

**BUILDING A ROBUST DATASET FOR INSCRIPTION CHARACTER
RECOGNITION USING DEEP LEARNING****Dr.R.Deepalakshmi¹****Professor
Department of computer
science,
Velammal college of engineering and
technology, Madurai.
rdl@vcet.ac.in****M.B.Varun Adhityan², N.K.Srekaravarshan³,
V.K.Jeganath⁴, C.Pradeesha⁵****varunadhityanbaskar@gmail.com
, srekaravarshan@gmail.com,
jeganath2003@gmail.com,
pradeeshac994@gmail.com
Velammal college of engineering and
technology,
Madurai.**

Abstract-- Even though inscription characters are a significant part of human history and culture, machine learning algorithms still struggle to recognize them. For the purpose of creating precise and dependable machine learning models for inscription character recognition, a high-quality dataset must be created. In this abstract, we provide an overview of the steps required in building a dataset for inscription characters, as well as the difficulties encountered during data collection and preprocessing. We explain the many methods for gathering inscription data, including digitization and photography, and offer suggestions for efficient data gathering. We also go over the many preprocessing methods that can be used to improve the dataset's quality, such as picture normalization, segmentation, and noise reduction. The creation of precise inscription character recognition models will be made possible by adhering to the best practices described in this abstract, and this will facilitate research in a variety of domains, such as linguistics, history, and archaeology.

Keywords—Dataset, Inscription, Machine Learning, Data Preprocessing

I. INTRODUCTION

The fascinating aspect of ancient Tamil culture has drawn the attention of scholars, historians, and enthusiasts all around the world. Tamil Nadu, a state in the south of India, has a lengthy history that goes back more than two thousand years. Tamil, one of the oldest and most sophisticated languages in the world, is spoken there and has a distinctive and complex writing system.

Tamil literature, art, and architecture from antiquity display the splendour and wealth of Tamil civilization. Tamil literature has a rich and diverse history which encompasses both secular and religious works. The ancient Tamil poems and songs that make up the Sangam literature, which is thought to have been written between the third century BCE and the third century CE, provides insights into the social, cultural, and political lives of the Tamil people.

The ancient temples of Tamil Nadu are among the best examples of Dravidian architecture, and Tamil art and architecture are also widely respected. The intricate carvings, towering gopurams (towers), and vast courtyards of these temples are well-known features. The bronze idols from the Chola era are among the most notable examples of the exquisite and detailed bronzes found in Tamil Nadu.

Tamil inscriptions are prehistoric texts that were inscribed using the Tamil script on rocks, metal plates, and other objects. They contain important details regarding the Tamil people's history, culture, language, and way of life. The Tamil people have lived in southern India for many centuries. One of the earliest and most important sources of knowledge about Tamil civilisation is thought to be Tamil inscriptions.

Tamil inscriptions have a long history that dates back to the third century BCE, when the language and script were originally created. The earliest inscriptions were discovered in the Tamil Nadu region, which was a centre of Tamil culture and learning. The inscriptions were primarily used to record royal edicts, land grants, and religious donations. They were frequently written in verse and were sometimes accompanied by illustrations and images. Tamil writing evolved over time to become more complex and sophisticated. They were used to document the

actions of rulers, the accomplishments of poets and academics, and the regular lives of people in society. The inscriptions were also employed to disseminate knowledge and learning, as well as to promote religious and cultural values.

Tamil inscriptions reached their pinnacle of complexity and sophistication during the mediaeval era. Large stone pillars that were placed in public areas like temples and markets were inscribed with them. These inscriptions were extremely elaborate, featuring dexterous calligraphy and exquisite designs, and they frequently came with in-depth depictions and images.

Tamil inscriptions are still a priceless source of knowledge about Tamil civilization. They offer insights into the Tamil people's social, cultural, and political lives, as well as their language, literature, and religious practises. Additionally, Tamil inscriptions have helped scholars in their understanding of the language's and script's historical development as well as the development of Tamil culture.

The wonder of datasets rests in the abundance of data and knowledge they contain. Individual data points or tiny data samples may not always show patterns, trends, or insights, but datasets can. Researchers and analysts can find hidden linkages, correlations, and causalities in massive datasets that can be used to guide decision-making, develop new products and services, or enhance current ones.

Datasets can also be used to train machine learning models, which are able to identify intricate data patterns and make predictions based on them. Large datasets can be used to train machine learning models to identify patterns and generate predictions that are more reliable and precise than those generated by people.

Moreover, datasets can assist researchers and analysts in testing theories, confirming or refuting hypotheses, and gaining fresh insights. Additionally, they can be used to compare performance and track advancement over time. Researchers and analysts can assess the efficacy of interventions, regulations, or other changes by comparing new data to previous databases.

The power of databases, in general, resides in their capacity to highlight information and insights that might otherwise go unnoticed. Researchers, analysts, and developers could improve decisions, create new technologies, and better the environment around us by using datasets to analyse data.

We are proposing to develop a system with a robust dataset and character recognition system to understand these inscriptions and learn more about our ancient history, civilization, and deeds carried out by our ancestors because technology provides solutions for any challenge we encounter.

II. LITERATURE REVIEW

The demand for high-quality Tamil language datasets is growing as more and more natural language processing (NLP) applications are created. In this review of the literature, we will look at the methodology employed, the many kinds of datasets that are available, and the applications that have been created utilising this dataset generation research in Tamil. Tamil native speakers manually annotating datasets is the way that is most frequently employed. In order to do this, a sizable corpus of text must be created and annotated with a variety of linguistic features, including part-of-speech tags, named entities, and sentiment labels. These annotated datasets can subsequently be used to train machine learning algorithms for a variety of NLP tasks. Another approach is to automatically create new datasets using already-existing resources like parallel corpora or machine translation tools.

[1] Giridharan.R,Vellingiriraj.E.K, and Dr.Balasubramanie have presented a method for character recognition and information retrieval of Brahmi,Vattezhuthu, and Grantha letters from temple epigraphy, as well as the conversion of these characters into digital form. They asserted that their method would address issues like maintaining language grammar and correcting the spellings.

[2]RajaKumar.s and Dr.V.subbaih Bharathi.s system tries to recognise and identify the characters used in Vatteluttu inscriptions from the 7th century, which are difficult to read due to age and damage. For historians and archaeologists who research the history and culture of the ancient Tamil people, this method can be helpful. This method is only focused on Tamil characters from the 7th century.

[3]Manigandan.T et al. have proposed a Tamil character recognition system based on OCR and NLP that focuses on Tamil characters from the 9th to 12th centuries. They pre-process and segment the inscription photos obtained from the Tamil Nadu Archaeological Department in their work. The colour photos were transformed to grayscale and binary images depending on threshold values during the segmentation process. For each letter, Scale Invariant Feature Transform (SIFT) methods have been used to extract image features such as the quantity of lines, curves, loops, and dots in order to precisely identify the character. Support Vector Machine (SVM) classifier will be used to classify and produce characters, and the patterns of the characters will be compared to those of well-known characters and predicted using the Trigram method. For further character identification and to improve the system's

ability to recognise the characters, each identified character will be given its matching Unicode value and updated in the image corpus. Thus, the major issues with reading the inscription images can be resolved by the proposed system. [4]Lalitha Giridhar, Aishwarya Dharani, and Velmathi Guruviah have presented a fresh approach to OCR that focuses on enhancing optical character recognition algorithms for the historic Tamil script, which was in use during the 7th and 12th centuries. While it is a difficult undertaking to thoroughly curate a useful data set for ancient Tamil characters, in this work a data set has been curated using cropped images of characters found on specific temple inscriptions, unique to this time period. A two-dimensional convolution neural network is created and used to train, classify, and recognise the historic Tamil characters after the image has been binarized using the Otsu thresholding method. The Pytesseract package in Python is used to connect the neural network to the Tesseract in order to create optical character recognition techniques. This effort also uses Google's text-to-speech voice engine as an extra feature to provide an audio output of the digitised text. Numerous samples of both contemporary and historic Tamil were gathered and processed. It is discovered that a combined efficiency (OCR and text to speech) of 77.7% may be reached for Tamil inscriptions analysed over the considered time span.

[5]Devi Priya R et al. introduced a self-adaptive Lion Optimisation Algorithm (LOA) that is used to optimise brightness and contrast in stone inscription images that have been pre-processed for noise removal, and then each character is separated by identifying contours. Transfer Learning (TL), a Deep Convolution Neural Network-based multi classification method, is used to recognise characters. When applied to images of stone inscriptions, the suggested hybrid model Self-Adaptive Lion Optimisation Algorithm with Transfer Learning (SLOA-TL) outperforms conventional methods in terms of accuracy and speed. The recognition of Tamil characters in stone inscriptions and the preservation of Tamil traditional knowledge are both accomplished by this method.

[6]Shalaka Deore Prasad and Albert Pravin has proposed a paper focusing on the analysis and fine-tuning of a cutting-edge Deep Convolutional Neural Network (DCNN) created for the classification of Devanagari Handwritten characters. They produced a brand-new Devanagari handwriting dataset that includes 5800 isolated images of 58 distinct character classes, including 12 vowels, 36 consonants, and 10 numbers. A two-stage VGG16 deep learning model was also used in conjunction with this database to recognise those characters using two cutting-edge adaptive gradient methods. To improve the overall effectiveness of the proposed Devanagari Handwritten Character Recognition System (DHCRS), a two-stage deep learning methodology was created. The initial model had a training loss of 0.18 and a testing accuracy of 94.84%. Additionally, the second fine-tuned model achieves cutting-edge performance on a very small dataset with a lot fewer trainable parameters and far less training time. Testing accuracy was 96.55% and training loss was 0.12 percent.

[7]Kavitha Subramani and Murugavalli Subramaniam have conducted a study that aims to separate out ancient Tamil palm leaf manuscripts related to the subject of medicine in order to develop a sizable amount of Tamil character datasets. In their investigation, the fictional characters are fed into expert systems as inputs in order to recognise context and content that are thought to be present in the chosen medical articles. Large numbers of the characters have been manually detected, and datasets are made using Gaussian distortion.

[8]Yong Haur Tay et al. described two offline handwriting recognition systems, the first employing standard discrete HMMs and the second a NN-HMM hybrid. Results on the IRONOFF-196, IRONOFF Cheque, and SRTP-Cheque databases are presented, showing the superiority of the hybrid recognizer. Finally, they showed how the hybrid recognizer can be automatically bootstrapped from the discrete HMM recognizer and how many training phases can considerably increase its recognition accuracy.

III. METHODOLOGY

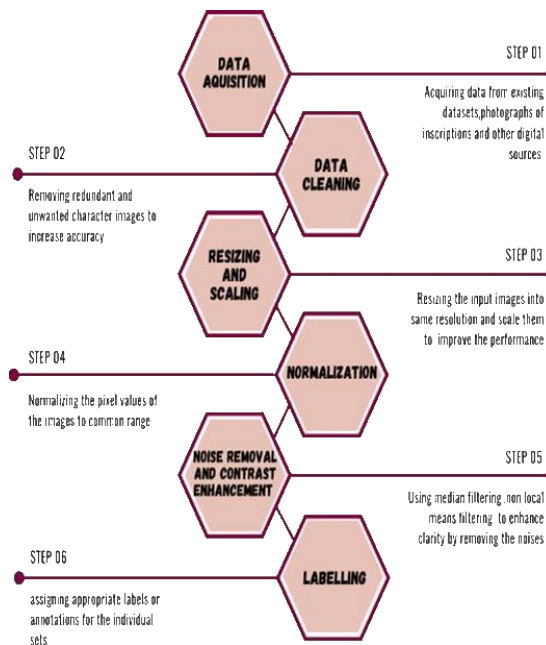


Fig. 1 Methodology for Creation of Image Dataset

IV. PURPOSE AND SCOPE OF THE DATASET

Our dataset's main objective is to provide as a source of data for Tamil character recognition systems. Given the vastly increased quantity of training set images, this dataset will be more effective than the current datasets.

V. PLANNING AND DESIGNING THE DATA COLLECTION METHODS:

A few sources and methods for the data gathering for the dataset include existing images of the inscriptions, accessible datasets, the collection of data images directly from the inscriptions by images, and organising a team to write multiple versions/copies of the characters.

VI. COLLECTION AND PREPROCESSING OF DATA:

Errors, missing numbers, outliers, and inconsistent data are frequently present in raw data. Data cleaning entails locating and addressing these problems. For greater accuracy, noise reduction and thresholding are applied to the data images.

A. Resizing and Scaling:

One of the major steps involved in pre-processing is image scaling. There are numerous benefits in using this strategy. A well-formed input data format must be taken into account in order to construct a successful data model and to make sure that all of the data inputs are the same size and ratio, resize. We must scale the images after making sure that all of the data inputs have been scaled. There are several suggested scaling techniques for the supplied image. In our work, the Lanczo Algorithm is built and used after extensive experimentation with numerous techniques.[9] The Lanczo interpolation algorithm preserves the edges in the scaled images since handwritten scripts have various dimensions and pixel counts. This is crucial for handwritten characters since they frequently feature delicate nuances and complicated strokes.



Fig. 2 Post result of Resizing and Scaling

B. Normalization:

Normalization is important when working with Inscription images that have varying pixel intensities, which can make it difficult to identify patterns, edges and feature of the characters. Normalization ensures that the pixel values of an image are within a specific range, which helps in representing the features of the image better. As the dataset will be fetched with enormous number of test images, captured under different lighting conditions can have varying pixel values, It will increase the effect of illumination. To reduce the effect of illumination and to visualize the outliers of the dataset, we used Z-Score algorithm is used.[10] Z-Score Algorithm allows data points to be standardized and compared against each other, regardless of their original scale. The main objective of this is to reduce the dimension of the original data image by normalizing the intensities.



Fig. 3 Post result of Normalization

C. Denoising:

Denoising an inscription image is a challenging task that requires special techniques to preserve the text while removing the noise and artifacts. Inscription images are often affected by noise due to various factors such as image capture conditions, scanner settings, and image processing algorithms. The presence of noise and artifacts can reduce the readability of the inscription and make it difficult to interpret. The efficient technique that is used in our dataset to denoise an inscription image is Fast Non-Local Means filtering. [10] It is a powerful technique that involves averaging the pixel values of the image within a window based on the similarity of their neighbourhoods. This technique is effective in removing noise while preserving the details and edges of the inscription. Mainly, it denoises the cracks and noised areas.



Fig. 4 Denoised image of normalized image

LABELING AND ANNOTATING THE DATA:

Labelling and annotating some forms of data, such as images or text, may be necessary. To enable supervised learning or other types of analysis, this entails giving the data relevant tags or labels. The dataset visuals are mapped to the actual words using Lableme. Here, we map the image folders for historical figures to the contemporary Tamil characters that correspond to them.

VII. SPLITTING AND DOCUMENTATION OF THE DATASET

A training set, a validation set, and a testing set were created from the dataset. Models are trained using training data, validated using validation data, and tested against testing data to determine the model's ultimate performance. After the dataset has been validated, we will produce a thorough documentation. Details including data sources, collection methods, variables, pre-processing stages, and other pertinent information are included in this documentation. This documentation is essential for comprehending, replicating, and sharing your study and dataset with others.

After dataset completion, we will post it in public spaces like GitHub, Kaggle, or the UCI Machine Learning Repository.

VIII. RESULT & DISCUSSION

This paper presents an efficient vattezhuthu dataset which will be useful for Tamil character recognition system that we are planning to develop in the near future. This dataset will be an improved version of the existing dataset and will provide more accuracy in the charter recognition system. the proposed dataset will contain 10000 images in the training set and 7000 images in the test set for providing an enhanced performance. M.A.Pragathi et al., have presented a solution for Tamil OCR that includes a database, algorithm and an application. Their dataset consists of 15,600 character images each of dimensions 224x224. Our proposed system for Tamil handwritten characters gives efficiency of 94.52%. Our dataset will be more efficient than the previously mentioned solution and will have the efficiency of 95.21%. After the deployment of our tamil character recognition system if the input data is not available in our dataset, then the image will be dynamically added into our dataset. Hence the accuracy of the dataset will also increase.

IX. CONCLUSION

In conclusion, the development of an Ancient Tamil Characters dataset is an essential step in protecting and upholding Tamil language. This dataset can be used for a variety of purposes, including machine learning-based recognition systems, digital archives, and historical research. The paper explains the pre-processing methods needed to improve the quality and legibility of the characters that are to be used for building the dataset. The paper also highlights the value of preserving these historic traits, which are a crucial component of Tamil's cultural identity. Researchers, historians, and linguists now have access to a comprehensive resource for researching and preserving the Tamil inscriptions. The creation of this dataset is a great illustration of the potential benefits that come from digitising and preserving the stone inscriptions. With the use of this dataset, we can learn more about the history of the Tamil language and its evolution, which will help us embrace Tamil Nadu's rich cultural heritage on a deeper level.

REFERENCES

- [1] Giridharan.R,Vellingiriraj.E.K,Dr. Balasubramanie.P, " *Identification of Tamil Ancient Characters and Information Retrieval from Temple Epigraphy Using Image Zoning*", International Conference on Recent Trends in Information Technology (2016), pp 1-5
- [2] RajaKumar.S, Dr.Subbiah Bharathi.V, "Eighth Century Tamil Consonants Recognition From Stone Inscriptions", International Conference on Recent Trends in Information Technology (2012),pp 40-43
- [3]Manigandan.T,Dr.Vidhya.V,Dr.Dhanalakshmi.V,Nirmal a.B,"Tamil Character Recognition from Ancient Epigraphical Inscription using OCR and NLP",International Conference on Energy,Communication, Data Analytics and Soft Computing (2017),pp 1008-1010
- [4] Lalitha Giridhar,Aishwarya Dharani,Velmathi Guruviah,"A Novel Approach to OCR using Image Recognition based Classification for Ancient Tamil Inscriptions in Temples",arXiv preprint arXiv:1907.04917(2019),pp 1-5
- [5] Devi Priya.R, Karthikeyan.S, Indra.J, Kirubashankar.S, Ajith Abraham, Lubna A. Gabralla,Sivaraj.R,

- Nandhagopal.SM,"Self-Adaptive Hybridized Lion Optimization Algorithm with Transfer Learning for Ancient Tamil Character Recognition in Stone Inscriptions",IEEE Access(2023), pp 1-13
- [6] Shalaka Prasad Deore,Albert Pravin,"Devanagari Handwritten Character Recognition using fine-tuned Deep Convolutional Neural Network on trivial dataset",Springer Nature(2020),pp 1-12
- [7]Kavitha Subramani,Murugavalli Subramaniam,"Creation of original Tamil character dataset through segregation of ancient palm leaf manuscripts in medicine",Expert Systems Volume 38,Issue 1(2020),pp 1-12
- [8] Yong Haur Tay,Pierre-Michel Lallican,Marzuki Khalid,Christian Viard-Gaudin,Stefan Knerr,"An offline cursive handwritten word recognition system",Proceedings of IEEE Region 10 International Conference on Electrical and Electronic Technology. TENCON (2001),pp 3-10
- [9] Pankaj S. Parsania,Dr. Paresh V. Virparia,"A Comparative Analysis of Image Interpolation Algorithms",International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 1(2016),pp 29-34
- [10] Chris Cheadle, Marquis P. Vawter,William J. Freed, Kevin G. Becker,"Analysis of Microarray Data Using Z Score Transformation",Journal of Molecular Diagnostics, Vol. 5, No. 2(2003),pp 73-81
- [11] Yan-Li Liu, Jin Wang,Xi Chen,Yan-Wen Guo,Qun- Sheng Peng,"A Robust and Fast Non-Local Means Algorithm for Image Denoising",JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 23(2) ,pp 270–279 (2008)
- [12] S.Gopal Krishna Patro,Kishore Kumar sahu,"Normalization: A Preprocessing Stage",arXiv:1503.06462 (2015),pp 1-3
- [13]Suganya Athisayamani, Dr.Robert Singh.A, Dr.Athithan.T,"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),PP 1-8
- [14]Vijaya Lakshmi.TR , Panyam Narahari Sastry , Rajinikanth .TV,"A novel 3D approach to recognize Telugu palm leaf text",(ELSEVIER)Engineering science and technology,an international conference(2017),pp 143-150
- [15]M.A.Pragathi, K. Priyadarshini ,S. Saveetha ,A. Shavar Banu ,K. O. Mohammed Aarif ,"Handwritten Tamil Character Recognition Using Deep Learning ",International Conference on Vision Towards Emerging Trends in Communication and Networking (2019),pp 1-5
- [16]Shankar Mahadevan, Rahul Ponnusamy, Prasanna Kumar Kumaresan,Prabakaran Chandran, Ruba Priyadharshini, Sangeetha Sivanesan,Bharathi Raja Chakravarthi,"Thirumurai: A Large Dataset of Tamil Shaivite Poems and Classification of Tamil Pann",Proceedings of the 13th Conference on Language Resources and Evaluation (2022), pp 6556–6562
- [17]Rapeeporn Chamchong and Chun Che Fung,"A Framework for the Selection of Binarization Techniques on Palm Leaf Manuscripts Using Support Vector Machine",Hindawi Publishing Corporation Advances in Decision Sciences Volume 2015, Article ID 925935, pp 1-6, <http://dx.doi.org/10.1155/2015/925935>
- [18] O. Akbani, A.Gokrani, M. Quresh, Furqan M. Khan, Sadaf I. Behlim, Tahir Q. Syed ,"Character Recognition in Natural Scene Images",International Conference on Information and Communication Technologies(2015) pp1-4
- [19]AMIR YAVARIABDI , HUSEYIN KUSETOGULLARI , TURGAY CELIK,SHIVANI THUMMANAPALLY, SAKIB RIJWAN, AND JOHAN HALL ,"CArDIS: A Swedish Historical Handwritten Character and Word Dataset", IEEE Access ,Volume: 10(2022),pp 55338-55347
- [20]Khoi NGUYEN-TAN, Romain RAFFIN, Marc DANIEL and Cung LE ,"B-spline surface reconstruction by inverse subdivisions ", IEEE-RIVF International Conference on Computing and Communication Technologies(2009),pp 1-4
- [21]Mrs.G.Bhuvaneshwari ,V. Subbiah Bharathi,"An Efficient Positional Algorithm for Recognition of Ancient Stone Inscription Characters ",Seventh International Conference on Advanced Computing(2015),pp 1-5
- [22]R.A.E. Coningham, F.R. Alicifrt, COM. Batt and D. Lucy ,"Passage to India? Anuradhapura and the Early Use of the Brahmi Script",Cambridge Archaeological Journal 6:1 (1996), pp. 73—97
- [23]Guang-Bin Huang, Qin-Yu Zhu, Chee-Kheong Siew," Extreme learning machine: Theory and applications",(ELSEVIER)Neurocomputing Volume 70, Issues 1–3(2006), Pp 489-501
- [24]Md Mahbubar Rahman,M. A. H. Akhand,M. M. Hafizur Rahman ,"Bangla Handwritten Character Recognition

- using Convolutional Neural Network",International Journal of Image, Graphics and Signal Processing (2015)
DOI: 10.5815/ijigsp.2015.08.05,pp 52-59
- [25]G. Janani, V. Vishalini, Dr .P. Mohan Kumar ," Recognition and Analysis of Tamil Inscriptions And Mapping Using Image Processing Techniques" ,2016 Second International Conference on Science Technology Engineering and Management,pp 181-184
- [25] Abdelrahman Abdallah, Mohamed Hamada and Daniyar Nurseitov ,"Attention-Based Fully Gated CNN- BGRU for Russian Handwritten Text",arXiv:2008.05373 (2020),pp 1-21
- [26]Manoj Kumar Mahto ,Manoj Kumar Mahto ,R. K. Sharma ,"Combined Horizontal and Vertical Projection Feature Extraction Technique for Gurmukhi Handwritten Character Recognition ",2015 International Conference on Advances in Computer Engineering and Applications ,pp 59-68
- [27]Dr.G.Bhuvaneswari, Dr.G.Manikandan
,"RECOGNITION OF ANCIENT STONE INSCRIPTION CHARACTERS USING HISTOGRAM OF ORIENTED GRADIENTS ",International Conference on Recent Trends in Computing, Communication and Networking Technologies(2019),pp 1-8
- [28]Somaya Al-ma'adee,Jihad Mohamad Aljaa,Abdelâali Hassaïne,"QUWI: *An Arabic and English Handwriting Dataset for Offline Writer Identification*",International Conference on Frontiers in Handwriting Recognition(2012),pp 1-8
- [29]LIANG XU , YUXI WANG, XIUXI LI, AND MING PAN ,"Recognition of Handwritten Chinese Characters Based on Concept Learning",IEEE Access ,Volume: 7(2019),pp 102039 - 102053
- [30] Verónica Romeroa , Alicia Fornés, Nicolás Serrano, Joan Andreu Sánchez, Alejandro H. Toselli, Volkmar Frinken, Enrique Vidal, Josep Llado's,"*The ESPOSALLES Database: An Ancient Marriage License Corpus for Offline Handwriting Recognition*",(ELSEVIER)Pattern Recognition Volume 46, Issue 6(2013), Pages 1658-1669
- [32]N.Sridevi ,"Combining Zernike Moments with Regional features for Classification of Handwritten Ancient Tamil Scripts using Extreme Learning Machine" , IEEE International Conference ON Emerging Trends in Computing, Communication and Nanotechnology (2013),PP 158-162,DOI:10.1109/ICE-CCN.2013.6528483
- [33]S. Thadchanamoorthy ,N. D. Kodikara; H. L. Premaretne ,Umapada Pal ,Fumitaka Kimura ,"Tamil Handwritten City Name Database Development and Recognition for Postal Automation ",12th International Conference on Document Analysis and Recognition(2013),pp 793_798
- [34]E.K.Vellingiriraj,Dr.M.Balamurugan,Dr.P.Balasubram anie,"Information Extraction and Text Mining of Ancient Vattezhuthu Characters in Historical Documents Using Image Zoning " ,2016 International Conference on Asian Language Processing (2016),pp 37-40.