Journal of Innovative Image Processing (ISSN: 2582-4252)
www.irojournals.com/iroiip/

An Empirical Analysis of Tamil Optical Character Recognition

B. Radhika

Assistant Professor, Computer Science Department, SNS College of Engineering, Coimbatore

E-mail: radhikabalamurugan1988@gmail.com

Abstract

The concept of visual cues in a paperless workspace is based on converting scanned images into machine-readable text. It requires the development of a variety of new applications, such as automated postal systems, banks, institutions, word processing, and library systems, to name a few. Artificial Intelligence (AI) is an area of computer science that focuses on making machines smart. Hand typing, script, and print text recognition are key study subjects because no 100% recognition can be achieved even if the scanned image is accurate. Text, numbers, and images can all be used to write text in a variety of ways. The basics of Tamil text, previous tasks in Tamil text attention, Tamil letter recognition algorithms, and recognition barriers are covered in this study.

Keywords: Artificial intelligence, optical character recognition, tamil language, tamil script, visual cues

1. Introduction

In Tamil, vowels and consonants make up the majority of Tamil grammar. As a result, there are 31 independent letters and 216 combination letters in the Tamil text, totaling 247 consonant-vowel combinations, non-verbal consonants, and vowels alone. The consonant is marked with a vowel mark to produce these related characters. Tamil text is written in a left-to-right direction. The Tamil text, as well as other Indian manuscripts, were created using the Brahmi text, also known as Tamil Brahmi. Tholkappiyam is an ancient Tamil grammar written in a brahmi-like form. Tamil letters had evolved into the original vanuatu shape. The dot in the statement's end did not fully exist until printing was developed, despite the fact that the sound itself persists and plays an important part in the Tamil alphabet. Throughout the nineteenth century, a few letters, as well as a few unique forms, were changed to make typing easier. Standard Tamil grammar, which is based on official speech, has very specific rules

DOI: https://doi.org/10.36548/jiip.2022.2.006
Received: 28.05.2022, received in revised form: 25.06.2022, accepted: 06.07.2022, published: 18.07.2022

regarding when and when not to express. Discover engine, morphological analyzer, grammar development tool, reading tool, word recognition, spelling check, speech tag section, speech recognition, font converters, word processor, search engine, unicode conversion etc. are a few existing Optical Character Recognition (OCR) tools.

This programme transforms scanned photos into machine-readable images. Tamil characters are visualized in this text. Without the use of artifact scanners, OCR is used to translate text, compress it, and employ typewriters, text to speech, and text-based algorithms. OCR has a sub-discipline called pattern recognition research. Based on character classification, OCR scans the webpage and produces informed guesses about the sequence of letters that make up words. OCR has a big impact in the paper industry; however, it has a hard time recognising sophisticated images in such scanned papers with complicated backgrounds, corrupted images, loud sound, paper skew, image distortion, low resolution, and grid-interrupted images. All these factors have an impact on precision. OCR is becoming increasingly used as a low-cost method of creating digital forms and handwritten documents, as well as a link between printing, and electronic storage and editing. Many offices have hundreds of pages of printed material, making it tough to keep track of it all. It's also difficult to refer to them when needed. In such offices, this programme is beneficial.

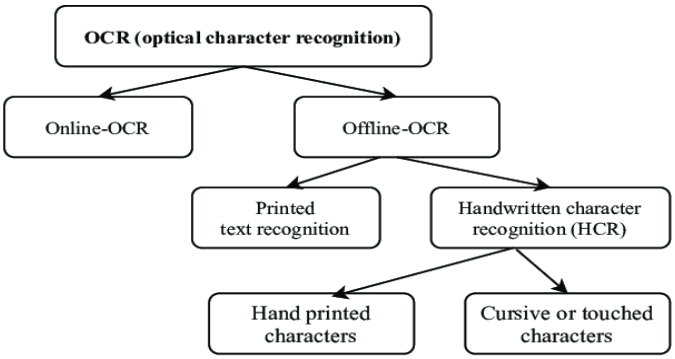

Figure 1. Optical Character Recognition - Types

Recognition methods include handwriting and written text. Figure 1 shows the categories of OCR. Tamil typed text recognition isn't always accurate, even when a clear image is available. Online character recognition detects a character's dynamic movement during handwriting, whereas offline character recognition recognises a static character. These device algorithms take advantage of the fact that each line segment's order, speed, and direction are known prior to installation. It's also possible to retrain the user to only use certain types of characters. Handwritten text identification is frequently an issue because

these methods are incompatible with scanning equipment for paper manuscripts. Excellent handwritten letters can achieve degrees of precision ranging from 80% to 90%, although that is the highest level. The technology is only applicable in a few applications due to the high amount of errors on each page. The accuracy measurement supplied is influenced by the precise measuring method utilized. A 1% error rate (99 percent accuracy) could be as high as 7% error rate (90 percent accuracy) if the scale is based on erroneous letters if the context of a word (really a dictionary) is used to correct software that detects missing words. In commercial OCR software, the accuracy of OCR letter to character conversion varies between 71% and 100%.

Integrated text recognition, perhaps more than handwriting, is a massive research facility with low recognition criteria. Higher levels of literacy are rarely achieved without the use of contextual knowledge or grammar. Individual letters from a script, for example, are easier to read than entire words from a dictionary. Using a small dictionary to read the value check line (which is usually a textual image) is an example of how using a small dictionary can help with recognition. Knowing grammatical grammar will help you figure out whether a word is a noun or an action, giving you more clarity. The characters of each handwritten character do not work. With this detailed introduction, Section II discusses the related works, Section III describes the contribution of OCR in multilingual context, and Section IV illustrates the methodology to implement OCR followed by the conclusion in Section V.

2. Tamil Script Recognition: Related Works

It's hard to be sure of the difference between Tamil old manuscripts and texts. The Tamil alphabet of the past differs from that of the twenty-first century. MATLAB is used to find and translate a century's worth of ancient Tamil characters into modern formats. The research [1] used contour-let transform, a recently released approach for recognising Tamil characters in stone texts. HMM and Statistical Dynamic Time Warping (SDTW) was used as separators to compare the results of the online Tamil handwriting test. A freelance writer might be identified thanks to the massive database that mixes a variety of handwriting styles [2]. Support Vector Machine, Editing Maps are some of the detection features. The supervised reading algorithms are used to separate the characters. According to the structural analysis, the proposed RCS system with a back distribution network has a high level of recognition. A multitude of methods are used to recognise the Tamil character in different systems. On the other hand, recognising Tamil characters on Tablet PCs is a difficult task. In

the past, SVM, pre-processing, translation, and other approaches were used, but they were insufficient in terms of ease of use [3].

A new pattern that was aligned with the Tamil Character Recognition algorithm was introduced. The digitally written character is confronted with a variety of functionalities in the technique. The pre-processing phase includes scanning, modification, grayscale conversion, and cutting. In the intermediate step, neural networks were used to implement background scattering, which is then followed by pattern-like algorithms [4]. A back distribution network was used for character recognition and how to do so with the MATLAB Neural Network toolbox. This is a more advanced version of the MATLAB Neural Network character recognition software included with the toolbox. The study looks at how effectively a neural network can handle distorted Tamil language characters [5]. Multi-phase system methods were described to recognise handwritten notes on the white board. As a result, the ROVER architecture was hired to integrate the output sequences of their cognizers. It is demonstrated how to recognise handwriting biometric text that is suitable for short text sequences (words) [6]. The script and traffic signal reading applications are also implemented on vehicle to infrastructure communications for enabling the safety of intelligent transportation systems, by merging it with a heterogeneous communication network [7, 9].

Episodes are handwriting structural units, with sentences are separated into two sequences. Unsupervised categorization using structured maps allows for the conversion of strokes into numbers and successful comparisons of sequential variables over time. In a recent phase, the results of each sequence were gathered. A new approach for detecting the Tamil character based on the Kohonen neural network-based Self Organizing Map (SOM) method was developed. This is considerably superior to a traditional neural network in terms of results [8]. The interclass relationship of printed Tamil characters was studied and a method for recognising them was suggested. The Multiclass Hierarchical Support Vector Machines are a new type of support vector machine [10]. In everyday life, handwriting is still used to write down information. Separation and recognition can be difficult, especially when working with handwritten materials in multiple languages. The proposed concept was a method for recognising Tamil script text, which is spoken in India, Sri Lanka, and Singapore. To recognise Tamil handwritten characters without cursive, Hidden Markov Models are utilized (HMMs). The system's tolerance is demonstrated by its ability to deal with obstacles posed by a wide range of writing styles while being adaptable and durable. On a large website, the method resulted in a high level of accuracy [11].

The introduction of an online recognition system for Tamil handwritten letters was given in paper [12]. It suggested that sequential stripes make up a handwritten letter. A stroke is a character unit of the shape features in a structure- or shape-based representation. To retrieve an unknown stroke, this character unit representation was compared to a stroke webpage using a dynamic character unit matching technique. The letter stroke is seen in its entirety. A finite state automaton was used to determine characterization. In other Indian scripts, the emergence of comparison systems was detailed [12]. For online identification of Tamil handwritten characters, the fractal coding strategy was proposed, and it offered a unique way to improve coding efficiency over time and effort. This approach takes use of data duplication. It also cuts down on the amount of time it takes to code and lowers distortion during reconstruction [13]. The recognition of handwritten characters is an attractive research topic thanks to the practical application explained in paper [14]. The procedure of detecting Tamil handwriting online was extended using Kohonen's Neural Network. Any input source can teach the system a unique writing style, which is then preserved as images on a website.

In [15], using a space called distance from the outline and the proper membership function, the odd Tamil handwriting was identified as one of the sample characters. It was pre-processed and tested to detect anonymous characters or prototypes. Using abstract principles, attempts were made to recognise Tamil handwritten letters. Characters were identified using the Kohonen Network technique after converting collected attributes to Self-Organizing Maps (SOM). Character recognition can be used in document analysis to turn a handwritten manuscript into an organized print document. The approach was used to identify and recreate handwritten documents in South African languages [15].

3. System Methodology

An optical scanner is used to scan a print. Sheet feed scanners are better at OCR than flatbed scanners since they can scan many pages sequentially. Most recent OCR software scans each page automatically, read the text, and then scan the following page. Figure 2 depicts the architecture for optical character recognition.

3.1 Preprocessing

Binarization is the initial step of OCR, and it entails converting a color or scanned gray material into a black and white translation (two colors / one bit). Any white will be a

part of the background. The first step in determining which text to process, is to convert the image to black and white, however this can cause issues. Five worldwide strategies were studied and a set of evaluation criteria has been come up with. Article [3] studied 19 distinct approaches, whereas compared only three. Several sorts of algorithms were discussed in [18]. A skew is any picture transfer from source paper. Skew identification is one of the first activities performed on scanned papers when transferring data to digital representation. Skew editing is still a crucial aspect of the document-processing process. Images recorded by both digital and classic film cameras from a range of sources will feature sound. A medium filter used for removing salt and pepper from pictures, has the disadvantage of distorting image angles and tiny lines. One of the greatest Median filters is the Center Weighted Median (CWM) filter [17]. The Progressive Media Transmission Filter was recommended by PSMF [16]. The majority of contemporary dynamic filters [17, 19] generate acceptable effects at low noise levels, although they aim to repair extremely deformed images. To extract salt and pepper, a complete modification of the standard was applied [20]. In the realm of picture restoration, algorithms are groundbreaking. Swedish wavelets of the second generation have shown to be beneficial in a range of image processing applications. The lifting approach was previously employed in continuous photography, as shown in [19]. [15 - 18] show how Lifting Scheme variables have been employed in image compression and reconstruction. Although it has lately gained popularity, the notion of noise reduction employing lifting filters is not new.

3.2 Segmentation

In the image of each page, segmentation recognises the text character by character, word by word, and line by line. Various clustering algorithms for image segmentation are investigated in the work [16]. There are three types of techniques to choose from:

- Projection profile
- Hough transform
- Thinning based

As a broad approach of splitting lines of text, a global horizontal examination of black pixels has been applied. To distinguish text pages in different languages, many scholars utilize partial or partly horizontal analysis of black pixels as a modified global hypothesis.

The Hough Transformation has been utilized in a variety of document analysis applications, including skew detection, slant detection, and text line separation.

3.3 Feature Extraction

Each glyph image is evaluated and downloaded for character length, breadth, horizontal lines, straight lines, slope lines, circles, arcs, and other descriptors. Choosing the appropriate technique to delete the feature is one of the most crucial aspects of earning a positive performance appraisal. Several strategies for extracting character identification traits have been published in the literature. Gabor is one of the system's most widely used features.

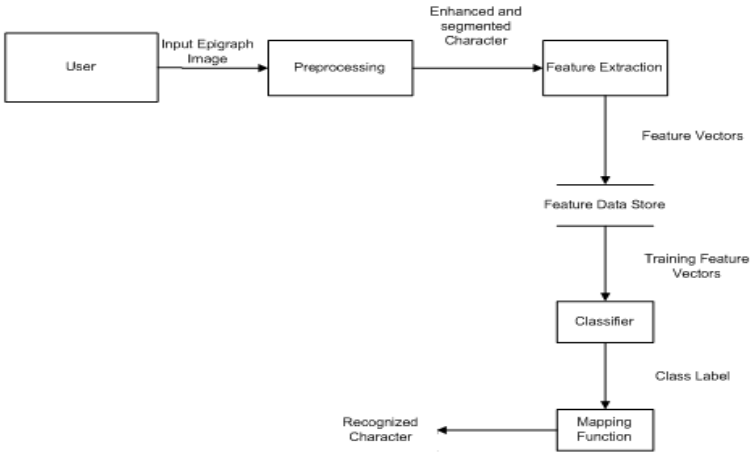

Figure 2. Optical Character Recognition Architecture

3.4 Classification

The difficulty of recognising a group where the fresh observations are a part of it while the identification of a small number is unknown, is based on a training data set that includes the visually impressive section of the population. Support Vector Machines, K nearest neighbors, hybrid Gaussian model, Gaussian, Naive Bayes, Decision Trees, and widely utilized RBF separators are all examples of neural networks (multi-layer perceptrons). Classification algorithms include line dividers (Fisher line discrimination, retrospective, Naive Bayes classifier, Perceptron), Neural Networks, etc. As a backend, synthetic neural has been used. Neural networks have developed as fast and reliable methods of segmentation in offline recognition systems, leading to high accuracy recognition. Since the 1990's, classification techniques have been used to improve the recognition of handwritten characters.

3.5 Basic error correction

Some systems allow you to preview and modify each page separately: they process a full page rapidly and employ a built-in spell checker to flag any misspellings that could cause a disagreement, allowing you to amend an error automatically. To aid in the detection of errors, advanced OCR systems contain extra error checking features. Some computers, for example, utilize a technique called neighbor analysis to find words that may appear nearby, thus a text misidentified as "barkingdog" may be automatically transformed to "barking dog" (because "barking" and "dog" are two words that often run together).

3.6 Layout analysis

Fine OCR algorithms automatically detect complex page layouts, such as several columns of text, tables, pictures, and so on. Images are converted to graphics automatically, tables are converted to tables (hopefully), and columns are properly categorized (so the first row of the first column is not automatically linked to the first row of the second column).

3.7 Proofreading

Even the most advanced OCR software is bad, especially if you work with older documents or lower quality printed text. That is why good, old-fashioned human testing should always be the final step in the OCR.

4. Applications of OCR in Indian Multilingual Context

In terms of research, design, development, and validation, character recognition programmes in India have various challenges. The eighth schedule of the Indian constitution lists 22 official languages. A vast variety of additional languages and dialects are also spoken. Some of these languages have similar texts (for example, Bengali, Bishnupriya Manipuri, Assamese use almost the same text). Many Indian languages have descended from a small number of common languages such as Brahmi.

4.1 Common Ancestry of Languages

The fragmented ancestry of Indian languages may appear to have influenced the alphabetic observers' evolution. Despite sharing the same alphabet, these languages have vastly different texts. In terms of structure, northern Indian text differs greatly from that employed in southern Indian languages. The shape of the glyphs and their 2D distribution which produces the text, vary greatly even within the region.

4.2 Collaborative and Distributed Research Activities

Effective inputs are required from geographically distributed experts with substantial knowledge to successfully construct OCRs in several Indian languages. This needs a development strategy focused on the distributed development of seamless integrated modules. Creating procedures that cater to a wide range of requirements necessitates a high level of adaptability. An ancient, stable, and consistent framework between the modules provided by each participant is required to provide the desired outputs.

4.3 Encoding and Representation

The majority of Indian literature does not comply with foreign standards such as UNICODE. As a result, a plethora of practical issues have arisen that must be addressed. In the late 1960s and early 1970s, vernacular translations also decreased in size to accommodate typewriters and computer programmes. As a result, an apparently random sign with no common denominator exists in numerous languages. While these discussions promote Unicode, they are supplemented by other well-known initiatives and presentations.

4.4 Representation and Recognition

In Indian languages, the shape of the letters varies depending on the situation. Matras are formed when vowels and consonants are mixed with other consonants. When two or more consonants are combined in Indian languages, unifying symbols are formed. In Latin texts, letters are very essential. In a document composed in that language, a character is a separate bit of data. Indian languages rely heavily on their members. Pattern separation at the character level is almost impossible due to separation issues. There are many different words in the basic Indian language (thousands). The internal technology allows to display names, akshara, Unicode, and symbols all at the same time.

4.5 Performance Evaluation

To evaluate OCR performance and comparable systems, a standard data set and virtual interface are necessary. A large and diversified set of benchmark data is required to make performance measurement statistically significant. The same structural representation that can handle a variety of document image categories is necessary to be representative and accessible. To make data more extensively available to users and applications, flexible accessibility solutions are required. Both of these goals can be achieved with XML representation.

4.6 Developing Multilingual Applications

If all connected parts share the same communication architecture, it is easier to design. As a result, system upgrades are significantly easier. The multilingual application demands the integration of numerous modules that process data using a variety of texts and processes. With a standard interaction, the study's findings will be easy to transform into industry-standard commercial assets. The following issues have been noted in numerous studies, and they may spark the researchers' interest in this topic more than other issues. It can be concluded that a small number of Tamil characters were recognised. Testing samples from a minimum number of different handwritten documents have been taken into consideration. It is challenging to distinguish the strange script and similarly shaped characters. There is no mention of font variation or sliding letters and low recognition accuracy rate. Old handwritten character sets that are present in stone sculptures, historical records, mother documents, and palm leaf are not currently dealt with.

5. Conclusion

Visual character recognition is a relatively young study area in which reaching 100% accuracy remains a problem. India is currently considered a big market for a variety of products. English, on the other hand, is not commonly spoken in many places of the world. As a result, certain people at the grassroots level, or locals, should be communicated using local or regional languages. Hence, vernacular literature is in demand today. Although OCR is a minor aspect of digital photography, it has a significant impact in the business world. It has applications in financial organizations, libraries, and the conversion of existing books to computer formats. This paper has focused on the Tamil text, previous attempts to recognize Tamil text, Tamil letter recognition methods, and obstacles related to recognition.

References

[1] K.H. Aparna, Vidhya Subramanian, M. Kasirajan, G. Vijay Prakash, V.S. Chakravarthy.(2004).Online Handwriting Recognition for Tamil.Proceedings of the 9th Int’l Workshop on Frontiers in Handwriting Recognition (IWFHR-9 2004).

[2] Enric Sesa-Nogueras n, MarcosFaundez-Zanuy, “Biometric recognition using online uppercase handwritten text”, www.elsevier.com/locate/pr.

[3] Gowri.N and R.Bhaskaran.(2011).Distortion Analysis Of Tamil Language Characters Recognition,IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 1, January 2011.

[4] R.IndraGandhi and Dr.K.Iyakutti.(2009).An Attempt to Recognize Handwritten TamilCharacter Using Kohonen SOM.Int. J. of Advanced Networking and Applications Volume: 01 Issue: 03 Pages: 188-192 ,2009.

[5] R. Jagadeesh Kannan and R. Prabhakar.(2008).Off-Line Cursive Handwritten Tamil Character Recognition",WSEAS TRANSACTIONS on SIGNAL PROCESSING, ISSN: 1790-5052 351 Issue 6, Volume 4, June 2008.

[6] Marcus Liwicki and HorstBunke "Combining diverse on-line and off-line systems for handwritten text line recognition ", www.elsevier.com/locate/pr.

[7] Kanthavel, D., Sangeetha, S. K. B., & Keerthana, K. P. (2021). An empirical study of vehicle to infrastructure communications-An intense learning of smart infrastructure for safety and mobility. International Journal of Intelligent Networks, 2, 77-82.

[8] S.Rajkumar and Dr.V.Subbiah Barathi.(2011).Century Identification and Recognition of Ancient Tamil Character Recognition" International Journal of Computer Applications (0975 – 8887)Volume 26– No.4, July 2011.

[9] Sangeetha, S. K. B., Dhaya, R., & Kanthavel, R. (2019). Improving performance of cooperative communication in heterogeneous manet environments. Cluster Computing, 22(5), 12389-12395.

[10] Rituraj Kunwar, A. G. Ramakrishnan.(2011).Online handwriting recognition of Tamil script using Fractal geometry, 2011 International Conference on Document Analysis and Recognition.

[11] Shashikiran K, Kolli Sai Prasad, Rituraj Kunwar, A. G. Ramakrishnan,Comparison of HMM and DTW for TamilHandwritten Character Recognition, Technology Development forIndian Languages (TDIL), DIT, Govt. of India

[12] C.Suresh Kumar and Dr.T.Ravichandran.(2010).Handwritten Tamil Character Recognition Using RCSAlgorithm, International Journal of Computer Applications (0975 – 8887) Volume 8– No.8, October 2010.

[13] Sivasubramani K, Loganathan R, Srinivasan CJ, Ajay V, Soman KP.(2007).Multiclass Hierarchical SVM for Recognition of Printed Tamil Characters,AND 2007.

[14] S. Peyarajan and R. Indra Gandhi,(2011).On-Line Tamil HandWritten Character Recognition Using Kohonen Neural Network, Vol 02, Issue 02, July, 2011 Research Journal of Computer Systems Engineering- An International Journal.

[15] Shareef, Shafana & Ragel, Roshan & Nanda Kumara, Titus. (2021). An effective feature set for enhancing printed Tamil character recognition. Journal of the National Science Foundation of Sri Lanka. 49. 195. 10.4038/jnsfsr.v49i2.9466.

[16] R.M. Suresh,(2008).Printed and Handwritten Tamil CharactersRecognition Using Fuzzy Technique, Proceedings of the International MultiConference of Engineers and Computer Scientists 2008 Vol I IMECS 2008, 19-21 March, 2008, Hong Kong.

[17] S. Thilagamani and N. Shanthi.(2011).A Survey on Image Segmentation Through Clustering, International Journal of Research and Reviews in Information Sciences Vol. 1, No. 1, March 2011.

[18] Dr.J.Venkatesh† and C. Sureshkumar,92009).Tamil Handwritten Character Recognition Using Kohonon's SelfOrganizing Map, IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.12, December 2009.

[19] Vladimir Crnojevic´, Vojin Senk and Zeljen Trpovski, 92004).Advanced Impulse Detection based on Pixel-wise MAD, IEEE Signal Processing Letters, Vol. 11, No. 7, July 2004.

[20] Zhou Wang and David Zhan, (1999).Progressive Switching Median Filter for the Removal of Impulse Noise from Highly Corrupted Images,IEEE Transactions on Circuits And Systems—II: Analog And Digital Signal Processing, Vol. 46, No. 1, January 1999.