Character Recognition of Ancient South Indian language with conversion of modern language and translation

A.Naresh Kumar [1] and Dr.G.Geetha[2]

[1] Research Scholar, Department of CSE, Anna University, Chennai- 600025, India.

[2] Professor & Head, Department of CSE, Jerusalem College of Engineering, Chennai- 600100, India.

Correspondence should be addressed to A.Naresh Kumar; nareshkumar5884@gmail.com

Abstract

The aim of this paper is to develop a system that involves translation of one of the Ancient south Indian language to modern language using images of stone inscriptions from several places for the usage of various applications. The technology makes users to understand the translation of ancient language characters to modern language characters with valuable information that can be used in various applications such as Education, Architecture, and Medical science. As many researchers have implemented various algorithms and techniques for ancient character recognition in different languages. Even though Ancient characters conversion still poses a big challenge, because character recognition technology has reached near perfection when it comes to scanning English and other language text. But conversion of Ancient language character to modern language character with high levels of accuracy is still elusive. Only few people are familiar with the ancient characters and make attempts to convert them into modern language characters digitally. The proposed system overcomes such a situation by converting all the ancient historical documents from stone inscriptions into modern language text format. Our algorithm comprises different stages: i) Image Acquisition ii) Pre-processing iii) Advanced Maximally Stable Extremal Regions and iv) Language Translation. The accuracy of the character recognition is 95.59% which is implemented by our algorithm.

1 Introduction

Pattern recognition theory which plays an important role in realizing automation of inputting character involves major role in ancient language character recognition of research and application. A significant amount of research has been carried out in the direction of reading inscriptions from monuments around the world. For the detection of text, localization and extraction of text from images of inscriptions numerous methods have been proposed.

Ancient Tamil character recognition from stone inscription is a challenging problem in pattern recognition area. The obscurity is mainly caused by the large variations of individual writing style. Robust feature extraction is very important to improve the performance of Ancient Tamil character recognition system. The commercially available Optical Character

Recognition (OCR) has very poor recognition accuracy of images of the inscriptions on monuments.

The images of stone inscriptions were passed through the commercial OCR for text extraction, but the OCR failed to recognize these images. These images can be recognized by OCR only after proper enrichment. Fast ICA based enhancement method has given good results for inscription images with a reasonable color difference between text and the background. Most of the ancient inscriptions do not have such reasonable color distinction between the two regions. Therefore to digitize such inscriptions we have to develop the difference between the two regions. This paper proposes a method to develop the minimal difference between text and non text regions of such inscription images.

Figure 1 shows the Development of Ancient Tamil Characters from history. Historic inscriptions are an outlook of the past and an essential part of social, economical and scientific studies. However, pertaining to various factors such as environmental change and human intervention, the quality of such inscriptions degrade with the passage of time. Therefore, there is a need to digitize these inscriptions in form of images, which should be free of any noise and unwanted background information and further, store such images in a digital library so that their retrieval with related information could be performed efficiently in real time.

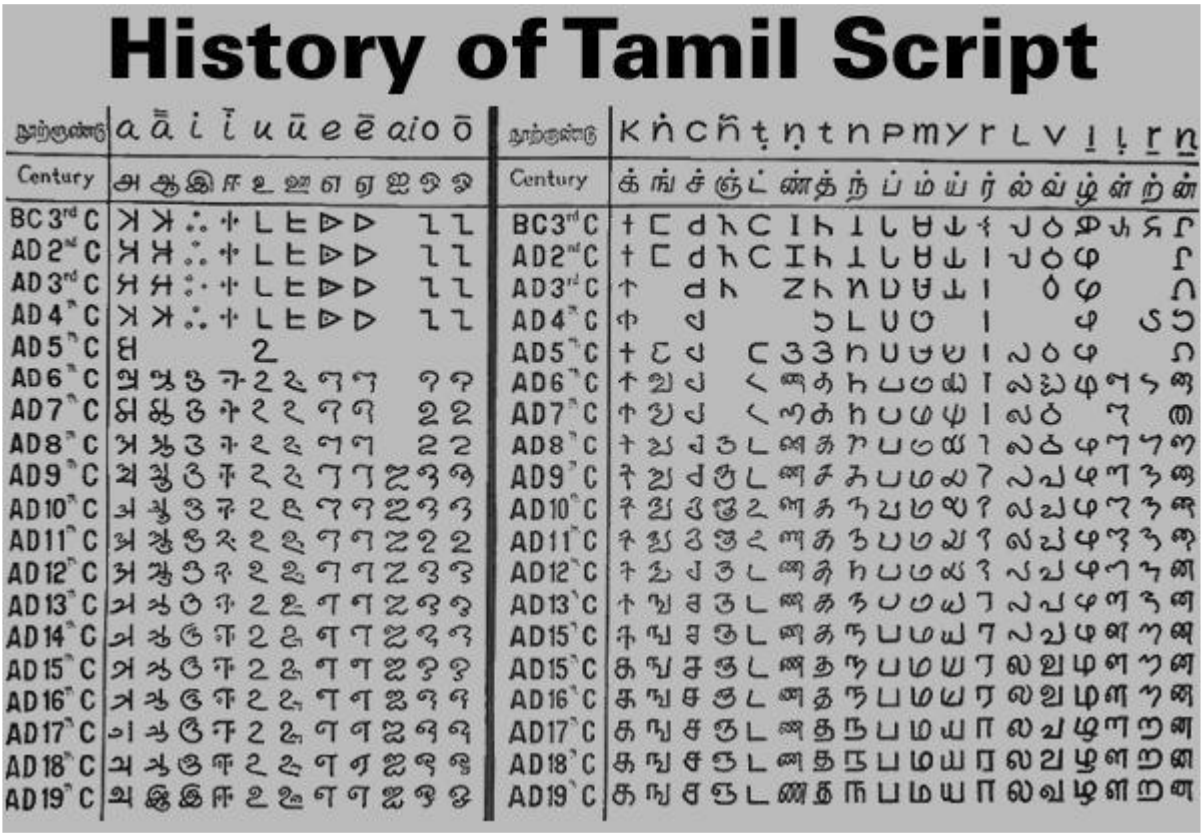

Figure 1: Development of Ancient Tamil Characters

We propose to improve the accuracy of character from images by detecting extremal region and characteristic function of the region using AMSER algorithm. When compared with existing methods, the proposed method is effective in terms of recall and precision.

2 Related Works

The character recognition from stone inscriptions has been done by three major activities: Image Acquisition, Pre-processing and Character Recognition. The basis of various activities discussed and image enhancement system is proposed by combining Modified Fuzzy Entropy-based Adaptive Thresholding (MFEAT) with degree of Gaussian membership function and iterative bilateral filter (IBF) [1].

The fuzzy system [3] helps predicting uncertainty among the character and the background pixels. The use of Shape and Hough transform for feature extraction using Group Search Optimization and Firefly algorithm for feature selection to recognize the ancient Tamil script.

Character recognition is the predominant technique that includes stages of preprocessing [2] the script images and undergo machine learning to recognize individual character from them and to analyze the potentials for developing a sys-tem for developing a system Tamil character recognition system.

The division of the content from the recorded debased picture reports is an extremely difficult undertaking on the grounds that the varieties between the closer view content and foundation content are difficult to decide. [4] Paper proposed canny edge detection algorithm to examine and remove the words from a corrupted picture; the procedure conveyed here is the picture contrast which is adaptively found to settle the issue.

Revealing the content of these inscriptions can be highly valuable to investigate the history of ancient Sri Lanka. Currently the content of each of these inscriptions are translated to modern Sinhala language manually by an archaeology expert who has specialized knowledge to understand the ancient scripts. So the Inscriptions letters read through human eye with great effort and this manual procedure would be time consuming and generate some uncertain outcomes due to lack of consistency [5].

[6] Our proposed system overcomes 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. [8] Our algorithm comprises different stages: i) image pre-processing, ii) feature extraction, iii) character recognition and iv) digital text conversion.

Binarization of degraded document images has been a challenge to many computer scientists. [9] This paper primarily deals with binarization of the palm-leaf manuscript images using Niblack, Sauvola, NICK and Bradley binarization algorithms. Binarization not only reduces the size, thereby protecting the textual content, but also helps in next level of character recognition which includes the segmentation.

The most important necessity in image processing lies in the conversion of ancient Tamil characters to modern text. The ancient Tamil stone inscriptions are the source for these ancient Tamil characters. [10] Analysing and recognising the ancient Tamil characters from the scripts called inscriptions is a difficult task for the present generation who learn to educate, read and write only through the modern Tamil characters. Learning the elaboration of Modern Tamil from ancient Tamil inscriptions is a time devouring process and thus we need a recognition system that helps to teach, understand and also to research the ancient cultures and heritages.

To design a good comprehend and better analysing system [10] we propose a method called noise removal which is also called as pre-processing which removes the entire disturbance in the input image. Next we use a method called morphological operation to perform dilation and erosion operations. Next step is to use the connected component to find the letters which are present in the binary image. Finally we will segment each and every character and match it with our current Tamil language using a process called correlation matching to produce the result of matching characters.

3 Proposed Methodology

The architecture in figure 2 shows the proposed system for the Ancient character recognition and Language Translation.

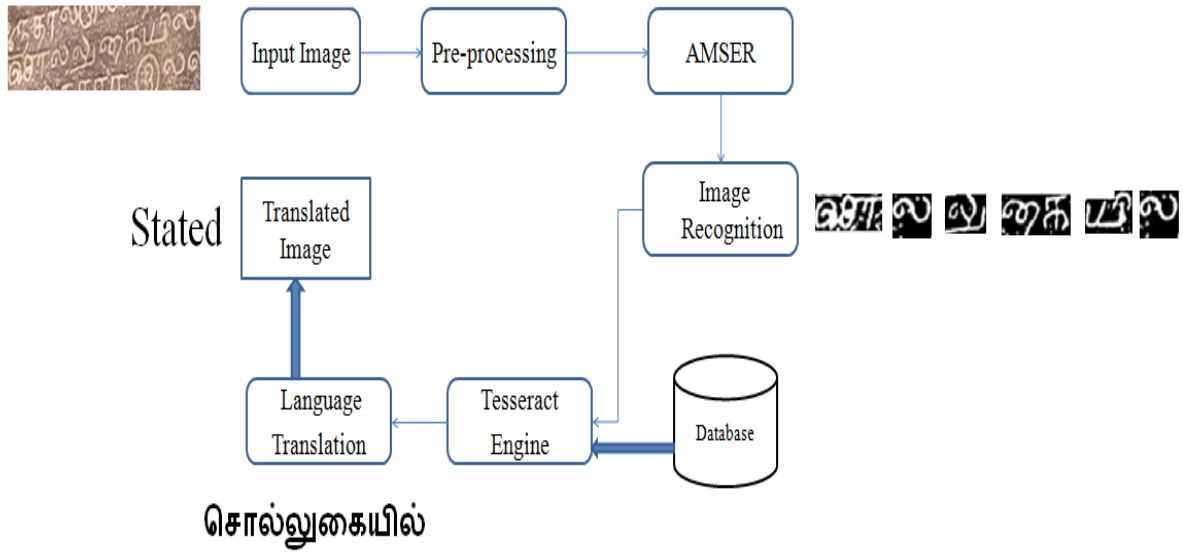

Figure 2: General Architecture for Character recognition and Language translation

3.1 Input Image

Camera – captured image is taken as input in this proposed method. The inscriptions in Garbarakshambigai temple, shown in figure 5, were engraved from King Raja Raja Chola's period who ruled between 985 and 1014 CE* as well as those from Parantaka Chola's period (early 10th century). The walls of the temple consist of 31 inscriptions dating from the Chola period from the reigns of KoperiVarman, Raja Raja I, RajendraChola I and

KulothungaChola. These inscriptions are captured by a Canon EOS-600D Digital SLR Camera.

3.2 Pre-processing

The pre-processing is a development of the input image data that reduce unwanted distortions or enhances some image features important for further processing by using following techniques:

i) Image Restoration

Restoration is a process to recover a degraded image by using a clear knowledge of the degrading phenomenon. Restoration is a function to remove blur images by using de-blurring technique, which is to remove blurring artifacts from images, such as blur caused by defocus aberration or motion blur. The blur is typically model as the convolution of a point spread function with a hypothetical sharp input image, where both the sharp input image and the point spread function are unknown.

ii) Geometric Transformation

A geometric transformation of an image coordinate system is the spatial transformation of an image. In a spatial transformation each points (a, b) of image A is mapped to a point (x, y) in a new coordinate system. The transformation equations are either known in advance or can be determined from known innovative and transformed images. Several pixels in both images with known association are used to derive the unknown transformation. The coordinates of input image are transformed into coordinates of output image using a transformation function.

iii) Pixel Brightness Transformation

Brightness transformations amend pixel brightness and the transformation depends on the properties of a pixel itself. Brightness correction considers original brightness pixel spot in the image. Gray scale transformations change brightness without observe to position in the image.

3.3 Advanced Maximally Stable Extremal Regions

Figure 3 shows an Architecture diagram for the character recognition using AMSER algorithm. It has been divided into five categories:

i) Region Detectors
ii) Affine Invariant Intensity Extrema Based
iii) AMSER Processing
iv) Sweeping Image Thresholds
v) Character Segmentation

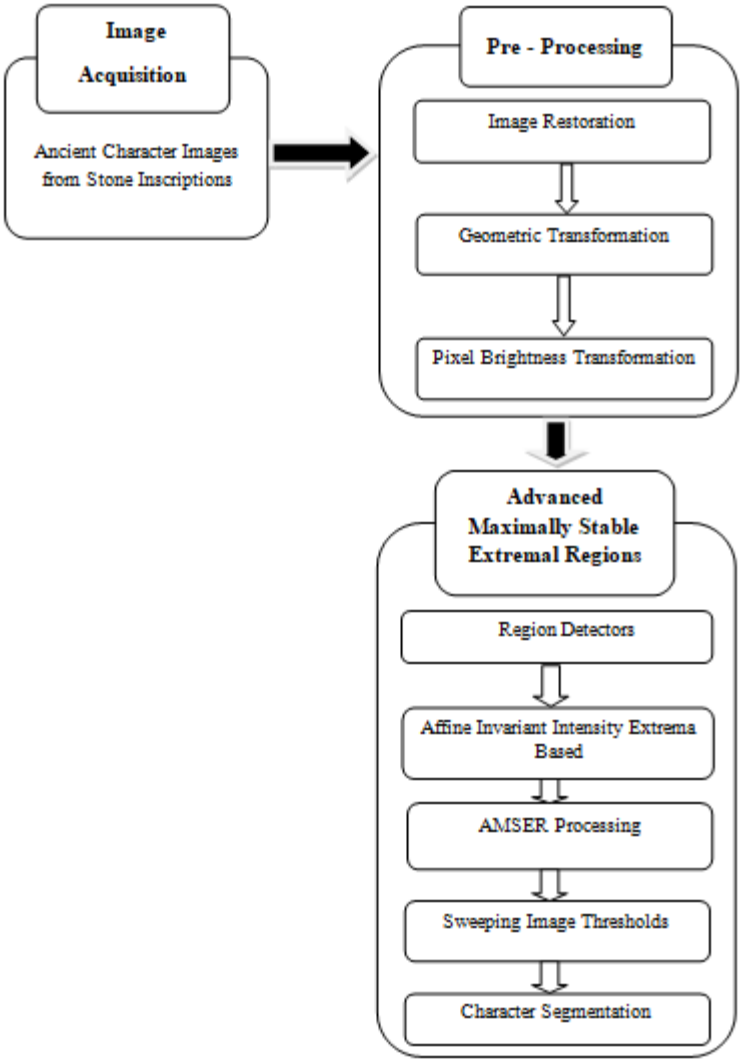

Figure 3: Architecture for Character recognition using AMSER Algorithm

i) Region Detectors

Region Detectors a set of regions was defined that is preserved under a broad class of geometric and photometric changes and yet has the same cardinality.

ii) Affine Invariant Intensity Extrema Based

Affine Invariant Intensity Extrema Based defined by receiving input image (I) through Advanced Maximally Stable Extremal Regions AMSER; it is used to detect extremal regions for each and every point along the ray.

Detecting extremal regions:

Detect anchor points (f.e. using Harris detector for corners). Anchor points detected at multiple scales are local extremas of intensity – explore image around rays from each anchor point. Go along every ray starting from this point until an extremum of function f is reached.

The Characteristic function of the region can be calculated by f (t), by using intensity as shown in Figure 4 AMSER – Affine Invariant Intensity Extreme Based to approximate the region from images. The ellipse of an affine-transformed region corresponds to the ellipse of the extremal local and global region under the same transformation.

- f characteristic function of the region (1 inside, 0 outside)

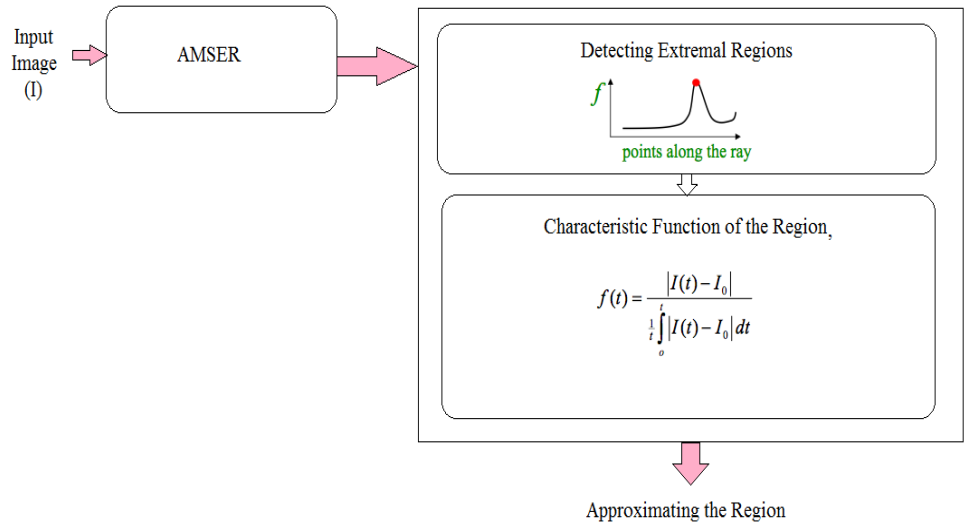

Figure 4: AMSER – Affine Invariant Intensity Extreme Based

All points create some irregularly-shaped region. Approximately corresponding regions are obtained for affine-transformed regions

iii) AMSER Processing

AMSER is a method for clump detection in images. The AMSER algorithm extracts from an image a number of co-variant regions, called AMSERs: an AMSER is a stable connected component of some gray-level sets of the image.

The AMSER extraction implements the following steps:

Step i) Sweep threshold of intensity for global minimum from black to white, performing a simple minimal luminance threshold values of the image.

Step ii) Extract connected components. (“Wide Extremal Regions”)
Find a threshold when a wide extremal region is “Maximally Stable”, i.e. global minimum of the relative growth of its square. Due to the discrete nature of the image, the wide Extremal region below / above may be coincident with the actual region, in which case the region is still deemed maximal.

Step iii) Approximate a region with an ellipse. Keep those regions descriptors as features.

iv) Sweeping Image Thresholds

Sweeping the threshold value from low to high to extract the entire image which is received from AMSER for approximating the region. Apply a series of thresholds – one for each grayscale level. Threshold the image at each level to create a series of black and white images.

v) Character Segmentation

Character segmentation is an action that seeks to crumble an image of a sequence of characters into sub-images of individual symbols. It may be one of the decision processes in a system for optical character recognition (OCR).

3.4 Image Recognition

Image recognized from AMSER, which shows the images into several characters based on the bag of words from the database.
After preprocessing of image document, OCR technology can be used to read and translate characters. Mainly Character Recognition consist of the basic stages to compare the characters in the scanned image to the characters in the learned set involving,

- The method of extraction and isolation of each character from an image
- Determined the properties of extracted character.
- Evaluation of the properties of extracted character and learned character.

3.5 Tesseract Engine

Tesseract – OCR is the most widely used open source OCR athwart the world. It supports English language as default and few more language and it is a command line tool. Tesseract OCR Engine has suppleness that it can be trained to any language. In Tesseract engine, the image will be received from Character Segmentation and by using image to string code will return the result of a Tesseract OCR run on the image to string and the converted string will be matched with data set of database. So that engine will convert the ancient language character to naive language character for language translation.

3.6 Language Translation

Language translation plays major role by converting one language to another while maintaining the original document format.

3.7 Translated Image

Image translation refers to an added service provided by translation applications where the languages from the stone inscriptions, name board, sign board etc will be displayed.

3.8 Inscription Application

The inscriptions record the grants of the temple by the kings. The inscriptions are found on the walls of the temple, sanctum and also in the halls. The inscriptions are recorded in numbers 100, 102 and 110 in South Indian Inscriptions.

3.9 Proposed AMSER Algorithm

Image I is a mapping I: $\mathcal{D} \subset N^2 \rightarrow S$. Extremal regions are well defined on images if:

1. S is totally ordered, i.e. reflexive, anti symmetric and transitive binary relation $\leq$ exists. In this paper only S = {0, 1,…, 255} is considered, but extremal regions can be defined on e.g. real-valued images (S = R).
2. An adjacency (neighbourhood) relation $A \subset \mathcal{D} \times \mathcal{D}$ is defined. In this paper 4-neighbourhoods are used, i.e. p,q $\in \mathcal{D}$ are adjacent (pAq) iff $\sum_{i=1}^{d} |pi - qi| \leq 1$.

Region $\mathcal{Q}$ is a contiguous subset of $\mathcal{D}$, i.e. For each p, q $\in \mathcal{Q}$ there is a sequence *p*,a1,a2,...,a*n*,*q* and *p*Aa_1, a_iAa_{i+1}, a_nA$_{\boldsymbol{q}}$.

(Outer) Region Boundary $\mathbf{o}\,\mathcal{Q}$ = {$q \in D \setminus \mathcal{Q}$: $\exists\ p \in \mathcal{Q}$: qAp}, i.e. the boundary $\mathbf{o}\,\mathcal{Q}$ of $\mathcal{Q}$ is the set of pixels being adjacent to at least one pixel of $\mathcal{Q}$ but not belonging to $\mathcal{Q}$.

Extremal Region $\mathcal{Q} \subset \mathcal{D}$ is a region such that for all $p \in \mathcal{Q}$, $q \in \mathbf{o}\,\mathcal{Q}$: $I(p) > I(q)$ (Maximum intensity region) or $I(p) < I(q)$ (minimum intensity region).

Advanced Maximally Stable Extremal Region (AMSER). Let $\mathcal{Q}_1,...,\mathcal{Q}_{i-1}, \mathcal{Q}_i$ be a sequence of nested extremal regions, i.e. $\mathcal{Q}_i \subset \mathcal{Q}_{i+1}$. Extremal region $\mathcal{Q}_{i^*}$ is maximally stable iff $q(i) = |\mathcal{Q}_{i+\Delta} \setminus \mathcal{Q}_{i-\Delta}| / |\mathcal{Q}_i|$ has a global minimum at i* (|.| denotes cardinality) . $\Delta \in$ S is a parameter of the method.

4 Results and Discussion

Digitally acquired stone inscription images from Garbarakshambigai temple are taken as input images, which belong to the tenth century, because the unique features and the style of writing of Tamil characters have their evolution in the tenth century and got standardized by the nineteenth century. All the inscriptions in the temple are captured using a digital camera. Clear scripts which are not deteriorated and focused in good light illumination using phone camera can also be handled. The stone inscription has hard surface, and hence light illumination will be the only parameter which differentiates the texture of the character.

Several images were handled in this work, which is subdivided into few categories, each containing images based on light illuminations.

Figure 5: Input Image

Figure 6: MSER Method

Figure 7: MFEAT Method

Figure 8: AMSER Method

The proposed AMSER algorithm outperformed other methods such as MSER and MFEAT as shown in figure 6, 7 & 8. Even though the captured images vary in light illumination and the color of the stone, our proposed method binarizes and extracts the foreground character, which leads to a good recognition rate. The translation of Ancient tamil character to modern tamil character is done using Tesseract engine which is shown in figure 9.

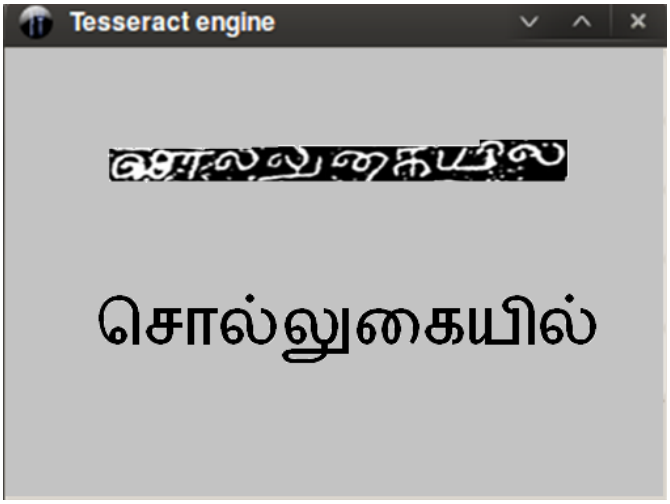

Figure 9: Translation of Ancient tamil character to modern tamil character

Figure 10 shows the Translation of Modern tamil character to other language which is done using tamil.changathi web services.

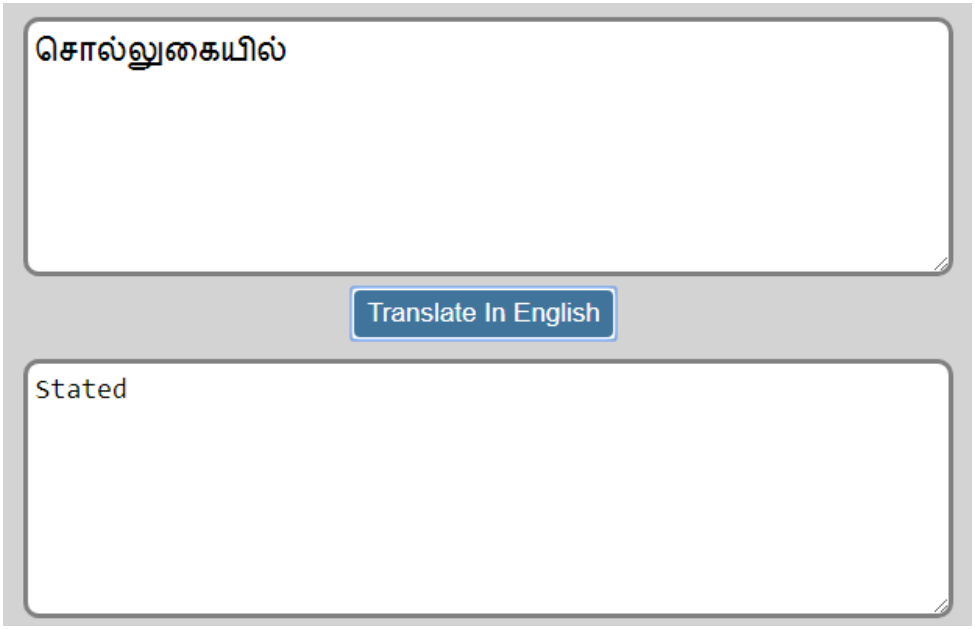

Figure 10: Translation of Modern tamil character to other language

5 Performance metrics analysis

The performance analysis for the Comparison graph for the proposed algorithm using various methods showed in figure 11 with PSNR in db.

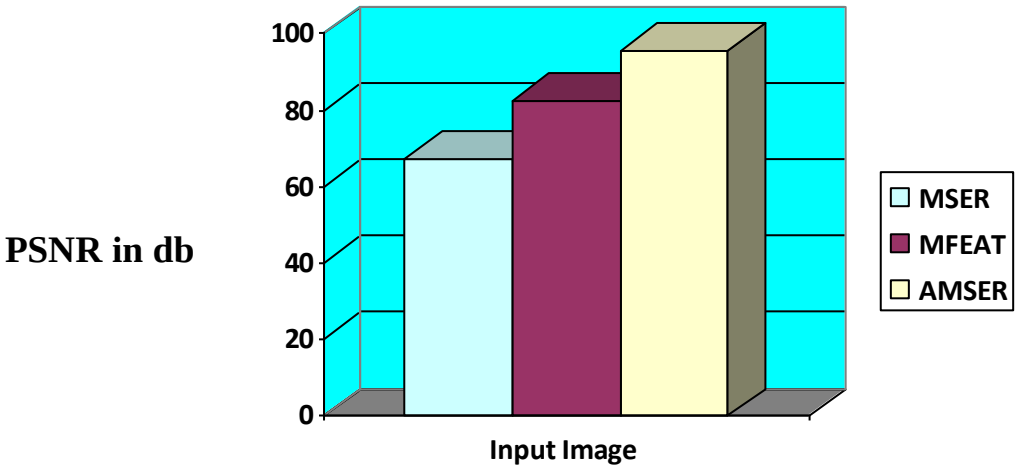

Figure 11: Comparison graph for the proposed algorithm with various methods

6 Conclusion

In this paper, we mainly focus on application oriented research from stone inscriptions by implementing algorithm for the process of segmenting each and every Tamil letter in the stone inscription we captured and matching it with the Tamil template and finding what the letter is according to modern Tamil language as well as for other language translation.

7 Future works

Light illumination differentiates the character well based on the inscribed depth on flat stone background. Since no dominating colors have been used for character, the major key to differentiate the character is light illumination by considering depth as a parameter. The other geometric features such as edges and blobs can be included in our future work.

This paper does not concentrate on deteriorated stone inscriptions. With this, the future work can be concentrated on segmentation, feature extraction, and efficient machine learning algorithm for recognizing which character it is.

References

[1] K. Durga Devi and P. Uma Maheswari, "Digital acquisition and character extraction from stone inscription images using modified fuzzy entropy-based adaptive thresholding," *Springer*, vol. 22, no. 4177, 2018.

[2] K. Durga Devi and P. Uma Maheswari, "Insight on Character Recognition for Calligraphy Digitization," *2017 IEEE International Conference of Technological Innovations in ICT for Agricultural and Rural Department (TIAR 2017)*, vol., no. pp. 78-83, 2017.

[3] T S Suganya, Dr. S Murugavalli, "Feature Selection for an Automated Ancient Tamil Script Classification System Using Machine Learning Techniques," *2017 IEEE International Conference on Algorithms, Methodology, Models and Applications in Emerging Technologies (ICAMMAET)*, 2017.

[4] P. Selvakumar, Dr. S. Hari Ganesh , "Tamil Character Recognition using Canny Edge Detection Algorithm," *2017 IEEE World Congress on Computing and Communication Technologies (WCCCT)*, vol., no. pp. 250-254, 2017.

[5] D.A.S. Ruwanmini, K.V. Liyanage , K.G.N.D. Karunarathne , et al., "An Architecture for An Inscription Recognition System for Sinhala Epigraphy," *International Journal of Research - GRANTHAALAYAH* vol. 4, no.12, pp. 48 – 64, 2016.

[6] E. K. Vellingiriraj, M. Balamurugan, and P. Balasubramanie "Text Analysis and Information Retrieval of Historical Tamil Ancient Documents Using Machine Translation in Image Zoning," *International Journal of Languages, Literature and Linguistics* vol. 2, no. 4, pp 164 – 168, 2016.

[7] Lisa Madlberger, Ade Romadhony, Ayu Purwarianti, "Gotong Royong in NLP Research a Mobile Tool for Collaborative Text Annotation in Indonesia," *2016 IEEE International Conference on Asian Language Processing (IALP)*, pp. 99 – 102, 2016.

[8] E. K. Vellingiriraj, M. Balamurugan, and P. Balasubramanie, "Information Extraction and Text Mining of Ancient Vattezhuthu Characters in Historical Documents Using Image Zoning," *2016 IEEE International Conference on Asian Language Processing (IALP)*, pp 37 – 40, 2016

[9] Sabeenian R S, Paramasivam M E and Dinesh P M, "Appraisal of Localized Binarization Methods on Tamil Palm-leaf Manuscripts," *2016 IEEE International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, pp 793 – 797, 2016.

[10] Giridharan.R, E. K. Vellingiriraj, and P. Balasubramanie, "Identification of Tamil Ancient Characters and Information Retrieval from Temple Epigraphy Using Image Zoning," *2016 IEEE International Conference on Recent Trends in Information Technology (ICRTIT)*, 2016.

[11] G. Janani, V. Vishalini, Dr .P. Mohan Kumar, "Recognition and Analysis of Tamil Inscriptions And Mapping Using Image Processing Techniques," *2016 IEEE International Conference on Science Technology Engineering and Management (ICONSTEM)*, pp 181 – 184, 2016.

[12] Kavitha Subramani , Dr S.Murugavalli, " A Novel Binarization Method For Degraded Tamil Palm Leaf Images," *2016 IEEE International Conference on Advanced Computing (ICoAC)*, pp 176 – 181, 2016.

[13] A.S. Kavitha, P. Shivakumara, G.H. Kumar, et al., "Text segmentation in degraded historical document Images," *Egyptian Informatics Journal (2016)*.

[14] Mrs.G.Bhuvaneswari, Dr.V.Subbiah Bharathi, "An Efficient Positional Algorithm for Recognition of Ancient Stone Inscription Characters," *2015 IEEE International Conference on Advanced Computing (ICoAC)*, pp 1 – 5, 2015.

[15] Shashaank M. Aswatha, Ananth Nath Talla, Jayanta Mukhopadhyay, et al., "A Method for Extracting Text from Stone Inscriptions using Character spotting" *Asian Conference on Computer Vision*, vol 9009. Springer, Cham, 2015.

[16] Chamila Liyanage, Thilini Nadungodage, Ruvan Weerasinghe, "Developing a commercial grade Tamil OCR for recognizing font and size independent text," *2015 IEEE International Conference on Advances in ICT for Emerging Regions (ICTer)*, 2015.

[17] H.S. Mohana, Navya.K, Rajithkumar.B.K , et al., "Interactive Segmentation for Character Extraction in Stone Inscriptions," *2014 IEEE International Conference on Current Trends in Engineering and Technology, ICCTET'14*, pp 321 – 327, 2014.

[18] R. Angelin Jeniffer, G. Bhuvaneswari, "Image Glazing for Thinning of Ancient Tamil Characters,"

International Journal of Scientific & Engineering Research, vol 5, iss 6, pp 1 - , 2014.

[19] Vivek Hanumante, Rubi Debnath, Disha Bhattacharjee, Deepti Tripathi, et al., "English Text to Multilingual Speech Translator Using Android," 2014 *International Journal of Inventive Engineering and Sciences (IJIES)*, vol-2, iss-5, pp 4-9, 2014

[20] M. V. Jeya Greeba, G.Bhuvaneswari, "Recognition of Ancient Tamil Characters in Stone Inscription Using Improved Feature Extraction," 2014 *International Journal of Recent Development in Engineering and Technology, ICMACE14*, vol 2, iss 3, pp 38 - 41 2014.

[21] T S Suganya, Dr. S Murugavalli, "Binarization of Ancient Tamil Scripts From Stone Inscriptions," 2014 *GLOBAL JOURNAL FOR RESEARCH ANALYSIS (GFRA)*, vol. 3, no. 11, pp. 72-744, 2014.

[22] Indu Sreedevi, Rishi Pandey, N. Jayanthi, et al., "NGFICA Based Digitization of Historic Inscription Images," *ISRN Signal Processing*, vol. 2013, ID. 735857, pp. 1 – 7, 2013.

[23] S. RajaKumar, Dr. V. Subbiah Bharathi, "Ancient Tamil Script Recognition from Stone Inscriptions Using Slant Removal Method," *International Conference on Electrical, Electronics and Biomedical Engineering (ICEEBE'2012) Penang (Malaysia)*, pp. 80 – 84, 2012.

[24] Rajakumar S, Dr.Subbiah Bharathi V, "7th Century Ancient Tamil Character Recognition From Temple Wall Inscriptions," *Indian Journal of Computer Science and Engineering (IJCSE)*, vol. 3, no.5, pp 673 – 677 Oct - Nov 2012.

[25] Nick Ceramella, "Is English a Killer Language or an International Auxiliary? Its Use and Function in a Globalised World," *International Journal of Language, Translation and Intercultural Communication (IJLTIC)*, vol. 1, pp 8 – 23, 2012.

[26] S. RajaKumar, Dr. V. Subbiah Bharathi, "Century Identification and Recognition of Ancient Tamil Character Recognition," *International Journal of Computer Applications*, vol. 26, pp 32 – 35, 2011.

[27] www.tnarch.gov.in

[28] http://www.agamaacademy.org/digital-library-en.php

[29] http://www.instituteofasianstudies.com/

[30] http://www.agamaacademy.org/digital-library-book/Kamika-Agamam

[31] http://thiruvenkidapuram.blogspot.com/2016/01/sri-garbharakshambikasameta-shri.html

[32] http://tamil.changathi.com