Original Article

A Survey on Recognition of Ancient Tamil Brahmi Characters from Epigraphy

A. Vidhyavani[1], Dr.T. Manoranjitham[2]

[1]Research Scholar, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India. E-mail: vidhyava@srmist.edu.in

[2]Assistant Professor (S.R), SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India. E-mail: manorant@srmist.edu.in

Abstract

The system which involves in character identification of Brahmi letters from epigraphy and converts to current Tamil character format. Identification of earliest Tamil characters is one of the hardest part. If the letters are on the walls, it is still more difficulty in recognizing the characters. Character recognition has reached near perfection in English and other language text. Identification of epigraphy brahmi characters is very difficult because brahmi characters has been used from 3rd Century BCE to 4th Century CE. The accuracy level is not high in Brahmi letters because of these reasons- First the consonants are similar to consonantal vowel characters, second the author doesn't inscribe the Brahmi letterings correctly and the fonts are inscribed in various styles and strokes. Only few people are known with ancient character if this remains, all the valuable data given by our forefather will not be identified by the future generation. In this survey paper, we describe and compare many techniques of identification through table.

Keywords: ICR, OCRs, NLP.

DOI: 10.47750/pnr.2022.13.S03.199

INTRODUCTION

Tamil is one of the eldest language in earth with rich poetry. In olden days the poet in Tamilnadu used to write in palm leaves, stones and inscriptions. Tholkappiyam which is written during 4th BC is the best example for palm leaf inscription which is the history of grammar book in Tamil. The ancient inscriptions contains valuable commentaries, some classics like saivam, vaishnavam, medical commentaries, food, numerology, astrology, music, siddha, dance and many. Palm manuscripts mainly used for grammer, science, astrology, land registration which is donated by the king, historical, places.

Old Tamil originated from Tamil brahmi script. Tamil script has been identified in wide range of area. Tamil Brahmi is the source for old Tamil. The prediction of Tamil brahmi has been in use from 3rd century BCE. But the recent days evidences shows that it has been still back from 5th century BCE, but it do not have majority acceptance. Tamil-brahmi had been in usage for many eras and gets adapted then gets mutated into Vaṭṭeḻuttu from 5th Century CE.

The current Tamil is not directly derived from Tamil-Brahmi. The current Tamil has been in vogue from 7th century Common era which is derived from Pallava grantha with mixture of Vaṭṭeḻuttu. Upto 11th century CE Vaṭṭeḻuttu has been used in Tamil Nadu.

Categories of Tamil-Brahmi

Tamil-Brahmi has two types of development:

Early (Tamil Brahmi)

Early Tamil Brahmi has been used from 3rd Century BCE to 1st Century CE. TB I and TB II vowel notational systems were in use.

Late (Tamil Brahmi)

It has been started from 2nd to 4th Century CE. The forms of the characters gradually became more cursive which gets converted into initial Vaṭṭeḻuttu characters. The schemes TBII and TBIII has been used.

Character recognition is the very complex work in pattern recocgnization system, because the researcher should know to separate the characters, and to identify the various fonts with different style, categorize the character of same shape and size.

OCR and ICR

OCR and ICR overall process looks same, but there is a main differences between the two systems. OCR which is mainly to converts scanned images of text, even printed or typewritten, and translates these photographs into machine-encoded text. It is used as a file-keeping system for businesses purpose and also used to post text through online.

OCR is usually used to convert books or any large documents into electronic files.

Whereas ICR technically resembles as OCR but it is very specific. An ICR is a method which is mainly to learn varies fonts and styles of handwritten characters. Using ICR, a system can read handwriting and it is also recognize it to increase accuracy and recognition. ICR is technically smarter than OCR and also very detailed and more involved.

ICR is a subclass of OCR software, where the OCR software is not set up to identify handwritten characters. Main difference is that OCR is used for printed documents which has been typed and translate into text and OCR text facilitates copy paste. ICR focuses mainly on handwriting materials which has most difficult fonts than OCRs.

Literature Survey

[1] The ancient history is being studied using the resource of inscription and principles of development in the nation. Recognizing and converting the early brahmi letters from the temple epigraphy is one of the hardest part for current age group. Brahmi letters to Sinhala language automation is not there. So manually they are translating the inscriptions. The inscription is on the rock wall it is taking much time to covert the character into sinhala letters. This research major attentions on identification of early Brahmi letters inscribed between 3rd BC and 1st AD the time period. First, it eliminate the sound, sector the characters from the epigraphy pictures are taken and transform into binary format using Image Processing techniques. Then, it identify the broken letters and convert that into correct Brahmi letters then recognize the century of the inscriptions using CNN in machine learning. At last, the Brahmi characters are coverted into Sinhala characters with the meaning of the letters using NLP.

The result of InceptionV3 model accuracy is 55.71. DenseNet121 model accuracy is 60% and damage is 2.41. Whereas the damage value of test set was greater than other model. Exceptions model accuracy is 46% and the damage is 1.21. ReseNet50 model accuracy is 55.00% and the damage is 1.97. VGG19 model accuracy is 50.12% and the damage is 1.25. VGG16 model accuracy is 52.86% and the damage is 1.72.

[2] It is a method for recognition of both handwritten and printed Brahmi characters which involve preprocessing, segmentation, feature extraction, and classification of Brahmi script characters. The geometric method was used for feature extraction into six different entities, followed by a newly developed classification rules to recognize the Brahmi characters based on the features. The method obtains accuracy of 91.69% and 89.55% for handwritten vowels and consonants character respectively and 93.30% and 94.90% for printed vowel and consonants character respectively.

Cropping, thresholding, and thinning method were used in the preprocessing, Line detection and character detection method for segmentation before implementing feature extraction and classify the characters. The accuracy of this method is 94.10% and 90.62% for printed and handwritten Brahmi character recognition respectively. As a whole, this method offers a satisfactory success rate but the results could be further improved by using NN and SVM techniques for classification of the Brahmi characters.

[3] Early Sri Lankan information's are written on stone surface are generally polluted with huge clatters such as cracks, scuffs, spaces, etc. The ancient letters can be identified using precise method of alphabet fonts that tends to automatic reading of ancient manuscripts by computers.

This research work involved in automatic identification of early Brahmi inscriptions by computer. Based on time period the shape of the letter gets changed. Also in a same period of time, one letter may get changed slightly. In this research, it established automatic identification of ancient letters. They proposed Modified correlation function method has been which is more sophisticated than the previous correlation peak method. Digital repository of Sri Lankan inscriptions is also trying to produce which has a role of automatic reading of letters by computers.

[4] In this research twelve vowels from the palm leaf manuscripts has been identified using B-spline curves. The uniqueness and robustness is the main advantage in B-spline curve. Each and every vowels in the Tamil character has more than one curves with various angle. The vowels can be identified by the combination of curves. Various narrators inscription of same letter has been identified using this B-spline technique with the high level of accuracy than any other method.

B-spline curve method consist of three points P1, P2, P3 Where P1 is the first point, P2 is the middle point and P3 is a variable point. Then the variable point and the first point has been connected and find the distance, then remove the point P3, when it is less than a threshold value. Repeat until all the middle points are changed into P3.

The data set are collected from Aagama academy it is executed in Python 3.0 and the different archives are tested for the performance of vowel characters. It works based on the neural network with the support of B-spline curve [5] It involves in identification of Tamil letters from 9th to 12th centuries First it pre-process the image then segmentation. Based on threshold value the segmentation involves in converting the images into binary image. Then the extraction has been done by Scale Invariant Feature Transform algorithms for each character to identify the correct letter.

Support Vector Machine classifier classifies the letters and predicted by Trigram technique. Each recognized letter is allocated with Unicode value and stored in the image warehouse

It takes 50 pictures of each era. Othu's Thresholding has been used for identification of character and Contorlet Transformation method for Recognition of characters.

[6] In this research the text can be identified using RTI in which it has the characteristics of surface reflection. It choose one view over the object and it can identify the surface based

on different lightings. Here the dome structure for RTI acquisition was done by 116 computers with controlled lights and a digital camera contains a stand. Every single set of captured images were processed to generate the RTI image.

In Cultural Heritage Imaging website RTI Builder open source software is available. It can be worked remotely using web interface system.

Markout has been developed for RTI visualization tool, since for recognition and tracing phase of RTI visualization it is needed.

[7] In order to maximize the speed and decrease the fault, adaptive backpropagation learning model were proposed. Network output contains error values and in each Iteration hidden layers in the network is adapted. The change in values and weights is noted in order to give efficient and less computation time.

For identifing Tamil Letters ABP Model has been used 12 vowel characters are selected. From that, 8 characters are selected for training and 4 characters are selected for testing. C- Programming language has been used for coding and verification. The character are taken as image format Then it is digitized as binary format of 0 and 1 for the input. It is segmented into 48 grids From 48 nodes, 8 were taken as hidden nodes and 4 were taken as output nodes, which is like 48-8-4.

For training the network model all the nodes has been used which is also called epoch Mean squared error is used for validation which can be identified by division of sum of squared linear error in single node by double the maximum number of node. The range of Network weights is between -3 to +3. Then the performance is tested with BP based on that Static parameter values and Elimination conditions was determined. C surve measurement indicates the union of ABP model.

[8] This research is to recognize hand written Devanagari letters by means of Deep Convolution Neural Network with transfer learning. It has been implemented in DenseNet, Vgg, AlexNet and Inception.

Inception V3 and ConvNet. Inception V3 gives 99 % accurate result than any other methods in 16.3 minutes and AlexNet performs with the accuracy of 98% at high speed of 2.2 minutes for sinlge epoch with a limited amount of data training the CNN model is very difficult task.

So that, transfer learning has been used for solving these kind of problem. This transfer learning which transfer its knowledge to small data and gives accurate result. So CNN take the same data for the beginning convolutional layers of the network and it is going to train only the last few layers.

[9] This research involves in digitizing the ancient Tamil characters. For feature Extraction it uses Shape and Hough transform. Group Search Optimization, Firefly algorithm is for feature selection. These algorithm is used for recognizing the ancient Tamil script.

Shape Transformation

In shape transformation algorithm the transformation of character from image A to B has been done by

1) Some pixels are substitution.
2) Some pixels are moved to the closest pixels in B, 3) Some pixels are deleted due to deletion, insertion.

Two possible cases occur in this algorithm

i) Pixels remains in A;
ii) Pixels which is not connected remain in B.
iii) Pixels which is left in A were deleted. Moving the pixel outside the character frame is eual to the cost of deleting a pixel A.

Defective images are identified using Hough Transform methods. In a parameter space the system of voting is taken place, in which local maxima and accumulator space is obtained. Straight lines can be identified in an image. Based on that broken lines can be identified for feature selection.

[10] This study was created for recognizing cursive handwritten texts. It takes the handwritten image as input. For feature extraction diagonal Feature Extraction is used, Euler Number is used for classification. High accuracy rate is achieved through Euler Number which is combined with Diagonal Feature Extraction.

After conducting a set of tests on Fallaria system, this has been decided this system is more efficient than Fallaria system. In this 100 characters has taken and got character accuracy of 88.78% and word accuracy of 50.4348%.

[11] This paper is based on the identification of special characters and alpha numerals. Intelligence is needed for identifying the characters. OCR works well in printed character but when it comes to handwritten or change in style, font etc it becomes very difficult to identify.

For learning attributes and classification of labels Supervised machine learning is used. By training the machine well it shows accurate result.

Even for large amount of data ICR gives accurate result also. It first classifies and send the data to excel sheet. The Using this method the result shows the accuracy of 95% for special characters and alpha numerals.

[12] This study involves in text identification of Brahmi and Vattezuthu characters from palm leaf and convert into Tamil digital text using the neural network and image zoning method. This algorithm contains image capturing, image preprocessing which consist of 4 process like.

i) Cropping

Usually the brahmi characters are inscribed with space and some noise over the image. Cropping remove the unwanted space and niose.

ii) Segmentation

There are three types

a) line

b) word
c) character

iii) Resizing

Each character is in different size. So it resized all the character in same size.

iv) Image thinning

A character can be converted in thin character.

v) Binarization

In binarization 1's are considered as a Dark pixel and 0's are considered as a Light pixel

Than *Data set training, Character Recognition, Unicode Text, Retrieve from the database is done*. Brahmi script accuracy rate is 91.57%. Vattezhuthu character accuracy is 89.75%.

So the Accuracy of Brahmi character is higher than Vattezhuthu.

AUTHOR	METHOD/TECHNIQUES USED	NUMBER OF CHARACTERS TAKEN FOR IMPLEMENTATION	ACCURACY
K.A.S.A. Nilupuli Wijerathna	CNN in deep learning, NLP	Early Brahmi Characters, late brahmi characters	Pretrained based model VGG 16-93.33% with loss-0.22, Pretrained based model with different time period, DenseNet121s
Neha gowtam	Optical character recognition using geometric method	Vowels and Consonants for handwritten and printed characters	Handwritten-90.62%, Printed character-9%s
Nalin Warnajith	Modified correlation function method, automatic correlation function method	Early brahmi characters	Early brahmi characters-55%
Suganya Athisayamani	B-spline Curve Recognition	Vowel 6 characters	85% accuracy
Manigandan T	**OCR and NLP techniques**	50 Images of each centuries between 9th and 12th century.	72% accuracy
Federico Ponchio	Reflection Transformation Images	20000 characters	90% accuracy with only 10% loss
M. Sornam	***Adaptive** Backpropagation **Learning method***	45 epochs, 50 epochs and 120 epochs	Identified 45 epochs in 56 milli second, 50 epochs in 67 milli second, 120 epochs in 89 milli second
Nagender Aneja	**Transfer learning for Deep Convolution Neural Network**	15 epochs	Inception 99% accuracy in the first epoch at 16.3 minutes Vgg11 99% accuracy in 45.6 minutes AlexNet with 98% accuracyin6.6 minutes with 3 epochs
T.S. Suganya	***Shape and Hough transform, Group Search Optimization, Firefly algorithm***	9 characters each contains 35 samples	J48-93.33, KNN-91.75,NN-93.97
Yosuke R. Matsuoka	*Intelligent Character Recognition* Diagonal Feature Extraction Vertical Feature Extraction	Word and charcters	DFE with euler number-88.7 DFE without euler number-80.4 VFC with euler number-87.2 VFC without euler number-79.1
Renuka Kajale	intelligent character recognition, supervised machine learning	200 samples taken	95% accuracy
E.K.VELLINGIRIRAJ	Image Zoning method has been used	Vowels, Consonants, Consonant vowels of both brahmi and vattezhuthu characters	Brahmi (vowels-93.45%, consonants-92.75%, consonant vowels-90.24%) Vattezhuthu (vowels-91.11%, consonants-90.75%, consonant vowels-89.12%) overall brahmi character-91.57%, vattezhuthu character-89.75%s

CONCLUSION

The conclusion of this paper is, the recognition of characters from handwritten, palmleaf, and epigraphy. Most of the identification is based on image preprocessing, segmentation, classification and final comparison. Optical character recognition plays a major role in identification of characters and deep learning also helps to identify the character. The main aim of this paper is to identify method which gives accurate results from the epigraphy. Where the intelligent

character recognition is mainly for handwritten and which is an advance method of OCR. The ancient characters has been identified in very few places in this literature review transfer learning system very suitable for small dataset which train the limited amount of data and gives accurate result.

REFERENCES

K.A.S.A. Nilupuli Wijerathna, Rashmi Sepalitha, Thuiyadura Indika, Harshana Athauda, P.D. Suranjini, J.A.D.C. Silva, Anuradha Jayakodi, Recognition and translation of Ancient Brahmi Letters using deep learning and NLP, 2019 International Conference on Advancements in Computing (ICAC), 978-1-7281-4170-1/19/$31.00 ©2019 IEEE.

Neha Gautam and Soo See Chai, Optical Character Recognition for Brahmi Script Using Geometric Method, Journal of Telecommunication, Electronic and Computer Engineering, e-ISSN: 2289-8131 Vol. 9 No. 3-11, https://www.researchgate.net/publication/322137737

Nalin Warnajith, Dammi Bandara, Sarkar Barbaq Quarmal, Masanori Itaba, Atsushi Minato and Satoru Ozawa, Computer Analysis of Photographic Data of Sri Lankan Early Brahmi Inscriptions, IOSR Journal of Engineering (IOSRJEN), e-ISSN: 2250-3021, p-ISSN: 2278-8719 Vol. 3, Issue 1 (Jan. 2013), ||V3|| PP 44-49.

Suganya Athisayamani, Dr.A. Robert Singh,_, Dr.T. Athithanc, Recognition of Ancient Tamil Palm Leaf Vowel Characters in Historical Documents using B-spline Curve Recognition, Third International Conference on Computing and Network Communications (CoCoNet'19) Procedia Computer Science 171 (2020) 2302–2309, Published By Elsevier.

Manigandan T, Dr. V.Vidhya, Dr.Dhanalakshmi V, Nirmala B, Tamil Character Recognition from Ancient Epigraphical Inscription using OCR and NLP, International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS-2017), 978-1-5386-1887-5/17/$31.00 ©2017 IEEE.

Federico Ponchio, Marion Lame, Roberto Scopigno, Bruce Robertson, Visualizing and Transcribing Complex Writings through RTI, 978-1-5386-4385-3/18/$31.00 ©2018 IEEE.

M. Sornam, Muthu Subash Kavitha, M. Poornima Devi, An Efficient Morlet Function Based Adaptive Method for Faster Back propagation for Handwritten Character Recognition, 2016 IEEE International Conference on Advances in Computer Applications (ICACA), 978-1-5090-3770-4/16/$31.00©2016 IEEE.

Nagender Aneja_ and Sandhya Aneja, Transfer Learning using CNN for Handwritten Devanagari Character Recognition, 2019 1st International Conf erence on Advances in Information Technology, 978-1-7281-3241-9/19/$31.00 © 2019 IEEE.

TS Suganya, Dr.S Murugavalli, Feature Selection for An Automated Ancient Tamil Script Classification System Using Machine Learning Techniques, PUBLISHED IN IEEE.

Yosuke R. Matsuoka, Gabriel Angelo R. Sandoval, Luis Paolo Q. Say, Jann Skyler Y. Teng, Donata D. Acula, Enhanced Intelligent Character Recognition (ICR) Approach using Diagonal Feature Extraction and Euler Number as Classifier with Modified One-Pixel Width Character Segmentation Algorithm, 2018 International Conference on Platform Technology and Service (PlatCon), 978-1-5386-4710-3/18/$31.00 ©2018 IEEE.

Renuka Kajale, Soubhik Das, Paritosh Medhekar, Supervised machine learning in intelligent character recognition of handwritten and printed nameplate, 978-1-5386-3852-1/17/$31.00 ©2017 IEEE.

E.K. Vellingiriraj, Dr.M. Balamurugan Dr.P. Balasubramanie, Information Extraction and Text Mining of Ancient Vattezhuthu Characters in Historical Documents Using Image Zoning, 2016 International Conference on Asian Language Processing (IALP), 978-1-5090-0922-0/16/$31.00_c 2016 IEEE.