Vol. 5 No. 2 October 2017 ISSN: 2321-788X UGC Approval No: 43960 Impact Factor: 2.114

TAMIL ANCIENT SCRIPTS COLLECTION AND PREPROCESSING USING BIG DATA

Article Particulars

Received: 18.8.2017 Accepted: 15.09.2017 Published: 30.10.2017

Mrs. L. SURIYA KALA, MCA., M.Phil.,
Research Scholar, Mother Teresa Women's University
Kodaikanal, Tamil Nadu, India

Abstract

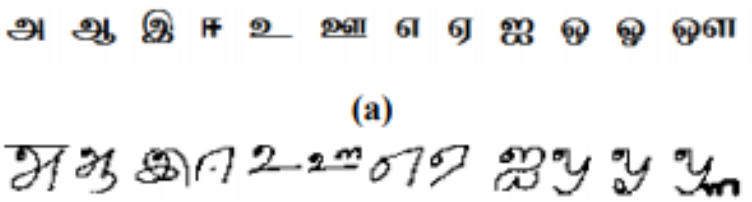
Character recognition of Brahmi, Grantha and Vattezuthu Characters from palm original copies of Historical Tamil Ancient Documents, analyzed the content and machine interpreted the present Tamil computerized content organization. In spite of the fact that numerous specialists have executed different algorithms and strategies for character recognition in various dialects, Ancient characters change still represents a big challenge. An offline transcribed Tamil Character Recognition system is typically created considering point-based highlights that depict distinctive geometric qualities of handwriting. Often, because of the wide varieties recorded as hard copy styles, the utilization of point-based highlights bring about high intra-class changeability in include space. In this paper, we have caused an endeavor to perceive ancient Tamil characters by utilizing To filter includes and introduced another and proficient methodology dependent on a pack of the key focuses portrayal.

Keywords: Ancient, Preprocessing, Big Data, Sift Algorithm, Character Recognition.

1. Introduction

The Tamil lyrics were separated from an ancient southern state of the Brahmi lyrics. These days, it is utilized to record the Tamil language in Tamil Nadu, a state in India. This strategy is followed in a nation like Sri Lanka as well. In southern India, Tamil is one of the most seasoned dialects contrasted with different dialects. Prior to the 1stcentury, the underlying language written in the southern option was Brahmi. Afterward, the lyrics were misshaped, and the Tamil lyrics were extricated into the present structure in the eighth century. In prior hundreds of years, the Traditional Medicinal System (TMS) was utilized by individuals to counteract the ailment event for robust and healthy living. This TMS system was extremely famous among individuals lives in Chinese, African, and Indian. In particular, India pursued just one of its sort strategies, to be specific Indian System of Medicines (ISM). The Ayurveda, Siddha, Unani, Homeopathy, and Naturopathy are accessible ISMs. The Siddhars are the outstanding researchers in ISM, who have accomplished Ashta-mahasiddi. They have hypothesized, hugely contributed and built up the idea of Tamil Medicinal Systems known as Siddha System

of Medicine (SSM). From the most punctual time, in Southern India, explicitly in Tamil Nadu, SSM has been rehearsed. From one viewpoint, the commencement of the present Medicinal System has massively partial the past SSM and arranged the philosophy of SSM, therefore it is evaporated. Then again, at present, a few gatherings of people are utilizing Siddha Medicine as a central restorative system for a sound life. In this point of view, numerous works manage the beginning of TMS system. At long last, later on, Siddha will be one of the most well known ancient native wellbeings rehearses despite the fact that it is in excess of a couple of pointed difficulties and issues, which is required to be hailed in a proficient way and to be saved and stimulated around the world.



(a)

Figure 1 Handwritten Tamil Characters

The character recognition is the most testing and tempting field, on the grounds that the big innovative work exertion that has gone into it has not unraveled all monetarily dire and mentally fascinating issues. Recognition of manually written characters has been a prominent research territory for a long time in light of its different application possibilities. A portion of its potential application territories are interpretation, postal robotization, bank check preparing, programmed data section, and so on. There are numerous bits of work towards written by hand recognition of Roman, Japanese, Chinese and Arabic contents, and different methodologies have been proposed by the scientists towards manually written character recognition. Character Recognition is a field of research in picture handling, artificial inelegance, design recognition and machine learning. Tamil is the most famous language on the planet and especially in Tamilnadu, India. In excess of 8 crore Tamils live in Tamil Nadu and Pondicherry. Around one crore Tamils live in different conditions of India. Outside India, Sri Lanka, Burma, Malaysia, Singapore, Indonesia, South Africa, Fiji, Mauritius islands are a portion of the nations having countless Tamil talking individuals. In this way, the work on Tamil content is helpful for the Tamil people group the world over. The letter set of the advanced Tamil content comprises of 12 vowels, one Aaydham, 18 consonants and 216 consonantal vowels and thus there is a sum of 247characters in Tamil. The essential characters of Tamil content are appeared in Fig.1. Composing style in Tamil content is from left to right. The idea of upper/lower case is missing in Tamil content. In Tamil content, a vowel following a consonant takes an adjusted shape. Contingent upon the vowel, its altered shape is put at the left, right (or both), upper

side or base of the consonant. A vowel following a consonant in some cases takes a compound orthographic shape, which we call a compound character. These compound letters are shaped by adding a vowel marker to the consonant. A few vowels require the fundamental state of the consonant to be modified in a manner that is explicit to that vowel. Others are composed by adding a vowel-explicit postfix to the consonant, yet others a prefix, lastly a few vowels require including both a prefix and an addition to the consonant.

2. Literature Survey

[1] E. K. Vellingiriraj, M. Balamurugan, and P. Balasubramanie (2016) the proposed framework beats such a circumstance by changing over all the ancient historical documents from engravings and palm manuscripts into Tamil digital text format. It changes over the digital text format utilizing Tamil Unicode. This is the better approach for moving toward Ancient vattezhuthu character recognition, the aftereffect of which is seen as above 90% for consonant and vowel letter recognition, though the consonantal vowel recognition is to some degree low because of the absence of likeness of letters. The general yield of the proposed algorithm is 89.75% exact which higher than that of the current frameworks. By and by, the Brahmi character and vattezhuthu recognition frameworks still contain numerous issues that require increasingly proficient algorithms to understand the three reasons.

Merits

- The first stage transformation precision of the Brahmi content pace of the algorithm is 91.57% utilizing the neural system and picture zoning strategy.
- The conversion precision of Vattezhuthu is 89.75%.

Demerits

- The greatest downside of machine interpretation is low quality, all the time mistaken, and interpretations.
- Machine interpreters can't get context or the imaginative utilization of language, for example, plays on words, analogies, trademarks, bringing about an in exactly the same words interpretation which has neither rhyme nor reason when converted into an alternate language.

[2] Mr R.Vinoth, Rajesh R., Yoganandhan P. (2017) the proposed framework defeats such a circumstance by changing over all the palm manuscripts into Tamil digital text format. The fundamental goal of the framework is to make the Script of Tamil language increasingly open. This framework can be utilized in an assortment of ways relying upon the prerequisite of the client. The understudies may not comprehend the ancient Tamil language that they find in the Palm-leaf manuscript. Be that as it may, this System does. Also, this will push them to a more noteworthy degree and will enable them to be independent. Given a Palm leaf manuscript of Tamil to any individual with the

fundamental information on Tamil language and that could conceivably comprehend the substance, by changing over it to the present Tamil language. This framework can likewise be valuable for chronicling purposes. It is useful in digitizing the Palm leaf manuscript by changing over and putting away it in a digital format. This has been accomplished for putting away the palm leaf manuscript information on the grounds that the palm leaf manuscript was troubled for protecting and putting away and utilizing this store in the framework if the palm leaf manuscript were decimated.

Merits

- The primary goal of the framework is to make the Script of Tamil language progressively open
- It is useful in digitizing the Palm leaf manuscript by changing over and putting away it in a digital format.

Demerits

- Generally, Ancient letter transformation still has a major test.

[3] Punitharaja.K and Dr.P.Elango Proposed a GMM is a model utilizing a lot of six novel features that got from directional vitality conveyances of the hidden picture. The proposed is utilized to the GMM approach towards the recognition of off-line Handwritten. The intricacy of handwritten character recognition system is used increments for the most part due to different composing styles of various people. A large portion of the mistakes in such a system emerges as a result of the disarray among the comparable formed characters. In Tamil, there are numerous comparable formed characters. Tamil Character Recognition, Here two classifiers (GMM and SVM) in view of inclination and bend features are utilized for the recognition.

Merits

- The adequacy of the proposed GMM back features is appeared for character and word recognition tasks, utilizing an SVM arrangement system.

Demerits

- If this is the situation with your information then you should attempt either the ANBC or Support Vector Machine characterization algorithms.
- Another inconvenience of the GMM algorithm is that the client must set the number of blend models that the algorithm will attempt to fit the preparation dataset.

[4] E.K. Vellingiriraj, Dr.P. Balasubrmanie (2017) the proposed strategy, perceiving and prediction of the ancient Tamil characters is finished by executing HMNL-PRS. This strategy performs in a superior manner and delivers better outcome by rectifying the prediction mistake an incentive in each layer at run time. At first, commotions present in the pictures are evacuated by utilizing a median filter in the MATLAB simulation environment. At that point feature extraction is done and it is given as a contribution to HMNL-PRS. In this technique the yield got from each layer is pre-handled to alter the

blunder before sending it to the following layer of the neural system. Thusly, exact learning is done with the goal that prediction exactness gets improved.

Merits

- The proposed strategy precisely perceives the ancient Tamil characters than the current techniques, in this way able answers for powerful restorative treatment can be distinguished rapidly and precisely.

Demerits

- Disadvantages incorporate its "discovery" nature, more noteworthy computational weight, the inclination to over fitting, and the exact idea of model advancement.

[5] S Venkata Krishna Kumar Poornima T V (2014) proposed to peruse the ancient Tamil characters having a place with different periods by testing a limited quantity of characters alluded to as inspected characters in the Tamil language. The Proposed system is intended to facilitate the manual hindrance by helping the PC to comprehend human handwritten characters through a computerized system. The structure for the most part points in the execution of the character period prediction system in which PC will have the option to comprehend a couple of basic directions and recognize the era of these characters. The period prediction of epigraphical content is where it is workable for the client to the PC and has it comprehend or perceive the character. A calculative methodology is proposed here for predicting the Tamil Scripts utilizing the Transductive Support Vector Machine (TSVM).To play out this; the proposed technique utilizes numerous tests like discovering whether the character is on which century.

Merits

- SVM works generally well when there is a clear edge of detachment between classes.
- SVM is progressively powerful in high dimensional spaces.

Demerits

- The primary weakness of the SVM algorithm is that it has a few key parameters that should be set accurately to accomplish the best arrangement results for some random issue.

3. Proposed Work

3.1 Preprocessing

Each time data is gathered for recognition, it is gathered as an optically filtered picture of the paper record. Content is changing over into computerized structure by utilizing a level bed scanner having goals somewhere in the range of 100 and 600 dpi and put away. These pixels may have values: 0 or 1 for twofold pictures, 0–255 for gray-scale pictures and three channels of 0–255 shadings, for example, RGB values for shading pictures. This gathered crude data ought to be additionally investigated to get

valuable data. Pre-processing basically improves the picture for the reasonable division. Such processing incorporates the accompanying:

A. RGB to Gray Conversion

The filtered picture is put away as a JPEG picture however pictures of different configurations like BMP, TIFF and so forth are likewise utilized for recognition. Every one of these pictures is in RGB position are changed over into grayscale, at that point the RGB esteems for every pixel and make as yield an unmistakable worth mirroring the force of that pixel. One such approach is to take the normal of the commitment from each channel: (R+B+C)/3. The estimation of a pixel lies under 0 to 1 or under 0 to 255 contingent on its group.

B. Thresholding/Binarization

Binarization is a method of changing over a grayscale picture into a double picture by utilizing a worldwide thresholding procedure. This picture course of action likewise stores a picture as a network yet can just shading a pixel dark or white. It appoints zero for dark and one for white. At that point, it is upset to acquire picture to such an extent that article pixels are spoken to by 1 and background pixels by0.

C. Noise Reduction

The optical examining gadget or the composing instrument presented the noise which causes disengaged line fragments, knocks, and holes in lines, filled circles and so on. The mutilation together with nearby varieties, adjusting of corners, widening, and disintegration, is additionally an emergency. Middle channel is a procedure that replaces the estimation of a pixel by the middle of gray levels in the area of that pixel.

D. Skew Detection and correction

Skew Detection alludes to the slope in the bitmapped picture of the examined picture. It is generally caused if the paper or palm content isn't bolstered straight into the scanner. Numerous analysts proposed an algorithm to appraise the skew point which is a position edge from the level or vertical course. To expel the skew present in the picture, the content is pivoted into the inverse heading. It must be a zero degree. The skew in the report pictures can be characterized into three distinct sorts, for example, worldwide skew, various skew and non-uniform content line skew.

E. Thinning

Thinning is one of the morphological activity that is utilized to wipe out the picked closer view pixels from the twofold pictures and skeletal the pictures to single-pixel width level with the goal that their shapes are brought out more seriously along these lines, the ascribes to be inspected later and it won't be influenced by the lopsided thickness of edges or lines in the image. Diverse standard capacities are currently accessible in MATLAB for the above activities.

3.2 SIFT Algorithm

The SIFT algorithm takes a character picture and changes it into a lot of nearby highlights, every one of which depicts a neighborhood part around a key point. Every one of these component vectors should be particular and invariant to any scaling, revolution or interpretation of the picture. The SIFT descriptor manufactures a portrayal for each key point dependent on a fix of pixels in its nearby neighborhood. For each example character, the dimensional SIFT highlights are determined. All the SIFT highlights of a specific character are linked to make bigger element space. In the component extraction process, resized singular character of size 120x120 pixels is additionally separated into 54 equivalent zones, every one of size 20x20 pixels. The highlights are removed from the pixels of each zone by moving along their diagonals. This method is rehashed for every one of the zones prompting extraction of 3 highlights for each character. These extricated highlights are utilized to prepare a feed-forward back propagation neural network utilized for performing arrangement and recognition errands. Broad reproduction considers show that the recognition system utilizing corner to corner highlights gives great recognition exactness while requiring less time for preparing.

The primary strides of our technique are:

- Perform different pre-handling tasks to upgrade the nature of ancient Tamil character stone engraving pictures.
- Detection of intrigue focuses in SIFT descriptors.
- Constructing visual codebooks by methods for bunching procedures (K-implies). The codebook is the arrangement of focuses of the educated bunches.
- Constructing a packs of key focuses, which checks the quantity of patches allowed to each bunch.
- Applying a SVM classifier, treating the pack of key focuses as the component vector, and along these lines figure out which character to relegate to the picture.
- Selection of a codebook and classifier giving the best by and large characterization exactness

4. Experimental Results

Intensity Ratio

Table 1 Comparison Table of Intensity RATIO

Transductive Support Vector Machine Algorithm	Gaussian Mixture Model Algorithm	Proposed SIFT Algorithm
57	69.5	83
59	69.9	84.8
62	69.5	87.9
66	70.9	90.2
69	72	93.6

The comparison table of Intensity Ratio shows the different values of Transductive Support Vector Machine Algorithm, Gaussian Mixture Model Algorithm and proposed SIFT Algorithm. When Comparing the Transductive Support Vector Machine Algorithm, Gaussian Mixture Model Algorithm and proposed SIFT Algorithm, the proposed SIFT Algorithm provides the better results. The Transductive Support Vector Machine Algorithm value starts from 57 to 69, Gaussian Mixture Model Algorithm values starts from 69.5 to 72 and the proposed SIFT Algorithm values starts from 83 to 93.6. The proposed SIFT Algorithm provides the great results.

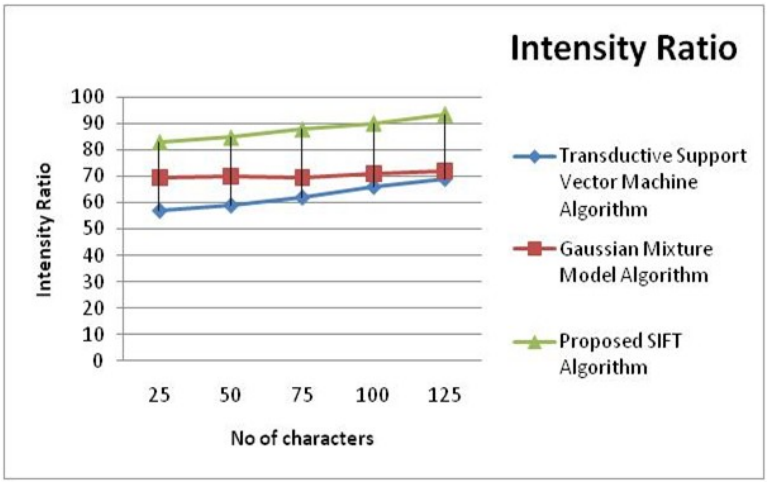


Figure 2 Comparison Chart of Intensity Ratio

The Comparison Chart of Intensity Ratio demonstrates the different values of Transductive Support Vector Machine Algorithm, Gaussian Mixture Model Algorithm and proposed SIFT Algorithm. In this Chart shows the No of Characters in X axis and the Intensity Ratio in Y axis. The Transductive Support Vector Machine Algorithm value starts from 57 to 69, Gaussian Mixture Model Algorithm values starts from 69.5 to 72 and the proposed SIFT Algorithm values starts from 83 to 93.6. The proposed SIFT Algorithm provides the great results compared than other two algorithms.

Detection Ratio

Table 2 Comparison table of Detection Ratio

Transductive Support Vector Machine Algorithm	Gaussian Mixture Model Algorithm	Proposed SIFT Algorithm
39	26.77	66
45	31.98	72
49	34.56	76.5
55	38.92	79.8
58	44.56	85

The comparison table of Detection Ratio shows the different values of Transductive Support Vector Machine Algorithm, Gaussian Mixture Model Algorithm and proposed SIFT Algorithm. When Comparing the Transductive Support Vector Machine Algorithm, Gaussian Mixture Model Algorithm and proposed SIFT Algorithm, the proposed SIFT Algorithm provides the better results. The Transductive Support Vector Machine Algorithm value starts from 39 to 58, Gaussian Mixture Model Algorithm values starts from 26.77 to 44.56 and the proposed SIFT Algorithm values starts from 66 to 85. The proposed SIFT Algorithm provides the great results.

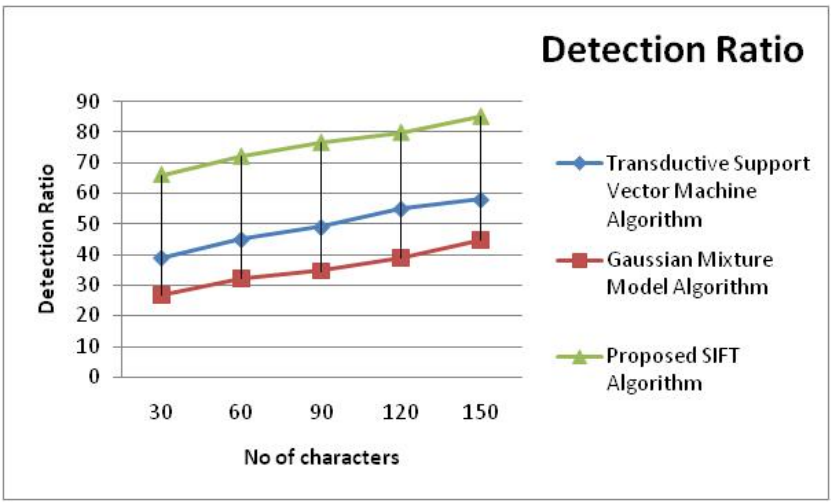


Figure 3 Comparison chart of Detection Ratio

The Comparison Chart of Detection Ratio demonstrates the different values of Transductive Support Vector Machine Algorithm, Gaussian Mixture Model Algorithm and proposed SIFT Algorithm. In this Chart shows the No of Characters in X axis and the Detection Ratio in Y axis. The Transductive Support Vector Machine Algorithm value starts from 39 to 58, Gaussian Mixture Model Algorithm values starts from 26.77 to 44.56 and the Proposed SIFT Algorithm values starts from 66 to 85. The proposed SIFT Algorithm provides the great results compared than other two algorithms.

Conclusion

Ancient Tamil character recognition is the most moved research field in reality which centers on perceiving the ancient Tamil characters with the goal that much data can be educated. The proposed SIFT algorithm is motivated by a straightforward perception that each content or language characterizes a limited arrangement of content examples ,each having a particular visual appearance, and consequently every character could be distinguished dependent on its segregating highlights.

References

1. 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, December 2016.
2. Mr R.Vinoth, Rajesh R., Yoganandhan P. "INTELLIGENCE SYSTEM FOR TAMIL VATTEZHUTTUOPTICAL CHARACTER RECOGNITION", International Journal of Computer Science & Engineering Technology (IJCSET) Apr 2017.
3. Punitharaja.K and Dr.P.Elango," Improving Handwritten Tamil Character Recognition using GMM", International Journal of Pure and Applied Mathematics Volume 118 No. 20.
4. E.K. Vellingiriraj, Dr.P. Balasubrmanie," A Novel Hybrid Neural Learning based Tamil Handwritten Character Recognition System in Palm Manuscripts for Siddha Medicine", Jour of Adv Research in Dynamical & Control Systems, Vol. 9, No. 3, 2017.
5. S Venkata Krishna Kumar Poornima T V," An Efficient Period Prediction System for Tamil Epigraphical Scripts Using Transductive Support Vector Machine", International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 9, September 2014.
6. E. K. Vellingiriraj and P. Balasubramanie, "Automatic digitization of ancient Brahmi characters into Tamil digital texts using image zoning from palm manuscripts and stone inscriptions," in Proc. International Conference on Digital Humanities (CDH2015), The Open University of Hong Kong, Hong Kong, vol. 1, issue 1, p. 49, Dec. 17-18, 2015.
7. Giridharan.R, Vellingiriraj.E.K, Dr. Balasubramanie.P, "Identification of Tamil ancient characters and information retrieval from temple Epigraphy using image zoning" ICRTIT 2016 International Conference on recent trend in Information Technology.
8. Chamila Liyanage; Thilini Nadungodage; Ruvan Weerasinghe "Developing a commercial grade Tamil OCR for recognizing font and size independent text" 2015 15th International Conference on advanced in ICT for emerging region.
9. Honey Mehta, Sanjay Singla, Aarti Mahajan "OpticalCharacter Recognition (OCR) System for Roman Script and English language using artificial Neural Network" 2016 International Conference on Research advanced in Integrated Navigation System (RAINS).
10. Pelin Gorgel, Oguzhan Oztas "Recognition of Handwritten Character using Neural Network" International Journal of innovative research in Computer and Communication Engineering. June 2016 Vol-4 issue-6.
11. Mari, S. Sobhana, and G. Raju. "Modified View Based Approaches for Handwritten Tamil Character Recognition." ICTACT Journal on Image & Video Processing 6.1 (2015).

12. Selvakumar, P., and S. Hari Ganesh. "Tamil Character Recognition Using Canny Edge Detection Algorithm." Computing and Communication Technologies (WCCCT), 2017 World Congress on. IEEE, 2017.
13. Thendral, T., M. S. Vijaya, and S. Karpagavalli. "Analysis of Tamil character writings and identification of writer using Support Vector Machine." Advanced Communication Control and Computing Technologies (ICACCCT), 2014 International Conference on. IEEE, 2014.
14. Urala, K. Bhargava, A. G. Ramakrishnan, and Sahil Mohamed. "Recognition of open vocabulary, online handwritten pages in Tamil script." Signal Processing and Communications (SPCOM), 2014 International Conference on. IEEE, 2014.
15. Punitharaja K and Dr.Elango P, "Accuracy improvement of Off-line Handwritten Tamil Character Recognition", Malaya Journal of Matematik, S (2)(2015) 504–512, 2015.