Text Analysis and Information Retrieval of Historical Tamil Ancient Documents Using Machine Translation in Image Zoning

E. K. Vellingiriraj, M. Balamurugan, and P. Balasubramanie

Abstract—**The aim of this paper is to develop a system that involves character recognition of Brahmi, Grantha and Vattezuthu Characters from palm manuscripts of Historical Tamil Ancient Documents, anaylsed the text and machine translated the present Tamil digital text format. Though many researchers have implemented various algorithms and techniques for character recognition in different languages, Ancient characters conversion still poses a big challenge. Because Image recognition technology has reached near-perfection when it comes to scanning English and other language text. But optical character recognition (OCR) software capable of digitizing printed Tamil text with high levels of accuracy is still elusive. Only a few people are familiar with the ancient characters and make attempts to convert them into written documents manually. The proposed system overcomes such a situation by converting all the ancient historical documents from inscriptions and palm manuscripts into Tamil digital text format. It converts the digital text format using Tamil Unicode. Our algorithm comprises different stages: i) image preprocessing, ii) feature extraction, iii) character recognition and iv) digital text conversion. The first phase conversion accuracy of the Brahmi script rate of our algorithm is 91.57% using the neural network and image zoning method. The second phase of the vettezhuthu character set is to be implemented. Conversion accuracy of Vattezhuthu is 89.75%**

Index Terms—**Character recognition, vattezhuthu, segmentation, image zoning, machine translation.**

I. Introduction

Tamil is one of the oldest languages in the world with a rich literature. In the ancient days, the writers, especially in Tamilnadu, have used palm leaves and inscriptions to encrypt their writings. A very good example of the usage of Palm leaf manuscripts is to store the history of Tamil grammar book named Tholkappiyam which is written during 4th B.C. The ancient literature includes many palm leaf manuscripts that contain rare commentaries on Sangam works, unpublished portions of classics, Saiva, Vaishnava and Jain works, poetry of all descriptions, medical works of exceptional values, food, astronomy & astrology, Vaastu & Kaama Shastra, jewelry, music, dance & drama, medicine, Siddha and so on. Palm manuscripts are utilized for 3 different categories. Document registration of land and building which are donated by the kings to the people are encrypted with palm manuscripts. The literary works, grammar, astrology, science and technology, etc., are encrypted in palm manuscripts. Historical moments of places and dominion are also encrypted. Character recognition is one of the most difficult tasks in the pattern recognition system. There are a lot of difficulties in image processing techniques. To solve these, one should know how to i) separate the characters in the segmentation process, ii) to recognize unlimited character fonts and sculpting styles in noisy image and iii) distinguish characters that have the same shape, but have different pronunciation in characters like (Ng) and (iT). Many researchers have tried to apply many techniques for breaking through the complex problems of character recognition. The most difficult thing is to recognize the sculpting of Brahmi characters in stone compared to the other character recognition from different sources. The Brahmi characters were sculpted in stones, clay pot, copper plate etc. In this paper, a character recognition system based on a statistical approach with a new feature set is proposed.

II. Brahmi and Vattezhuthu Scripts

Brahmi script is one of the most important writing systems in the world by virtue of its time depth and influence. It represents the earliest post-Indus corpus of texts, and some of the earliest historical inscriptions found in India. The best-known Brahmi inscriptions are the rock-cut edicts of Ashoka in North-central India, dated to 250–232 BCE. This elegant script appeared in India most certainly by the 5th century BCE, but the fact is that it had many local variants even in the early texts which suggest that its origin lies further back in time. There are several theories on the origin of the Brahmi script. Brahmic scripts are descended from the Brahmi script. The most reliable of these are short Brahmi inscriptions dated to the 4th century BCE and published by Coningham *et al.* [1] but scattered press reports have claimed both dates as early as the 6th century BCE and that the characters are identifiably Tamil Brahmi though these latter claims do not appear to have been published academically.

The Vatteluttu (or Vattezhuttu) is an alphabet used in the southern India mainly in the states of Kerala and Tamil Nadu. It has grown from the Brahmi script of southern part from around 6th century CE, and utilized to write Tamil and Malayalam languages. The main historical development in it is that the insignificant signs obtained from Brahmi to

Manuscript received September 14, 2016; revised December 5, 2016.

E. K. Vellingiriraj and P. Balasubramanie are with the Department of Computer Science and Engineering, Kongu Engineering College, Perundurai, Tamil Nadu, India (e-mail: girirajek@rediffmail.com, balu_p@kongu.ac.in).

M. Balamurugan is with the Department of Computer Science and Engineering, Christ University, Bangalore, Karnataka, India (e-mail: balamurugan.m@christuniversity.in).

doi: 10.18178/ijlll.2016.2.4.88

represent Tamil alphabets are removed from Vatteluttu. From the 8th century CE, Tamil language has been written in both Vatteluttu and Tamil scripts. The reason is that different kingdoms employed different scripts in the era of scattered political nature of southern India. During 15th century CE, Vatteluttu script was supplanted by Tamil script in Tamil-speaking areas. Similarly, in 12th century CE in Kerala, the Malayalam script was developed from the old Grantha script and so it became phonetically suitable to the Malayalam language. This situation led to the misuse of Vatteluttu instead of the Malayalam script. Fig. 1 shows the family of Ancient Brahmi and Vattezhuthu script that belong to 500BCE.

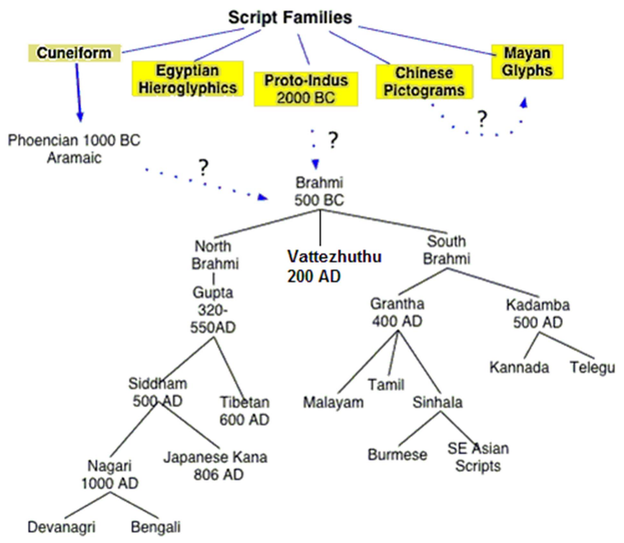


Fig. 1. Ancient script family.

III. Literature Review

Three major approaches used commonly for recognition of handwritten characters are statistical, structural or syntactic and neural network based approaches.

A. Statistical Approach

Statistical or probability functions are used for formulating a recognition algorithm in statistical approach. Features to be given as input are derived from a set of measurements used for pattern recognition. There prevails a difficulty in representing the pattern of classification on the basis of structural information in this statistical approach [2]-[8].

B. Structural or Syntactic Approach

Syntactic approach employs syntactic or structural information about patterns to obtain information relevant to patterns. It extracts similar patterns and formulates pattern syntax or structural rules. The information about the pattern syntax rules helps highlight, classify and identify unknown patterns. The structural approach is favorable for evolving a recognition system for handwritten characters. The reason is that it uses structural rules to build much pattern syntax for handwritten characters, but it is not easy to formulate learning structural rules [9]-[13].

C. Neural Network Based Approach

Neural pattern recognition is a pattern recognition approach that is based on the storing and manipulation of information by a biological neural system. This is an artificial neural system and it is named as "neural networks". This system is expected to solve all the problems with automatic reasoning, including pattern recognition problems. It classifies patterns on the basis of predictable properties of neural networks. However, it represents only a little amount of information about semantic from a network [14]-[16].

IV. Methodology

All the details of the proposed system design are presented below. Initially, the framework of the ancient Brahmi character recognition system is started. Then the component details are given. The Fig. 2 shows the vattezhuthu set. The system gets the input image of Brahmi characters and converts it into equivalent current Tamil digital format.

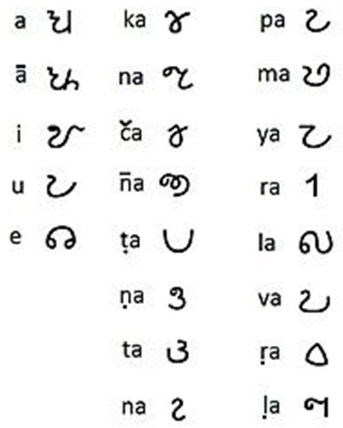

a
ā
i
u
e
ka
na
ča
ña
ṭa
ṇa
ta
na
pa
ma
ya
ra
la
va
ṛa
ḷa

Fig. 2. Vattezhuthu character set.

A. Overview of System Architecture

Fig. 3 shows the overall architecture of Tamil Brahmi character recognition system for stone inscriptions. The major components of this system are image preprocessing, feature extraction and character recognition. The architecture of the system is given below.

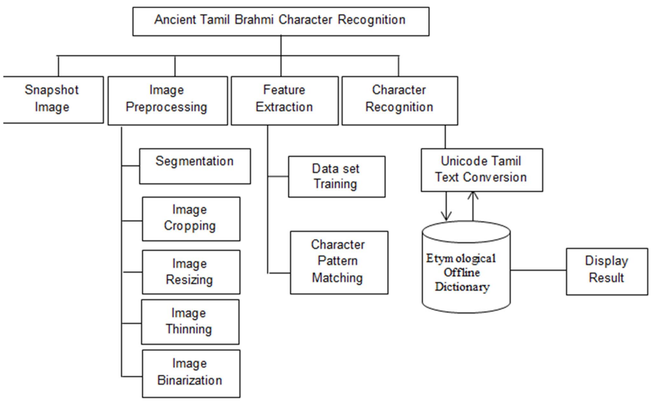


Fig. 3. Architecture of ancient Tamil Brahmi character recognition system for stone inscription.

B. Structure of the System

Based on the system framework formulated, the Brahmi image is converted into Tamil text format [17]. The framework involves i) capturing of an image, ii) preprocessing, iii) feature extraction, iv) character recognition and v) text conversion.

C. Image Capturing

In the first stage, the Brahmi Inscription images are collected from various places. The snapshot images are captured with high quality or high resolution High Definition / Digital Single Lens Reflex (HD / DSLR) camera and stored in JPEG format (Fig. 4 and Fig. 5).

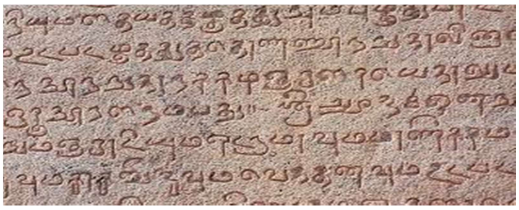


Fig. 4. A Vattezhuthu script in temple epigraphy.

D. Image Preprocessing

During the image preprocessing stage, a Brahmi character image is prepared for feature extraction. This stage consists of four sub-processes related to an image: a) cropping, b) segmentation, c) resizing, d) thickening and e) binarization. All these sub-processes are explained below.

a) Image cropping: The Brahmi Inscription image has the white space for each character, which means that in the inscription image, the Brahmi characters have spaces in between characters. In the stone image, the characters are easily identifiable from each pixel color. The image is changed to grayscale or black and white image and the noisy and unwanted spaces are removed.

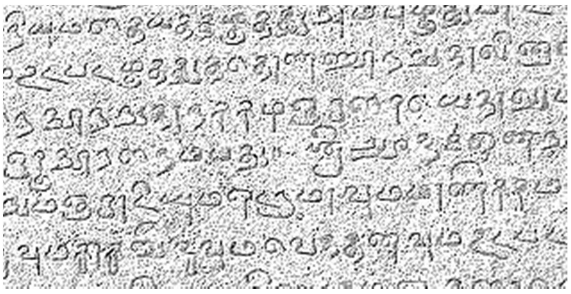


Fig. 5. Using Sobel's edge detection method for segmentation.

b) Segmentation: Using the grapheme extraction [18] technique, each character can be segmented by the individual characters. Segmentation is divided into three groups: line segmentation, word segmentation and character segmentation. In the present approach, the line (Fig. 6) and character segmentation are applied (Fig. 7).

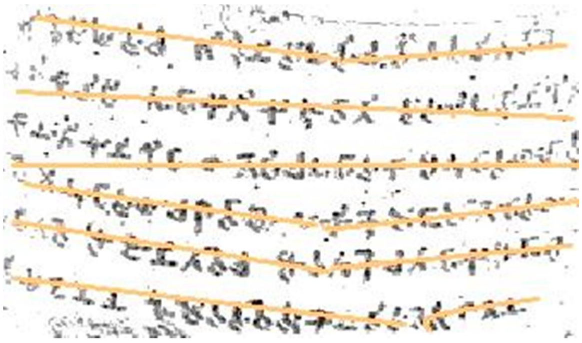


Fig. 6. Line segmentation.

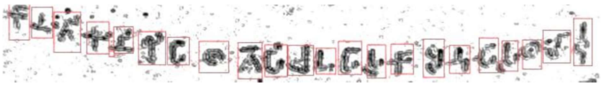


Fig. 7. Character segmentation.

c) Image resizing: After cropping the particular character, each character may be of different size. So, each character has to be changed to be of equal size. The character image is resized as 100×100 pixel image.

d) Image thinning: The dark pixel (i.e. A character) is converted to be the thinning character. A thickened character extracts easily the thin character by using a nearest pixel darken to lighten color change to a particular range.

e) Image binarization: A character stores the Boolean matrix that is used to store either 0's or 1's. The dark pixel is stored as 1's and light pixel is stored as 0's using the image zoning technique.

E. Data Set Training

Different classes of Brahmi characters and Vattezhuthu from various writers of all the 237 Vattezhuthu characters are collected and stored in the training data sets.

F. Character Recognition

A data set of Vattezhuthu characters is matched with the current user data set. The edge detection algorithm is used to identify each character in the training datasets using image zoning.

G. Unicode Text

After matching the character, the equivalent current Digital Tamil text is converted using Unicode.

H. Retrieve from the Database

After converting the Unicode text, spelling and meaning are checked in the dictionary database. If a word does not give the meaning in the sentence, the nearest meaning is to be searched from the database.

V. Proposed Algorithm

For Zoning, let *ZM*= {*z*1, *z*2, *z*3,..., *zM*} be a method of zoning. The zoning based classification deals with the way in which each feature is detected from a pattern *x* that has an influence on each zone of *ZM*. Let the classification of a pattern *x* into one class of $\Omega = \{C_{1,}C_{,2} \ldots\ldots C_K\}$ be considered using the feature set $F = \{f_{1,}f_{,2} \ldots\ldots f_T\}$. In this, *x* can be described by the feature matrix, A_x, of *T* rows (features) and *M* columns (zones).

$$Ax = \begin{bmatrix} A_x(1,1)A_x(1,2) \ldots\ldots A_x(1,j) \ldots A_x(1,M) \\ A_x(2,1)A_x(2,2) \ldots\ldots A_x(2,j) \ldots A_x(2,M) \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ A_x(i,1)A_x(i,2) \ldots\ldots A_x(i,j) \ldots A_x(i,M) \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ A_x(T,1)A_x(T,2) \ldots\ldots A_x(T,j) \ldots A_x(T,M) \end{bmatrix}$$

with $A_x\,(i,j) = \sum_{\text{instances of } fi \text{ in } x}^{wij}$
where *wij* represents the weight that determines the degree of influence of an instance of feature f_i on zone z_j. Different decompositions of the character of an image and the density of foreground pixels are measured in each zone to obtain

some features so as to create the first feature subset. The number of black pixels is divided by the total number of pixels in order to find the density of each zone.

VI. Results and Discussion

A. Database

To measure the effectiveness of the proposed OCR system, the Vattezhuthu database is used as a resource for training and testing processes. The database includes 5000 characters stored for training and 4000 handwritten Vattezhuthu characters produced by 60 writers, who wrote 5 samples of each 30 classes for testing. The Vattezhuthu characters are old one and so the writers have not written the test characters clearly and correctly. After some practice, the writers have written 5 samples of X 30 characters, totally 150 characters from each writer. The training set is composed of 11 vowels, 18 consonants and 198 consonantal vowels, totally 227 characters of all class and the testing set is composed of approximately 30 characters of each class.

The lack of character recognition is due to three reasons. One is that the consonantal vowel characters are similar to consonants as shown in Fig. 8.

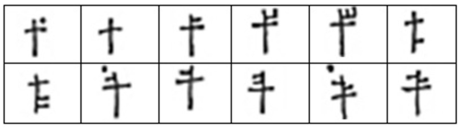


Fig. 8. Similarity of consonantal vowels.

The second one, the writers do not write the Brahmi characters properly. Moreover, the characters are stored in different strokes and styles with ambiguity as revealed in Fig. 9.

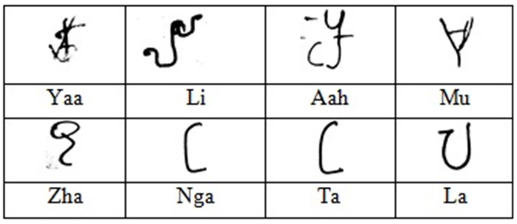


Fig. 9. Unknown characters in the data set.

The same character with different style and stroke from different writers is given in Fig. 10.

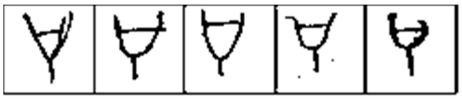
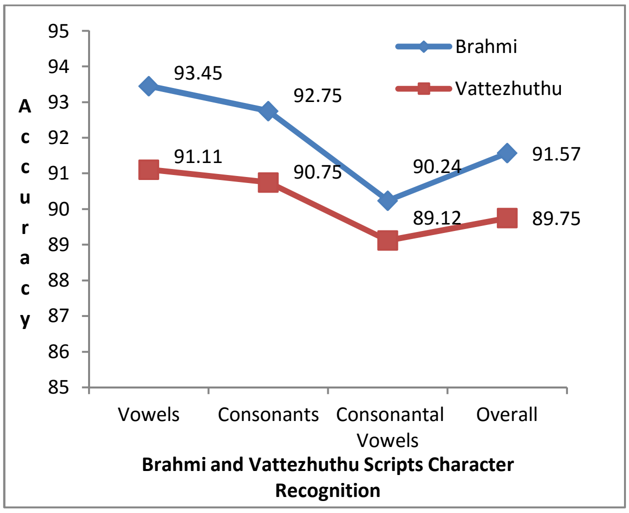


Fig. 10. Maa character from different writers.

Third one, the data set stores only a minimum number of data, i.e. only 6000 characters are stored in the database for training set.

TABLE I: Comparison of the Results of Proposed Algorithms

	Vowels	Consonants	Consonantal Vowels	Overall
Brahmi	93.45	92.75	90.24	91.57
Vattezhuthu	91.11	90.75	89.12	89.75



In the Table I shows that the comparison of both Brahmi and vattezhuthu ancient character's accuracy of proposed algorithms for Machine translation of the Tamil digital text. The conversion rate of vowels in Brahmi character is 93.45% [19] and the Vattezhuthu is 91.11%. The Consonants is 92.75% of Brahmi characters and 90.75% of vattezhuthu. The consonantal vowels of Brahmi are 90.24% and the vattezhuthu is 89.12%. The overall character conversion is 91.57% of Brahmi characters and the 89.75% of vattezhuthu. So the comparison of these results is Brahmi character set is more accuracy with the vattezhuthu.

VII. Conclusion and Scope for Future Work

In the present paper on Phase II, an approach for character recognition of vattezhuthu characters using the image zone based on classification and recognition is proposed. The next phase implements this approach to Grantha. This is the new way of approaching Ancient vattezhuthu character recognition, the result of which is found to be above 90% for consonant and vowel letter recognition, whereas the consonantal vowel recognition is somewhat low due to the lack of similarity of letters. The overall output of the proposed algorithm is 89.75% accurate which higher than that of the existing systems. Nevertheless, the Brahmi character and vattezhuthu recognition systems still contain many problems that require more efficient algorithms to solve the three reasons mentioned above. In future, the system can be modified with the incorporation of better feature extraction methods to improve the accuracy rate. The writers shall get trained in writing the Brahmi letters to create the data set. By that, the error rate of ambiguous letters will be reduced. In addition, the system in future will recognize the Brahmi letters and vattezhuthu from the stone inscriptions for improving the result in using some hybrid algorithms.

References

[1] R. A. E. Coningham, F. R. Allchin, C. M. Batt, and D. Lucy, “Passage to India? Anuradhapura and the early use of the Brahmi Script,” *Cambridge Archaeological Journal*, vol. 6, no. 1, pp. 73-97, 1996.

[2] S. Seth, G. Nagy, and M. Viswanathan, “A prototype document image analysis system for technical journals,” *Computer*, vol. 25, no. 7, pp. 10-22, Jul. 1992.

[3] F. Shafait, T. M. Breuel, and D. Keysers, “Performance evaluation and benchmarking of six-page segmentation algorithms,” *IEEE Trans. Pattern Analysis Machine Intelligence*, vol. 30, no. 6, pp. 941-954, Jun. 2008.

[4] Y. Liang, M. C. Fairhurst, and R. M. Guest, “A synthesisd word approach to word retrieval in handwritten documents,” *Elsevier Pattern Recognition*, vol. 45, pp. 4225-4236, Jun. 2012.

[5] G. Pirlo and D. Impedovo, “Adaptive membership functions for handwritten character recognition by voronoi-based image zoning,” *IEEE Trans on Image Processing*, vol. 21, no. 9, pp. 3827-3836, Sep 2012.

[6] C. Pornpanomchai, S. Jeungudomporn, V. Wongsawangtham, and N. Chatsumpun, “Thai handwritten character recognition by genetic algorithm (THCRGA),” *IACSIT Journal of Engineering and Technology*, vol. 3, no. 2, Apr. 2011.

[7] Q.-F. Wang, F. Yin, and C.-L. Liu, “Handwritten Chinese text recognition by integrating multiple contexts,” *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 34, no. 8, Aug. 2012.

[8] T. M. Jose and A. Wahi, “Recognition of Tamil handwritten characters using Daubechies wavelet transforms and feed-forward back propagation network,” *International Journal of Computer Applications*, vol. 64, no. 8, pp. 0975-8887, Feb. 2013.

[9] J. Chen and D. Lopresti, “Model based ruling line detection in noisy handwritten documents,” *Pattern Recognition Letters*, 2012.

[10] A Bharath and S. Madhvanath, “HMM-based lexicon-driven and lexicon-free word recognition for online handwritten Indic scripts,” *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, Apr. 2012.

[11] C. Pornpanomchai, D. N. Batanov, and N. Dimmitt, “Recognizing Thai handwritten characters and words for humancomputer interaction,” *International Journal of Human-Computer Studies*, pp. 259-279, 2001.

[12] C. Pornpanomchai, N. Pisitviroj, P. Panyasrivarom, and P. Prutkraiwat, “Thai handwritten character recognition by Euclidean distance,” in *Proc. the 2nd International Conference on Digital Image Processing (ICDIP 2010)*, 2010, pp. 53-58.

[13] C. Pornpanomchai and M. Daveloh, “Printed Thai character recognition by genetic algorithm,” in *Proc. The International Conference on Machine Learning and Cybernetics*, 2007, pp. 3354-3359.

[14] B. Kruatrachue, N. Pantrakarn and K. Siriboon, “State machine induction with positive and negative for Thai character recognition,” in *Proc. The International Conference on Communications, Circuits and Systems*, 2007, pp. 971-975.

[15] P. Sanguansat, S. Jitapunkul, and W. Asdornwised, “Online Thai handwritten character recognition using hidden Markov models and support vector machines,” in *Proc. the International Symposium on Communications and Information Technologies*, 2004, pp. 492-497.

[16] R. Budsayaplakorn, W. Asdornwised, and S. Jitapunkul, “On-line Thai handwritten character recognition using hidden Markov model and fuzzy logic,” in *Proc. the IEEE 13th Workshop on Neural Networks for Signal Processing*, pp. 537-546, 2003.

[17] E. K. Vellingiriraj and P. Balasubramanie, “Recognition of ancient Tamil handwritten characters in palm manuscripts using genetic algorithm,” *International Journal of Scientific Engineering and Technology*, vol. 2, issue 5, pp. 342-346, 2013.

[18] K. Siriboon, A. Jirayusakul, and B. Kruatrachue, “HMM topology selection for on-line Thai handwriting recognition,” *The First International Symposium on Cyber Worlds*, pp. 142-145, 2002.

[19] E. K. Vellingiriraj and P. Balasubramanie, “Automatic digitization of ancient Brahmi characters into Tamil digital texts using image zoning from palm manuscripts and stone inscriptions,” in *Proc. International Conference on Digital Humanities (CDH2015)*, The Open University of Hong Kong, Hong Kong, vol. 1, issue 1, p. 49, Dec. 17-18, 2015.

E. K. Vellingiriraj completed his BSc (computer science) degree in 1999, then received his M.C.A. degree in 2002 and completed M.E. degree in software engineering from Anna University, Coimbatore, India in 2010. He completed MBA (HR) in Indira Gandhi National Open University (IGNOU), New Delhi in 2010. He is very much interested in Tamil literature which led him to completed M.A. Tamil in TNOU, Chennai. Currently he is doing his research in Natural Language Processing of Character Recognition from Historical Palm Manuscripts and Stone Inscriptions in Anna University, Chennai. He is also completed master degree in psychology, sociology and political science in Tamilnadu Open University, Chennai. Totally, he has 13 years’ experience including 3 years industrial experience and 10 years of teaching. He presented his research work in international conferences held at Malaysia, Thailand, Mauritius and Hongkong. The Tamil Nadu Government recognized him for presenting his paper on 12th International Tamil Internet Conference held at Malaysia. And also UGC Granted to present his paper on First International Conference on Tamil Diaspora, Mauritius on coming July 2014. Erode Tamil Mandram and Pavendar Bharathidasan Tamil Mandram have announced the cash award for his research work. He has created his own Tamil Blog Spot and developed various Android apps for Tamil people and others.