International Journal of Engineering & Technology, 7 (3.34) (2018) 975-979

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

A Survey on Handwritten Character Recognition Techniques for Tamil Language

Babitha Lincy R[1]*, Gayathri R[2]

[1]*Research Scholar, Electronics and communication Engineering, Sri Venkateswara College of Engineering, Sriperumbudur*
[2]*Professor, Electronics and communication Engineering, Sri Venkateswara College of Engineering, Sriperumbudur*
**Corresponding author E-mail:rblincy@gmail.com*

Abstract

Nowadays the most remarkable and fascinating success of optical character recognition system for the printed Tamil text leads to extremely interesting challenges in the part of handwritten identification tool for Tamil text because of its various research solicitations. Recognition of Tamil character has been broadly studied in the previous year by the academic laboratories and research companies. Presently the researchers have taken their attention towards various recognition methodologies for Tamil handwritten character recognition due to its open challenges. The aim of this survey is to exchange the different aspects of Tamil OCR and resolving the issues of Tamil handwritten OCR by conferring performance of the various techniques.

Keywords: *Deep Learning, Handwritten Recognition, OCR survey, Optical Character Recognition, Tamil language.*

1. Introduction

Optical Character Recognition (OCR) also known as Optical Character Reader, which is conferring the mechanized recognition of photographic document for many languages. Practically the OCR reader tool transfigures the script, which is written by hand or typewritten, printed script format into machine copy-edit text or computer processable format (ASCII), whether from a resemble scan document or a camera image. [1]. The OCR is used in following applications such as automatic data entry, baking, voice synthesizer, the reading device for visually challenged people [2]. Pattern recognition includes two types, they are online and offline [3] as shown in Fig.1

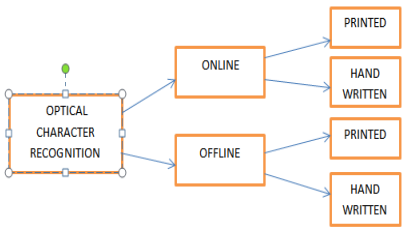


Fig 1: OCR types

Tamil is one of the oldest- preponderant surviving, and predominance has spoken language among the Dravidian language family by the Tamil peoples of India, Sri Lanka, Singapore, and Malaysia. The Tamil alphabets are be formed with 12 vowels, one Aayudham, 18 consonants, and 216 compound characters and hence Tamil has the total of 247 characters, also about 6 Grantha characters are also present in the Tamil language.[3]. The some of the basic Tamil characters are shown in Fig.2. Some of the formations of compound characters are shown in Table.1. Similarly, all the compound characters are formed.

Table 1: sample of compound character formation

Vowel + Consonant	Compound character
அ + க்	க
ஆ + க்	கா
இ + க்	கி
ஈ + க்	கீ
உ + க்	கு
ஊ + க்	கூ
எ + க்	கெ
ஏ + க்	கே
ஐ + க்	கை
ஒ + க்	கொ
ஓ + க்	கோ
ஔ+ க்	கௌ

The objective behind the Tamil Handwritten Recognition is pointed out the Tamil characters from a scanned image of the handwritten document, which is more strenuous task than printed Tamil character recognition due to miscellaneous writing styles. Samples of Tamil handwritten character are shown in Fig.3.



Fig.2: samples of Tamil language



Fig.3: Tamil Handwritten samples

In this paper section II narrate the basic system architecture of OCR recognition concept, section III give report about literature survey, section IV will give brief about future scope and open challenges, section V will give the conclusion about this review.

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

2. Architecture

The main intent of the Tamil handwritten recognition system is the conversion of the handwritten photographic document into digital format in order to make its data explorable and editable. The basic transfiguration architecture [4], as shown in Fig.4. The representative architecture design is composed of the series of four stages, which cover preprocessing, segmentation, feature extraction, and classification stages.

2.1. Image Acquisition

To recognize the Tamil handwritten character firstly, the images for the system might be acquired by scanning the handwritten document, which is stored in some of picture file format preferably such as TIF, BMP, GIF or JPG. The image acquisition also known as digitization, which digitized image, is given as input for next processing stages.

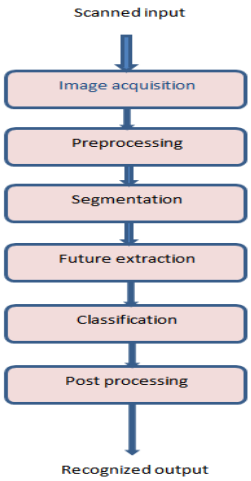


Fig.4: Basic architecture

2.2. Pre-Processing

The preprocessing step is used to adapt the digitized image into a suitable format, which involves a series of computations and operations to make it suitable and properly furnished image with free of any irregularities for processing. The preprocessing approach is handling the necessary steps to make data that help to the easy accessing and good accuracy for the system. The preprocessing approach has some of the following necessary operations. They are:

- RGB to grey conversion
- Binarization
- Skew detection and correction
- Image cropping
- Resampling
- Slope correction
- Noise removal
- Skeletonization
- Image resizing
- Image thinning
- Smoothing
- Normalization
- Slant removal
- De-blurring
- Image complement

In this step, RGB image is transfigured to greyscale image, with modern computer's parallel programming, due to it's not possible to perform simple pixel- by- pixel processing with high speed for character recognition by using the RGB image, since which has three color channel. But in the grey scale, image processing time is three times lesser than RGB image processing time, since which has a single color channel. When if we were analyzing thousands of images from a dataset, it is great to save the processing time. The output is shown in Fig.5 (a). In the next binarization step to make more intelligible the recognition work, the grey scale image [0,255] converted to true black or white image, called the binary image with only two unique value (0 or 255),[5]. The binarized output shows in Fig.5 (b).

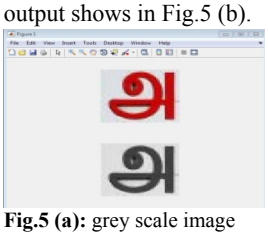


Fig.5 (a): grey scale image

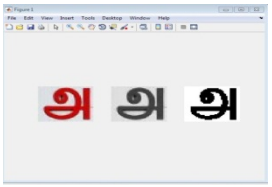


Fig.5 (b): binary image

If the content is not aligned satisfactory, this is slant i.e. possibilities of the image getting right or left orientation, due to improper scanning of the input document, which has some angle variations. So we can modify the photographic image with simple rotation in a particular angle by applying skew detection and correction algorithm [6]. By manipulating image cropping technique we can do the act of selecting and extracting the Region of Interest (or simply, ROI), which is the part of the image we are interested in; it will crop the surplus pixels and blank spaces from the input image. Resampling engaged a small but important place in the field of preprocessing, which is the action of transfiguring a sample image from one coordinate system to another.

The image intensity variation or unwanted pixel the presence due to image acquisition process, which is called presence of noise in the image. The main aim of this step is removing the imperfect or undesired information from the digital image by using different types of filters based on our needs and importance [7]. Skeletonization is the help to obtain the structure of character without affecting general shape. The thinning preprocess is the key point to get the single pixel structure width to recognize the character easily. After some preprocessing step the image size will vary, so we will perform resizing the image based on our needs.

Smoothing is the process of reducing the number of high-frequency components by the Gaussian filter and deleting the bumps and spurs from text character. The main objective of normalization is acquiring the standard size data, which will transform random size document into a systematized document. To do thinning operation we should do image complement, which involves image conversation.

2.3. Segmentation

We have studied for the occasion of the development of segmentation algorithm which produces an excellent high precision segmented output and high recall of all actual character boundaries [8]. Overall a variety of segmentation techniques are introduced which is used to subdivide the document image into lines, words, characters, individual symbols, and those effects are examined. The accuracy of the recognition system is based on the accuracy of the segmentation.

2.4. Feature Extraction

In the character recognition application selection of a suitable feature is a difficult task, which will vary from application to application.[9]. Feature extraction is the requisite part of the recognition system. The categories of feature extraction are local and global.[10]

2.4.1. Local Features

The local features are extracted from the small textual components.[10]

Types:

a) Statistical,

b) Structural,

c) Hybrid

2.4.1.1. Statistical

This pluck out the features like width, height, and area of the pixel. Some of the statistical feature techniques are Horizontal projection profiles, Water reservoir based feature, bounding box, character pitch, zoning [10].

2.4.1.2. Structural

Using this approach we can identify dimensions, oops, cusps, endpoints etc.., which is also known as geometric features. Some of the common structural features are Headline feature, Fractal feature, Topological feature, Morphological feature [10].

2.4.1.3. Hybrid

The hybrid features are strongly sensitive to the slant and strip.

2.4.2. Global Feature

The global features consider DCT, DWT, Gabor etc., usually these global feature derived from the texture.[10] Histogram of projection based on mean distance, pixel value, Vertical zero crossing are the other major significant feature extraction techniques [2].

2.5. Classification

Classification stage basically deals with the categorizing the character into the particular class, which will find out the best matching class input by comparing the input features with stored pattern. Some of the major classification approaches are Neural Network, Decision Tree, and SVM.

2.6. Post-Processing

Post-processing prints the recognized characters in the editable text. Moreover, it will increase accuracy by correcting the misclassified output. Some of the post-processing approaches are linguistic knowledge and dictionary-based approaches.

3. Literature Survey

This survey assigns the various approaches regarding the Optical Character Recognition tool for the Tamil language document. Over the few preceding years, Indian scientists presented much expertness to optimize the performance of the OCR tool. No significant work is directed towards the overlapped and blurred document of the Tamil language. Earlier efforts have been more concentrated on online OCR. Some of the previous works are discussed below.

Daniel Keysers [11] et al, this paper deals with the online handwriting recognition tool for Google. In this, the authors have mainly focused on the online handwriting recognition tool for 22 scripts and 97 languages including the Tamil language also. Moreover, it is applying the novel techniques, such as unified time- and position-based input interpretation, trainable segmentation, minimum-error rate training for feature combination, and a cascade of pruning strategies. However, it was concluded that at the present time this tool is accessible in several Google products.

S. M. Shyni, M. Antony Robert Raj et al [12], in this exploration Sub Line Direction and Bounding Box are advised for recognition of Tamil Handwritten Character in offline mode. In this, Zoning and Chain Code policies has been recommended for feature selection and Sub Line Direction and Bounding box policies are applied for extracting the features. The key value of this recognition instrument was based on the Support Vector Machine (SVM) for best learning. A recognition precision value of 88% has been reported on 30 Tamil character sets.

In [13], a commercial grade Tamil OCR device to acknowledge the Tamil characters for dissimilar font and size document is proposed. Tesseract engine has been used for recognizing the Tamil characters. Also, the characters are segmented by creating box files in Tesseract in this work. Anyhow this work carried out 81% accuracy for using 20 scanned images taken from 20 ancient Tamil books.

In [14], machine learning techniques are preferred for automated ancient Tamil script recognition. It is based on the K-Nearest Neighbor (KNN) model. Shape and Hough transform are chosen for feature extraction, Group Search Optimization and Firefly algorithm are used for feature selection to identify the ancient Tamil script. It is found that classifiers such as K-Nearest Neighbor (KNN), NN and J48 classifier are used for identification.

A. G. Ramakrishnan, Bhargava Urala K[15] et al, in this paper the recognition of Handwritten Numerals and Tamil Characters in online mode is designated which is based on Global and Local feature selection tactics. Handwriting recognition experiments are using the global features, local features and the combination of both for identification of characters. The SVM classifier is considered to develop the ideal identification tool. It is found that the precision rate for recognition more than 95% on the dataset.

Giridharan.R, Balasubramanie.P, Vellingiriraj.E.K, et al [16], serviced on the identification of Tamil antique characters and retrieval of character from temple epigraphy by image zoning approach. They proposed some hybrid algorithms for increasing the recognition rate and they obtained more than 90% accuracy for temple epigraphy. The proposed system converts the recognized Tamil digital text with meaning.

Suresh R.M, Jagadeesh Kannan.R, in [17], utilized the innovative technology, Hidden Markov Models (HMM) for Cursive Handwritten Tamil Character identification in offline manner. On experimentation with the collected dataset of different no of sample dataset the overall precision rate observed was better than the previous works for Cursive Handwritten Tamil Character identification in offline mode.

Alejandro H. Toselli, Moises Pastor, and Enrique Vidal [18] et al, this paper describes that identification tool for Tamil language handwritten characters in online mode generally based on the feature selection by time and frequency domain approach. The system was trained and tested with the labeled "hpl-tamil-iwfhr06-train-online" for training/validation purposes and the unlabeled "hpl-tamil-iwfhr06-test-online" for the final test. Further improvements are achieved by finding the error rate is around 9%.

R.Seethalakshmi.R, 2005 [19] et al, this study preferred the Optical Character Recognition for printed Tamil character using the Unicode way, where the documents are classified by Supervised Learning Algorithm. The suggested system is utilizing the Support Vector Machine (SVM) for the classification and finally, these classes are plotted onto Unicode for identification of printed Tamil text. Furthermore, the acknowledged character is reassembled by the Unicode fonts.

Elakkiya.V, in her work on [20], presented a new approach for classification, which is known as KNN classifier for Tamil Text Recognition, which has increased the speed and accuracy of character recognition system. Finally the experimental result Tamil text is translated into the English language. The overall identification rate of Tami text was found that 91%

Various approaches have been explained in this review for Tamil handwritten character recognition system. From the study done so far, a sort of comparison between different techniques among with their result which have been proposed by the researchers is shown in Table 2.

Table 2: Comparative analysis of various recognition techniques

Reference no	Methodology
[11] Daniel Keysers, Thomas Deselaers, Henry A. Rowley, Li-Lun Wang, and Victor Carbune	Image acquisition : Mobile phone camera Data set : Synthetic(synth) Concept : Scene text recog-

	nition, Optical music recognition Preprocessing :Resampling, slope correction Segmentation : Trainable segmentation Future extraction : Pointwise features, Character-global features Classification : ConvolutionalRecurrent neural network Post processing : - Accuracy : good
[12] S. M. Shyni, M. Antony Robert Raj and S. Abirami	Image acquisition : Printed document Data set : HP India lab Concept : Offline Tamil Handwritten Character Recognition Preprocessing : Binarization, skeletonization Segmentation : - Future extraction : Bounding box algorithm Classification : SVM Post processing : - Accuracy : 88%
[13] Chamila Liyanage, Thilini Nadungodage, Ruvan Weerasinghe	Image acquisition : 20 scanned images Data set : - Concept : recognize the Tamil characters with font and size independent text Preprocessing : Training the data with Tesseract engine Segmentation : creating box files in Tesseract Future extraction : - Classification : - Post processing : - Accuracy : 81%
[14] T S Suganya, S Murugavalli	Image acquisition : Ancient Tamil script image Data set : - Concept : Automated Ancient Tamil Script Classification System Using Machine Learning Preprocessing : Noise removal, Binarization Segmentation : line segmentation Future extraction : Shape and Hough transform Classification : Neural Network, J48, Naïve Bayes and KNN Post processing : - Accuracy : good accuracy(proved)
[15] A. G. Ramakrishnan, Bhargava Urala	Image acquisition : online data Data set : HP Labs Concept : Recognition of Online Handwritten Numerals and Tamil Characters Preprocessing : - Segmentation : - Future extraction : global, local features and combined Classification : SVM Post processing : - Accuracy : 95%
[16]Giridharan.R, Vellingiriraj.E.K, Dr. Balasubramanie.P	Image acquisition : camera image Data set : collected dataset Concept : Identification of Tamil Ancient Characters Preprocessing : image cropping, image resizing, image thickening, image binarization. Segmentation : grapheme extraction
	Future extraction : - Classification : image zone Post processing : - Accuracy : more than 90%
[17]R.Jagadeesh Kannan, R.Prabhakar and R.M. Suresh	Image acquisition : scanned image Data set : collected data Concept : Off-Line Cursive handwritten Tamil character recognition Preprocessing : Binarization, Noise Removal Skew Correction Segmentation : lines, words, and characters Future extraction : Time-domain, frequency-domain features Classification : HMM Post processing : - Accuracy : good
[18] Alejandro H. Toselli, Moises Pastor, and Enrique Vidal	Image acquisition : online data Data set : HPL online data Concept : On-Line Handwriting Recognition System for Tamil Preprocessing : noise reduction, normalization Segmentation : trace segmentation Future extraction : Time-domain, frequency domain features Classification : HMM Post processing : find the error rate Accuracy : moderate
[19] Seethalakshmi.R, Sreeranjani T.R, Balachandar T	Image acquisition : scanned image Data set : Unicode Standard (http://www.unicode.org) Concept : Recognition for printed Tamil textusing Unicode Preprocessing : Binarization, skewing Segmentation : vertical histograms, horizontal histograms Future extraction : character height, character width, curves, the number of circles, number of slope lines Classification : SVM Post processing : the errors, cost, and delay of manual data entry are reduced Accuracy : good
[20] V.Elakkiya, I.Muthumani, M.Jegajothi	Image acquisition : scanned image Data set : - Concept : Tamil character recognition Preprocessing : Binarization Segmentation : lines, words, and characters Future extraction : local binary pattern Classification : KNN Post processing : translation to the English language Accuracy : 91%

4. Future Scope and Open Challenges

The Optical Character identification (OCR) tool will make a prosperous solution to the many real-world problems such as data entry problems and the computer vision industry. Therefore, OCR tools are being evolved for almost all major languages in the world level and the Tamil language is also no exception to it. During the past years, considerable research and development works have been done towards the development of an efficient Tamil character recognition system.

Handwriting also an important tool for communication between people. Nowadays the handwritten documents were replaced with

digital and mechanical technology. The Tamil handwriting identification (HWR) tool can integrate with tablets and other devices to give the benefit of handwritten notes and letters, but not the storage of paper and overflowing file cabinets.

A comprehensive study on Tamil character recognition proposed modes by various researchers has been made, whereas the proposed model includes printed, handwritten recognition. As we have seen that handwritten document recognition has more open challenges than printed text recognition due to various writing styles, low-resolution input, braked images, and overlapped image. Hence, there could be ways to increase the recognition accuracy of handwritten recognition which may be carried out in the future. Consequently, future research should exploit this survey work and benefit the good effect of each approach while removing the unwanted effects to increase the system efficiency and proposed a novel system that meets the expectations.

5. Conclusion

This is a brief description of Tamil handwritten recognition techniques and includes its various approaches in recognition. From the review of various research papers, we have concluded that the selection of suitable classification technique plays a vital role in recognition rate. This survey will guide and provide the update for researchers working in the Tamil character recognition field. By studying Tamil character recognition we can decide that as per as technology is developing day by day the need for deep learning is increasing because of only good performance. If we talk about our proposed research presents an offline Tamil handwritten character recognition system techniques using the deep learning algorithm to improve the performance of overlapped and low-resolution handwritten images.

References

[1] Sukhpreet Singh, Ashutosh Aggarwal, Renu Dhir, "Use of Gabor Filters for Recognition of Handwritten Gurmukhi Character " International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 5, May 2012.

[2] Divakar Yadav, Sonia Sanchez-Cuadrado, Jorge Morato, " Optical Character Recognition for Hindi Language Using a Neural-network Approach" Journal of Information Processing System,Vol.9, No.1.January 2013

[3] K. Punitharaja, and P. Elango, Tamil Handwritten Character Recognition: Progress and Challenges I J C T A, pp. 143-151 2016.

[4] Jagruti Chandarana, Mayank Kapadia, Optical Character Recognition, International Journal of Emerging Technology and Advanced Engineering, Volume 4, Issue 5, May 2014.

[5] T.S. Suganya, Dr. S Murugavalli, Feature Selection For An Automated Ancient Tamil Script Classification System Using Machine Learning Techniques, Algorithms, Methodology, Models and Applications in Emerging Technologies (ICAMMAET), Feb. 2017.

[6] Seethalakshmi R, Sreeranjani T.R, Balachandar T, Optical Character Recognition for printed Tamil text using Unicode, Journal of Zhejiang University SCIENCE, pp. 1297–1305 November 2005.

[7] Dimple Bhasin, Gulshan Goyal, Maitreyee Dutta, Design of an Effective Preprocessing Approach for Offline Handwritten Images, International Journal of Computer Applications (0975 – 8887) Volume 98– No.1, July 2014.

[8] Daniel Keysers, Thomas Deselaers, Henry A. Rowley, Li-Lun Wang, and Victor Carbune, Multi-Language Online Handwriting Recognition, IEEE transactions on pattern analysis and machine intelligence, vol. 39, no. 6, june 2017

[9] Sheikh Faisal Rashid, Marc-Peter Schambach, Joerg Rottland, Low-resolution Arabic recognition with multidimensional recurrent neural networks, Proceedings of the 4th International Workshop on Multilingual OC August 2013

[10] Kurban Ubul, Gulzira Tursun, Alimjan Aysa, Donato Impedovo, Giuseppe Pirlo, and Tuergen Yibulayin, Script Identification of Multi-Script Documents: A Survey, IEEE access, March 30, 2017.

[11] Daniel Keysers, Thomas Deselaers, Henry A. Rowley, Li-Lun Wang, and Victor Carbune, Multi-Language Online Handwriting Recognition, IEEE transactions on pattern analysis and machine intelligence, vol. 39, no. 6, june 2017

[12] S. M. Shyni, M. Antony Robert Raj and S. Abirami, Offline Tamil Handwritten Character Recognition Using Sub Line Direction and Bounding Box Techniques, Indian Journal of Science and Technology, Vol 8(S7), 110–116, April 2015

[13] Chamila Liyanage, Thilini Nadungodage, Ruvan Weerasinghe, Developing a commercial grade Tamil OCR for recognizing font and size independent text2015 International Conference on Advances in ICT for Emerging Regions (ICTer) :24th - 26th August .

[14] T S Suganya, S Murugavalli, Feature selection for an automated ancient Tamil script classification system using machine learning techniques, 2017 International Conference on Algorithms, Methodology, Models and Applications in Emerging Technologies (ICAMMAET), 16-18 Feb. 2017.

[15] A. G. Ramakrishnan, Bhargava Urala K, Global and Local Features for Recognition of Online Handwritten Numerals and Tamil Characters, Proceedings of the 4th International Workshop on Multilingual OCR,Article No. 16, August 24 - 24, 2013.

[16] 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 (ICRTIT), 8-9 April 2016.

[17] R.Jagadeesh Kannan, R.Prabhakar and R.M. Suresh, Off-Line Cursive Handwritten Tamil Character Recognition, International Conference on Security Technology, 2008.

[18] Alejandro H. Toselli, Moises Pastor, and Enrique Vidal, On-Line Handwriting Recognition System for Tamil Handwritten Characters, Proceedings of the 3rd Iberian conference on Pattern Recognition and Image Analysis, Part I, July 2007.

[19] Seethalakshmi.R, Sreeranjani T.R, Balachandar T., Optical Character Recognition for printed Tamil text using Unicode, Journal of Zhejiang University science, 2005.

[20] V.Elakkiya, I.Muthumani, M.Jegajothi, Tamil Text Recognition Using KNN Classifier, Advances in Natural and Applied Sciences, 2017 May 11(7): pages 41-45.