is the identification level of the complete technique. The best and classic method for thinning is ***Zhang – Suen Thinning algorithm.***

In E.K.Vellingiriraj et al.technique [2], the step before preprocessing is to first scan the image of the Tamil palm manuscript, which is gathered from different places of different centuries using the 4800 dpi scanner. This scanned image is then stored as the JPEG image. This scanned JPEG image is preprocessed for the feature extraction phase. This step starts from image cropping which trims the character from the script. The trimmed position in the image is in a different color than the remaining spaces. This is followed by graphemes extraction or edge detection method for segmentation of characters. As characters in the stone inscriptions vary in size, these characters are resized to a unique size say 100*100 pixels image, followed by thinning of characters. The last step will be image binarization where each character is stored in the form of 1's and 0's using ***the Image Zoning technique***.

In S. Rajakumar et al. technique [3], the input image is sent for noise removal followed by skewing. This correct skew image checks for an angle of orientation of +/- 15 degrees and matches with a horizontal axis. The next step is smoothing which removes again the unnecessary noise using *Fuzzy median filter method* followed by *Gaussian filter* for effective noise reduction. And then will be the thresholding that extracts the foreground from the background. The last step is the segmentation by clustering where top down or bottom up segmentation takes place based on the arrangements of the scripts. The segmentation is by the selection of mean points where mean of each cluster is computed and then by minimizing the squared distances to the center.

In N. Sridevi et al. technique [4], the preprocessing step is by first understanding the Tamil scripts and text lines without any methodology. Learning about the Tamil scripts and the physical structure of a historical document image is the preprocessing step to

recognition. A base line, median line, upper line, and lower line, touching components and overlapping components are the structure of a text line. Comprehending such a structure is the first step to segmentation wherein it is an important activity for any character recognition system.

Table 3.1: Comparison Study on the Image Processing techniques used.

S. No	Properties	R.Angelin Jennifer et al. [1]	E.K. Vellingiriraj et al. [2]	S.Rajakumar et al. [3]	N. Sridevi et al. [4]
1	Background Subtraction through Image cropping	Yes	Yes	Yes	No
2	Noise Removal using thresholding and filters	Yes	Yes	Yes	No
3	Segmentation techniques followed	No	Yes	Yes	Yes
4	Thinning Technique	Yes	Yes	No	No
5	Binarization Technique	No	Yes	Yes	No

IV. METHOD OF RECOGNITION

This is the next phase of processing, where in the identification system helps in pinpointing the primordial Tamil scripts. In these four papers, there are various methods of recognition, which achieves one specific goal that is the translation.

In R. Angelin Jennifer et al. methodology [1], the feature extraction followed by pattern matching is the steps in recognition system. The term used for feature extraction is universe of discourse where a matrix is developed and calculated such that the shortest one makes the character skeleton. It is mainly based on the regions of different geometric line segments positioned in the script. Once this reference points are identified, the next step is to understand the " ***Zone*** ", where the image file is additionally disseminated into equal sized matrices in order to have a feature selection on the individual matrix. This can obtain the intricate details of the script skeleton. The overall total of straight lines, overall total of upright lines, the total of right slanting lines, the total of left slanting lines, the length of all straight, upright, right slanting and left slanting lines and area of the skeleton are the feature vector associated to each and every zone. Knowing these features, line type normalization that is the pixel value and normalized length can be calculated from which pattern matching is appreciated. It helps to correlate all possible inputs to the trained data sets, which enhances supervised learning. The identical collection of data then corresponds the data that is tested, with the closely related script.

In E.K.Vellingiriraj et al methodology [2], the feature extraction followed by character recognition has a set of following sub-processes. It starts with Image to Boolean matrix process wherein the pixels where 1's and 0's are available are pointed to represent the nodes in the graph. Then the graph is generated using the graph notation of vertex and edge. Using the ***BFS***, it is noted that every node travels via the edges of the network nodes from the start node to the end node. When crossing all the nodes in the network, it is the responsibility of the algorithm to have relative comparison with all the existing ancient scripts as the data set loaded in the database. While the result of the comparison tends to be equal, a corresponding script in the contemporary Tamil is delivered from the database. Thus, the last step to enhance the recognition process, image grouping of different styles and strokes that are stored as matrix along with character modeling, to analyze every character set as a single unique character is recorded.

In S. Rajakumar et al. methodology [3], the recognition is through a method of neural networking, which requires artificial intelligence. It is a circuit of artificial nodes called ***neurons***, which are connected functionally to form a particular physiological function. Artificial neural networking is a study of solving artificial intelligence problems that are highly complex. In this, the nodes are in the form of a graph where in the left most node has low weight and as the nodes are connected to the center, the weights are added up and the highest weight is at the rightmost node. Finally, based on the weight feature, the character is compared in the database of the test data; thereby the equivalent character is displayed. The fundamental of neural network is all possible data needed to train more accurately such that the weights are not unique for different data inputs that are fed.

In N. Sridevi et al. methodology [4], the method of recognition is by the segmentation of an image into regions or objects. This paper deals with segmentation at various levels. Henceforth, segmenting scripts is an inevitable task for any Character recognition system. Segmentation is a means of dividing the document image into text lines, followed by words and then into characters, which is used for further processing of the scripts. To overcome the problem of overlapping lines and characters in segmentation, the first method called the projection profile and ***PSO*** followed by connected components with the nearest neighborhood method is used. Thereby, minute differentiation such as removal of consonant and vowel modifiers are appreciated.

Table 4.1: Comparison Study on the Character Recognition techniques used.

S. No	**Properties**	**R.Angelin Jennifer et al. [1]**	**E.K. Vellingiriraj et al. [2]**	**S.Rajakumar et al. [3]**	**N. Sridevi et al. [4]**
1	Feature Extraction	Universe of Discourse	Graph Generation	Artificial Intelligence	Various Segmentation techniques uses Computation Intelligence
2	Algorithms Used	Determining Line type Normalization and Normalized Length with Pattern Matching.	Finding Boolean Matrix to Graph Generation followed by BFS with Pattern Matching.	Neural Networks with Artificial Intelligence using corresponding weights at the nodes.	Segmentation at various levels from line to word and then to character.
3	Trained Data Set needed	Yes	Yes	Yes	Yes
4	Performance Indicators	Increased Recognition Rate.	Predictability is achieved.	Better Accuracy.	Better Robustness.

V. FINDINGS

These four research papers have gathered a comparative study of the knowledge in the field of epigraphy. By understanding these papers, a substantial rise has been achieved to increase the quality of the techniques. The comparative study of the methodologies is prepared to comprehend below.

Table 5.1: Comparison Study on the techniques used and the problems associated with the techniques.

S. No	Properties	R.Angelin Jennifer et al. [1]	E.K. Vellingiriraj et al. [2]	S.Rajakumar et al. [3]	N. Sridevi et al. [4]
1	To convert ancient script to modern script	Yes	Yes	Yes	No
2	To identify century information	No	No	Yes	No
3	Main aim is to identify ancient script	Yes	Yes	Yes	Yes
4	Main Techniques used	Zhan- Seng Thinning Algorithm, Universe of Discourse	Boolean Matrix to Graph, BFS	Contour- let transform, Segmentation by clustering, Artificial Intelligence and Neural Networks	PSO, Projected Profile, Connected components,
5	Pre- processing methods available	Yes	Yes	Yes	No
6	Conclusion	Image glazing produces a fine tuned image for thinning the characters by which the rate of recognition is increased.	Digitizing the Tamil Manuscripts and converting them into modern Tamil.	A contour let based method on offline text independent of ancient Tamil handwritten character identification is presented with high rate of accuracy.	A new method is proposed using PSO in which an optimal threshold for segmenting lines and connected components to segment characters. This method is done in a robust way producing good results.
7	Problems associated as future work	Nil	Difficult in dealing with the Cursive writing of the Tamil script or by joining of two Boolean matrices. Damage in the palm leaf manuscript.	Enlarging the database with more handwritten scripts.	If the breaks amongst the basic character and modifiers are more, segmenting the touching lines and characters are difficult.

VI. CONCLUSION

The analysis work has gathered a comparative study of the data within the field of epigraphy. By understanding these studies, a considerable rise has been achieved to extend the standards of the techniques. The other papers include various techniques such as TSVM (Transductive Support Vector Machine) for foreseeing the Tamil Scripts using a calculative approach, which helps in finding which character belongs to which century, Haar wavelet features with SVM (Support Vector Machine) is used to represent the overall shape and details of the image, Zoning with structural feature extraction method are also used to have an enhanced feature extraction method for script recognition, and so on. Thus, the research discoveries have been made to understand that most of the procedures require basic step of processing the image, and of course, most of the methodologies are very uncontrollable in time, writer reliant and script reliant. It's created to differentiate that numerous techniques will be utilized in a similar field for the similar purpose to reinforce the accurateness, consistency, expectedness, and efficiency of the system in use.

REFERENCES

1. R. Angelin Jennifer and G. Bhuvaneswari. June 2014. " Image Glazing for Thinning of Ancient Tamil Characters " in International Journal of Scientific & Engineering Research.
2. E.K. Vellingiriraj and Dr. P. Balasubramanie. March 2014. " Recognition of Ancient Tamil Handwritten Characters in Historical Documents by Boolean Matrix and BFS Graph" in International Journal of Computer Science and Technology.
3. S. Rajakumar and Dr. V. Subbiah Bharathi. July 2011. " Century Identification and Recognition of Ancient Tamil Character Recognition" in International Journal of Computer Applications.
4. N. Sridevi and P. Subashini. August 2012. " Segmentation of Text Lines and Characters in Ancient Tamil Script Documents using Computational Intelligence Techniques" in International Journal of Computer Applications.
5. Antony Robert Raj.M, Abirami.S, Murugappan.S. Pages from INFIT 2013. " Analysis of Statistical Feature Extraction Approaches Used in Tamil Handwritten OCR", V7-27.
6. M.V. Jeya Greeba, G.Bhuvaneswari. February 2014. " Recognition of Ancient Tamil Characters in Stone Inscription Using Improved Feature Extraction", International Journal of Recent Development in Engineering and Technology, ISSN: 2347-6435, Volume 2, Special Issue 3.
7. T. S. Suganya and Dr. S. Murugavalli. 2014. "Binarization of Ancient Tamil Scripts From Stone Inscriptions", International Global Journal For Research analysis, Volume-3, Issue-11, Nov Special Issue-2014, ISSN No 2277-8160.
8. Soumya A and G Hemantha Kumar. April 2011. "SVM Classifier For The Prediction Of Era Of An Epigraphical Script", International Journal of Peer-to-Peer Networks (IJP2P), Vol.2, No.2.
9. K. Kaviya Selvi, and R. S. Sabeenian. April 2015. "Restoration of Degraded Documents Using Image Binarization Technique", ARPN (2006-2015 Asian Research Publishing Network) Journal of Engineering and Applied Sciences, Vol. 10, No. 7, ISSN 1819-6608.
10. S. Venkata Krishna Kumar, T.V. Poornima. September 2014. "An Efficient Period Prediction System for Tamil Epigraphical Scripts Using Transductive Support Vector Machine", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue 9, ISSN: 2278-1021.
11. Digital Image Processing by Rafael C.Gonzalez.