Grenze International Journal of Engineering and Technology, Special issue

Ancient Tamil Character Recognition using Ensemble Classifier

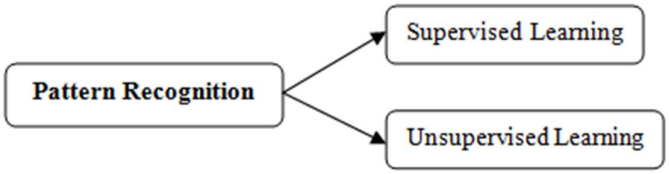
Merline Magrina M[1] and Dr. Santhi M[2]
[1-2]Saranathan College of Engineering, Dept of ECE, Trichy, India
Email: merlinemagrina96@gmail.com, santhiphd@gmail.com

***Abstract*—Ancient Tamil character recognition of the stone manuscripts are rich in the source of art, music, literature, management & administration. It therefore becomes critical to access those stone manuscripts belong to 12th century period by the epigraphers to share the contents present in the manuscripts with the people. In this paper the Ensemble based classification of OCR (Optical Character Recognition) is proposed to give more accuracy of recognizing those characters inscribed on the stone manuscript. The testing sample is preprocessed to remove noises present in the inscriptions using Median filtering. The noise free images are segmented using Bounding Box technology to segment each character. The global and local features are extracted from each character i.e 79 features for each character in the dataset. The one third of the extracted features are applied to the Bagging based KNN & Discriminant and also Boosting technique. The Bagging based technique correctly classifies the characters among 66 classes and the correctly classified character is mapped to Unicode of the Modern Tamil character. The performance measure of Segmentation rate & Recognition rate is calculated.**

***Index Terms*—12th century Tamil characters, OCR, Bounding Box, Character recognition, Bagging technique, Boosting Technique, Unicode mapping.**

I. INTRODUCTION

Pattern Recognition is the process of recognizing patterns by using Machine Learning Algorithms. PR is defined as the classification of data based on the knowledge gained on the statistical feature information extracted from the patterns and on their representation. One of the important aspect of PR is its application potential in Image classification, Character Recognition, Face Estimation, Pose Estimation, etc.



Optical Character Recognition is a field of research in PR, AI, & CV. OCR is the electronic conversion of printed/ handwritten character into a computer editable format. OCR eliminates the staff involvement & store large data for future use. Advanced OCR is of 2 types namely OFFLINE & ONLINE Character Recognition as shown in Fig 1.

Grenze ID: 01.GIJET.5.1.15

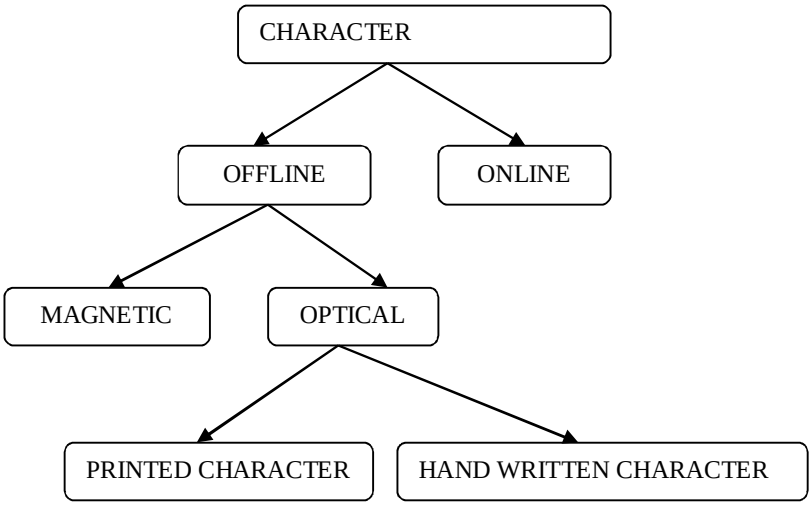


Fig 1: OCR Types

Offline Character Recognition is a conversion of text into codes directly and usually used in many computer applications.

Online Character Recognition is a conversion of text writing i.e. successive movements of the pen on the tablet that are transformed into series of electronic signals that are analyzed by the computer.

Testimonials are important in the formation of Historical Documents. The historical information is written only in stone manuscripts, copper-plate inscriptions are given first preference by the epigraphers to bring out the information to the people. The survey shows that Chola's showed social justice and their period is considered as a Golden Period. From the stone inscription of Ashoka we came to know about their humanity, dharma & religious activities. The inscriptions can be from 1 to 100 lines and each line represents the historical significance of the Chola Period 9th-12th Century. It is not easy to read the stone inscriptions since they are written in Ancient Tamil characters as shown in Fig 2 that can be read only by studying the linguistics & practice. The inscriptions were mostly in Tamil-Bhrami letters (Tamil mixed with Grantha Scripts) as shown in Fig 3. Only if we know the Grantha scripts (Fig 4) it is easy to read the stone manuscript. The ancestor of this stone manuscript is Palm script (Fig 5) because there is pulli characters present in the words & they are written as bare consonants.

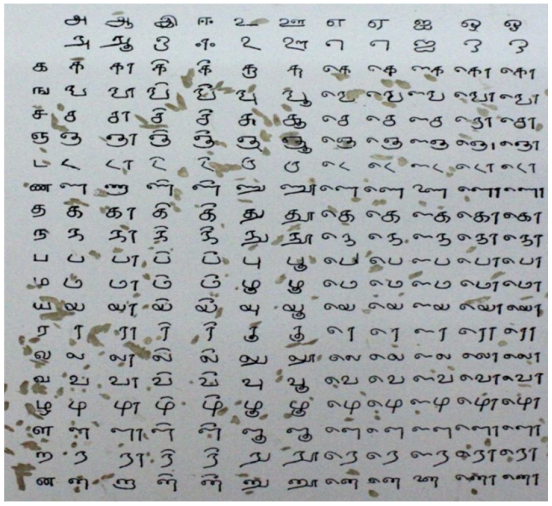


Fig 2: 12th Century Tamil characters found in stone manuscripts

Fig 3: Tamil- Bhrami script

Fig 4: Grantha Script

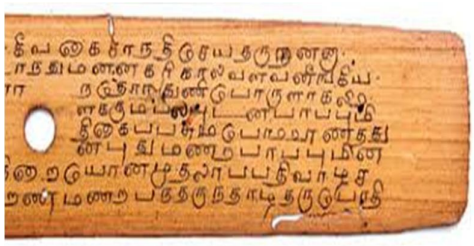


Fig 5: Palm Script

Unique Features of 12th Century Tamil Characters:

1. Bare consonants are present in the stone manuscript.
2. The sentence is written continuously without any separation & full stop at the end of the sentences.
3. The sentences are written like a poem that describes the offerings to the temple, land separation followed, administration & humanity.
4. No difference is followed between the egara, eegaara, yegara, yegaara, ugara, uugara, ogar, oogaara vowel letters.
5. The letter ழ’ & ’ழு’ and also ‘ற’ & ‘ர’ are carved & identified as same letters.
6. If the consonant comes twice continuously the first character is recognized as a pulli consonants.
7. A sample stone manuscript with explanation is given as follows

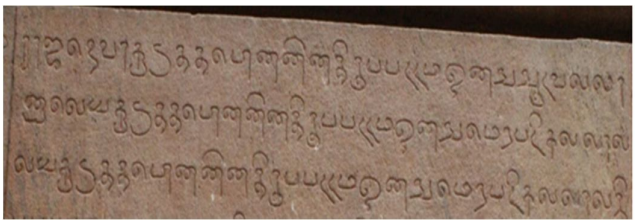


Fig 6: Sample stone manuscript

Explanation of the above stone manuscript

ராஜ தேவர் குடுத்த பொன்னின் திருப்பட்டம் ஒன்று ஆடவல்லா

னுலெய குடுத்த பொன்னின் திருப்பட்டம் ஒன்று மேற்படிக்கல்லால்

லய குடுத்த பொன்னின் திருப்பட்டம் ஒன்று மேற்படிக்கல்லால் நி

Objectives of our work:

1. Recognizing the characters using OCR.
2. The ancient Tamil character is mapped with the modern Tamil character using Unicode.
3. The performance measure of Segmentation rate & Recognition rate is calculated for the testing samples.

II. Literature Survey

OCR’s are available for languages like Arabic, Kannadam, Chinese, Devanagari, etc. Training a Neural Network for implementing an OCR is time consuming. Character matching includes extracting features like height, width, number of horizontal and vertical lines, curves, loops, slope lines and special dots. It leads to identify the centuries of characters using image processing technique and to identify the characters efficiently. OCR converts the printed characters to system editable format using Unicode values. The different inscription images are tested and observe the performance & accuracy of the character level recognition.

The OCR performance for noisy English characters by Inception V3 network is analyzed by Tan Chiang Wei [1]. Generally the character recognition is complicated for noisy images. To achieve the performance of the

testing the constructed network is trained using Transfer learning method. The reduction error rate of 21.5% is achieved.

SIFT (Scale Invariant Feature Transform) is an algorithm that extracts the feature data from an input image and comprises of characteristics that prevent image transformations such as image size and rotation in matching of feature points [5] by Hyun Bin Joo and Jae Wook Jeon.

V.Vaishali et al. [7] developed a system effectively for recognizing the Tamil inscriptions from the stone. The main process deals with removing the noise from the inscription by pre-processing & using many filters for noise removal such as Weiner filter & Median filter. Among these the Median filters are best to remove SALT & PEPPER noise which does not affect the smoothness & sharpness of the characters present in the manuscripts.

The characters are recognized using ANN [8] is proposed by Rajasekar.M and Celine Kavitha. The ANN network consists of input layer, hidden layer & output layer.. The input document is read pre-processed, feature extracted and text is recognized. OCR eliminates the difficulty by making the data to be available in printed format. Then this is applied for different samples in the testing process and the accuracy of the network is calculated.

Robi Polikar [10] proposed an ensemble system consists of 2 key components. (1) Need to build an ensemble that is diverse in nature. Some are bagging, boosting, AdaBoost, stacked generalization. (2) Need to combine the outputs of individual classifiers that the ensemble to make correct decisions. Meta algorithm [12] combines several machine learning (ML) techniques into a predictive model in order to reduce variance (bagging), bias (boosting) & improve prediction (stacking). The main idea is to improve the accuracy of supervised learning task. The challenges are how to construct ensemble & use individual hypotheses of ensemble to produce classification. Increased accuracy is obtained using Bagging Technique for OCR as it uses weighted average voting method.

III. METHODOLOGY

Our proposed work concentrates on the way of digitizing the 12th Ancient Tamil Characters to Modern Tamil Characters that describes the beauty of their dynasty, art, music, literature and justice. The architecture of our work is shown in Fig 7. The testing sample is subjected to pre-processing step of gray scale conversion, enhancement through Image adjustment, binarization through Image Binarization & noise removal through Median Filter. The noise free image is applied for Segmentation called Character segmentation by Bounding Box technique. From the segmented characters the features like Region properties through Region Props, Corner Points (Key points) through Harris Corner Detection and Orientation points & Magnitude points through SIFT algorithm. The features are applied for classification process through Ensemble learning method called Bagging & Boosting methodology. The characters are classified among 67 classes. The ancient Tamil characters are mapped to Modern Tamil Characters using Unicode and the performance measures as Segmentation rate 98.24% & Recognition rate 96.52% is calculated.

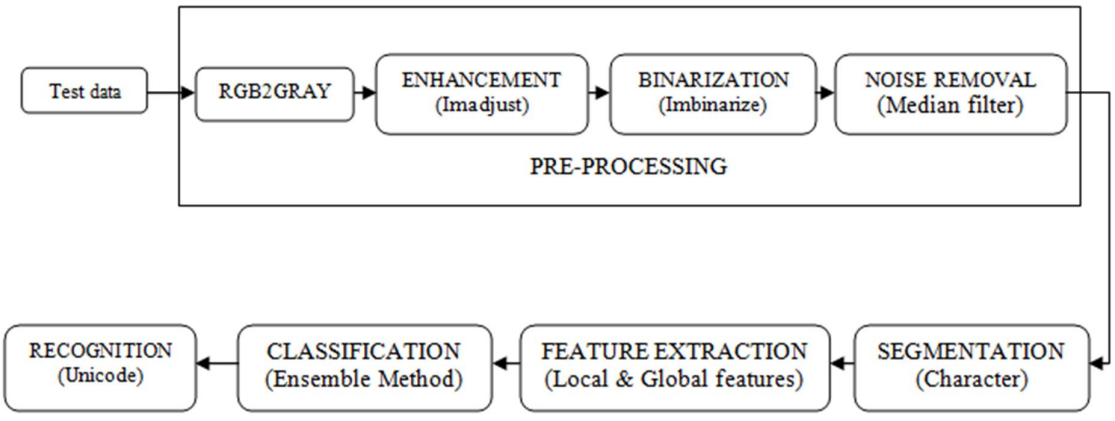


Fig 7: Proposed Block Diagram

Classification

Ensemble Classifier: Ensemble Learning Method [10] helps to improve machine learning (ML) results by combining several models. This model produce better predictive performance compared to a single model. They

are meta-algorithm that combines several machine learning techniques into one predictive model in order to decrease variance (**BAGGING**), Bias (**BOOSTING**) or improve predictions (**STACKING**).

- ✓ **Bagging**: Bagging stands for bootstrap aggregation. The variance of an estimate can be reduced by averaging multiple estimates. It uses voting for classification & averaging for regression. Computed the ensemble as

$$f(x) = {}^{1}/_{M} \sum_{m=1}^{M} f_m(x) \quad (1)$$

Accuracy for decision tree = 0.63 & KNN = 0.70

In random forest, each tree in the ensemble is built from a sample drawn with replacement from the training set. A random subset of features is selected instead of using all the features thereby randomizing the tree. The bias value of the forest increases slightly, but due to averaging of less correlated trees, its variance decreases, resulting in an overall better model.

Algorithm

1. Let the number of training points be N and the number of features in the training data be D.
2. Choose L to be the number of individual models in the ensemble.
3. For each individual model I, choose d_I be the number of input variables I. It is common to have a single value of d_I for all the individual values.
4. For each individual model I create a training set by choosing d_I features from D with replacement and train the model.

Recognition

The **Unicode standard** reflects the basic principle that emphasizes each character code has a numerical value. Unicode text is simple to define & process. After classification the characters, they are recognized by Unicode. The exact character is mapped & the recognition rate is calculated.

Ancient Tamil Character	Modern Tamil Character
	கு
	க
	ஓ
	அ
	னா
	ரு
	ட

Fig8: Sample Character Mapping

Performance Measures

1) *Segmentation rate:* The mathematical expression of the segmentation rate is given as follows

$$\boldsymbol{SR = \frac{no\ of\ characters\ correctly\ segemented}{total\ no\ of\ characters} * 100}$$

2) *Recognition rate:* It is the measure of number of samples that are correctly classified is as follws

$$\boldsymbol{RR = \frac{no\ of\ characters\ correctly\ classified}{total\ no\ of\ characters} * 100}$$

3) *Accuracy:* The mathematical expression for the classifier performance is as follows

$\boldsymbol{Accuracy = \frac{TP+FP+TN+FN}{TP+FP}}$ where TP= number of characters which are correctly assigned to the given category, TN= number of characters which are correctly assigned not to belong to category, FP= number of characters which are incorrectly assigned to the category, FN= number of characters which are incorrectly not assigned to the category.

4) *Specificity*: $\boldsymbol{Specificity = {}^{TN}/_{TN+FP}}$
5) *Sensitivity*: $\boldsymbol{Sensitivity = {}^{TP}/_{TP+FN}}$

6) *F1_measure*: The performance of the classifier is given as

$$\boldsymbol{F1\_measure} = \boldsymbol{2 * (Precision * Recall)} \Big/ \boldsymbol{(Precision + Recall)}$$

IV. Experimental Results

The original inscription image is collected from the dataset. It can be any document of different dimensions. This captured image (Fig 9) of dimension 401*147 in jpeg format is fed to pre-processing section.

Fig 9: Input image

To eliminate further interference and make the edge smooth Dilation operation is done as in Fig10.

Fig 10. Post procesed image

It is a process of distinguishing characters of an inscription image and extracts meaningful regions for analysis. Rectangular bounding box is used to bind each character into smallest possible rectangle as in Fig11.

Fig11. Segmented Character images

The exact characters are recognized to map the characters with corresponding Unicode values [56, 27, 2]. The recognized modern Tamil characters are shown in Fig 12.

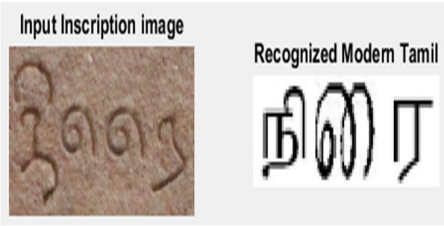


Fig10. Recognized Modern Tamil Character

The recognition rate is calculated as follows

$$\boldsymbol{RR} = \frac{\boldsymbol{no\ of\ correctly\ classified\ charcters}}{\boldsymbol{total\ no\ of\ characters}} * \boldsymbol{100} = \frac{\boldsymbol{513}}{\boldsymbol{532}} * \boldsymbol{100} = \boldsymbol{96.52\%}$$

The segmentation rate is calculated as follows

$$SR = \frac{no\ of\ correctly\ segmented\ charcters}{total\ no\ of\ characters} * 100 = \frac{520}{532} * 100 = 97.79\%$$

Fig 11 & 12 shows the percentage of accuracies for ensemble learning bagging classifier

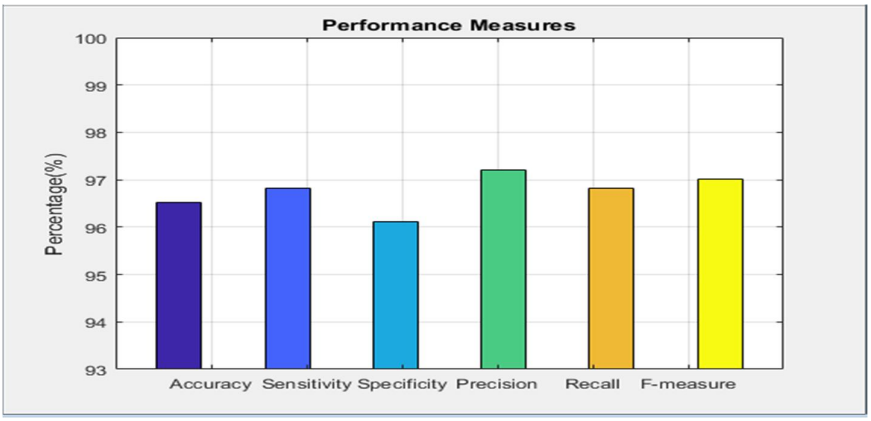
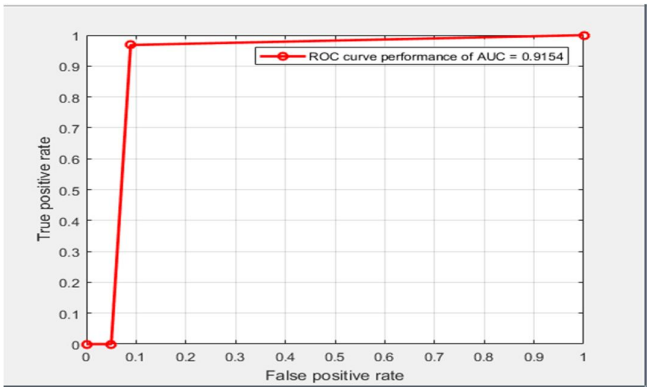


Fig 11: Performance Metrics

TABLE I: PERFORMANCE MEASURES

Metrics	Range
TP	243
TN	173
FP	7
FN	8
Accuracy	96.51%
Sensitivity	96.81%
Specificity	96.11%
Precision	97.20%
Recall	96.81%
F-measure	97%



ACCURACY = 91.5%

Fig12: ROC Performance of the classifier System

TABLE II: PERFORMANCE OF ENSEMBLE METHODOLOGY

Methodology & Type		Segmentation rate	Recognition rate
Bagging	**KNN**	***98.24%***	***96.52%***
	Discriminant	***98.14%***	***96.52%***
Boosting	AdaboostM2	8.7%	10.2%
	RUSBoost	7.9%	7.55%
	LPBoost	6.9%	7.27%
	TotalBoost	11.66%	11%

V. CONCLUSION

Thus the proposed work focuses on the recognition of ancient Tamil characters from epigraphical inscriptions to modern Tamil characters. The traditional methods of inscription identification are time consuming & laborious to digitize. Hence an innovative technique is developed to identify the exact modern Tamil characters from epigraphical Tamil characters. For this the testing samples are subjected to pre-processing by median filters, segmentation through bounding box, features are extracted by SIFT algorithm. The extracted features are applied for classification using Ensemble learning classifier. The modern Tamil character is mapped using Unicode. The performance metric calculated for segmentation & recognition rates of 98.24% and 96.52% are obtained respectively using KNN based random forest method & it is found to be the best among the ensemble learning boosting methods.

ACKNOWLEDGEMENT

I owe my sincere thanks to **Dr. M. BHAVANI, Assistant Professor**, Department of Epigraphy & Archaeology, Tamil University, Thanjavur and **Mr. M. RAJENDRAN, Record clerk**, Department of Archaeology, Thanjavur who provided me the dataset and their suggestion & guidance throughout my project.

REFERENCES

[1] Tan Chaing Wei, Sheikh U.U and Ab-Al Hadi Rahman, "Improved OCR with Deep Neural Network", IEEE 14th International Colloquium on Signal Processing and its application (ICSPA 2018), Mar 2018, pp: 245-259.

[2] Beihai Tan, Chao Hu and Zepei Zhang, "Character Recognition based on Corner Detection", IEEE International Conference on Natural computation, Fuzzy System and Knowledge Discovery (ICNC-FSKD), Mar 2017, pg: 503-507.

[3] Bhavani.B (2017), "Tamilnadu Historical Documents (Epigraphical Inscriptions)".

[4] DurgaDevi. K and Uma Maheswari. P, "Insight on Character Recognition for calligraphy digitization", IEEE International Innovations in ICT for Agriculture and Rural Development (TIAR), Jun 2017,pp: 78-89.

[5] Ellakiya. V, Jegajothi. M and Muthumani. I, "Tamil Text Recognition using KNN classifier", Advances in Natural and Applied Sciences, Open AccessJournal, Jan 2017, vol:7, issue:11, pp:41-45.

[6] Hyun Bin Joo and Jae Wook Joen , "Feature point extraction based on improved SIFT algorithm", International Conference on Control Automation and System (ICCAS), Oct 2017, pp: 345-350.

[7] Karunarathne K.G.N.D, Liyanage K.V, "Recognizing Ancient Sinhala Inscription Characters using Neural Network Technologies", International Journal of Scientific Engineering & Applied Sciences (IJSEAS), vol:3, issue:1, Jan 2017, pp: 37-49.

[8] Janani. G, Vaishali. V and Mohan, "Recognition and Analysis of Tamil Inscriptions & Mapping using Image Processing Technique", IEEE International Conference on Science and Technology Engineering & Management, Apr 2016, pp: 181-184.

[9] Rajasekar. M, Celine Kavitha and Anto Bennet.M , "Performance and analysis of Handwritten Tamil Character Recognition using Artificial Neural Network", International Journal of Recent Scientific Research (IJRST), vol: 7, Jan 2016, pp: 8611-8615.

[10] Preethi. N and Gunasekaran. T , "Language Specific Offline character recognition using Neural Network Classifier", International Journal of Advanced Research in Electronics & Communication Engineering (IJARECE), Feb 2015, vol: 4, pp: 218-221.

[11] Anitha Mary, Chacko and Dhanya, "Mutiple Classifier System for Offline Malayalam Character Recognition", International Conference on Information & Communication Technologies, vol: 46, Jan 2014, pp: 86-92.

[12] Robi Polikar, "Ensemble based systems in decision making", IEEE Circuits & Systems Magazine, vol:6, Sep 2010, pp:21-45.

[13] Warren Cheung and Ghassan Hamarneh, "n-SIFT: n dimensional scale invariant feature extraction", IEEE Transactions on Image Processing, vol: 18, Sep 2009, pp: 720-723.

[14] https://Tamil_UnicodeCode_block
[15] https://UnicodeCharacterdatabase
[16] https://TamilcharacterEncodingSchemes
[17] https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/