ANALYZING DIFFERENT ALGORITHMS AND TECHNIQUES TO FIND OPTICAL CHARACTER RECOGNITION FOR TAMIL SCRIPTS

Rajkumar N\textsuperscript{1}, A. B. Rajendra\textsuperscript{2}, Janhavi V\textsuperscript{3}

\textsuperscript{1}Research Mentor, Accendere Knowledge Management Services Pvt Ltd, New Delhi India
\textsuperscript{2}Department of Information Science and Engineering, Vidyavardhaka College of Engineering, Mysuru, Karnataka, India
\textsuperscript{3}Department of Computer Science and Engineering, Vidyavardhaka College of Engineering, Mysuru, Karnataka, India

\textsuperscript{1}rajkumar.mnm@gmail.com, \textsuperscript{2}abrajendra@vvce.ac.in

Corresponding Author: Rajkumar N

https://doi.org/10.26782/jmcms.2020.02.00029

Abstract

Tamil is one of the world's ancient languages. This paper focuses mainly in particular on OCR for the digitalization and conservation of texts and inscriptions in the Tamil language. A system that does not include obtaining either Standard size and shape or the color difference between background and foreground to recognize Palm Leaf Manuscript and stone inscriptions and obtaining information. A variety of algorithms have been analyzed for OCR texts for Tamils, and ancient letter conversion still has a big challenge to convert ancient Tamils into today's digital text format for Tamils.

Keywords: Tamil, OCR, Manuscript, Script, Optical Character Recognition, Tamil Language, Tamil Script

I. Introduction

Tamil is a southern Language spoken primarily in India in Tamilnadu. Among the Dravidian languages, Tamil has the longest unbroken mystical history. The Brahmi script was hereditary in Tamil. Tamilnadu has been preserved for centuries as the Palm Leaf Manuscript. Preserving the palm leaf manuscripts is a difficult and very careful method. Damage to the handwriting of palm leaves can result in data loss. Therefore, when moving to digital text, we need to retain the palm leaf book. Most of them use laptops and smart devices. This approach allows user-friendly conversion and conservation of information [XXV]. A system that scans and
stores the palm-leaf manuscript in the form of an image and converts it into a digital text format [XXVII].

Types of Ancient Language

Ancient languages such as Brahmi, Vatteluttu, Pallava, Grantha, etc. are available in different types. [XXVII].

About the Tamil language

Tamil has 18 consonants and 12 vowels. Vowels, which in the Tamil region are short (Kuril) and long (Nedil) form are otherwise known as UyirEzhuthu. For three classes consonants were categorized into six types as Vallinam, Idaiyinam, and Mellinam. The total of 247 (12 + 18 + 216+ 1) characters and 1 special character (Aayuthazhuthu)[XXV] were combined.Fig. 1 suggests a moderate Brahmi script development to the present Tamil language script. Fig. 2 is a handwritten manuscript with a palm leaf.

Fig. 1: History of Tamil script from Brahmi script to Current Script (Source: Wikipedia)

Fig. 2: Thirukkural Palm Leaf Manuscripts (Source: Wikipedia)

OCR Functional Block Diagram

The OCR frame diagram consists of multiple states as shown in Fig. 3. The test steps are preprocessing, segmentation, extraction of features, regression and classification, post-processing and verification of output [XXV].
Fig. 3: OCR Functional Block Diagram

The research community was interested in the extract of text from document images via a tenner, but very little research was done in the digitalization of inscription images of historical monuments. Fig. 4 demonstrates the Tamil Script test method for OCR.

Fig. 4: Example of OCR for Tamil Script

II. Literature Review

Agnihotri and Dimitrova's red color component [XIII] gives a strong contrast between texts and backgrounds. In the high-contrast video framework "uniform color" blocks are chosen to remove text regions correctly [XXVIII]. The first of the 64 clustered color channels was introduced by Kim et al. for text identification, with cluster colors centered on Euclidean RGB distances[XII]. This approach focused on variation in text and non-text regions were based on variability in Babu et al.[V]. At the text edges the variation is high and vice versa. The variation solution to the removal of the textual matter was inadequate in keeping with the blurry edges of the document and the minimal distinction between text and non-text regions [V]. Text in the inscription images is not uniformly colored and the text and backdrop have poor contrast.
contrast, rendering the use of the image inappropriate [XXVIII]. Simple edge-based approaches are often seen as successful for identifying areas that are especially clustered and long-term. This procedure is efficient if the context is not complex, the inscription images have difficult backgrounds, and in the case of inscriptions, such methods cannot be implemented [IV].

In comparison to threshold volume Seeger et al. [XIV] estimate the strength of the non-text region (background) and execute binarization. In order to improve text images Laplacian of Gaussian Sobolev filters, using various agents for various images is used [XX]. Besides, to describe the destroyed historical documents, curvelet transformations are proposed [II]. For Palm Leaf Manuscripts Cheralu et al. [XXI] used an adaptive binarization technique. Wiener Noise Removal Filter and the adaptive Contrast binarisation for context text segmentation implemented a wavelet-based enhancement/smearing method to eliminate strokes from the corpus of manuscript documents [III]. In [I] the authors proposed a hybrid way of clearing background noise from the old documents using local and global thresholding technologies. The results show that the improvement is made, but OCR can't even read it. The methods listed above were focused on variables, or methods of edge detection, on binarization text extraction. Both strategies are based on the pixel threshold value based on the difference from the foreground to the side.

Garain et al. [XXVI] clarify how the Fast ICA algorithm raises the image to three different layers that suit the text input. The method is an advanced operation, but can't improve weak or spatially-related inscription images. Its convergence has recently slowed down or even failed with the presence of saddle points, especially for short block sizes [XVII]. Fast ICA also has a hard time isolating outlets for insecure or locked sources [XVI].

In the case of unsound and complicated archeological images, the contrast between front and background is not clear. Natural gradient dependent separate variable analytical algorithms, as defined in [XXIII], [XXII], are more effective to minimize the dependency of coupled signals. This algorithm provides better outcomes than other algorithms.

In the course of the coordinating transformation of the pattern and the extract from the warped or other documents, which is then recognized and converted into an editable form, Hemantha Kumar [IX] presented a new methodology for dewraping the text from camera-catches images in the warping document. Chethan [XX] proposed a methodology for contrasts with different edge-based approaches. The Gaussian Kernel is used to identify text, and for placement of text fragments, it imposes dilation and logical AND operations. Jian Yuan [XXI] initiated the production for the extraction and identification of text in natural scene images captured on a smartphone and digital cameras, which were subsequently traduced into Chinese, of an immersive software application called ‘Textract’. In order to perform text extraction, Sachin Grover [XXIV] suggested the simple edge-based method to detect textual areas from the document and to isolate it from the graphics component. Chowdhury [XVIII] has brought forward a new technique that incorporates the
methods of color segmentation and observations of trends dependent on spatial distribution. Rama Mohan Babu [VII] has suggested an algorithm of edge detection utilizing basic mathematical morphology operators that is not sound-sensitive, skew and text-driven.

For the classification by neural networks, Wunsch and Laine [XV] used wavelets derived from the contour of handwritten characters. Lee et al. [XIX] defined using multilayer cluster neural network wavelet features derived from handwritten numerals. The multiwave analysis of the handwritten numbers contour using neural networks was created by Chen et al.[VIII]. Nevertheless, all these plays include very few sections as compared to the current major problem of classification. All the 20 basic characters of the Tamil Vettezhutthu were considered in this study. In [VI] used zero crossings of wave packets in a similar study to identify 20 groups via feed-forward backpropagation network.

Although a large proportion of the inscriptions are lined up in lavish prose, inscriptions provide insights into world dynastic history that otherwise neglect current historical records where knowledge from other outlets, such as established temples and ruins, can be verified. For protection and transparency, digital archiving of these photographs is important. The main technical obstacles for such images are flouted document edges and minimal separation between text and non-text. As in the case with the above inscription pictures[XXIII], the NGFICA algorithm deals with Gaussian, sub-gaussian and super-gaussian source signals.

III. Discussion

In relation to the cursive handwritten and palm leaf textbook and documents, the vehicle number plate in Tamil script we have identified four different algorithms for classification of OCR process. OCR has become increasingly popular in recent years in the use of Support Vector Machine (SVM) classification systems. For different pattern recognition systems, SVMs have obtained excellent results. Equal or even equivalent to expectations such as Bayesian classifiers or multilayer experiences have been shown. SVMs are classifications focused on Vapnik's theory of systemic risk reduction. The NGFICA algorithm is a good method in each stage of improved images to isolate the pitch of the signal. In this way, the OCR system has improved its accuracy by 65.3% (from 10.1% to 75.4%) and by 54.3% (from 32.4% to 86.7%), respectively. Classification Text / Non-Text is based on the threshold value of the setpoints. In the case of cursive handwritten text, binarization and correction in this algorithm are limited. Ancient handwritten character in Tamil, which is based on the basic characteristics of handwritten scripts, namely 1) circle, 2) lines AND 3) combination of circle and line. The system produces a chromosome of a 66-bit string which is a written character. The machine also uses the chromosome of the 66-bit string to classify any handwritten word. There are still many difficult problems in the handwritten recognition system for Tamil, which require more complex algorithms.

SVMs can enforce robust judgment boundaries in high-dimensional function areas as contrasted with other algorithms and classification techniques. The explicit regularization of the form of the classifier inhibits premature adaptation and creates
generalizations. Some other real-world properties are often seen as explanations for SVMs' popularity. The outcome of the training is secure, fast training algorithms are accessible and little information is required.

IV. Conclusion

Tamil is an ancient language. Many well-known authors have published millions of books. It is very difficult to maintain and receive the contents of the books. OCR creates an ecosystem without paper which promotes the exchange of knowledge. In this paper in Tamil, we have talked about different OCR algorithms. Relative to other recognition algorithms, SVM was found to be stronger. To distinguish individual characters, the recognition system requires a more powerful segmentation algorithm and a more reliable detection algorithm to classify ambiguous written types. For all written characters have unlimited forms, patterns, and styles. It is therefore difficult to define a standard structure to explain uncertain written characters in a general algorithm. Although the researcher interested in accessing the content is created to find information on rich Tamil content, people of various categories can access it easily and comfortably. If OCR becomes usable, it becomes easier to process and manage the information.

References


