VOL. 13, NO. 3, FEBRUARY 2018 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

# TAMIL CHARACTER RECOGNITION USING ANDROID MOBILE PHONE

K. Jayasakthi Velmurugan[1] and M. A. Dorairangaswamy[2]
[1]Department of Computer Science Engineering, Sathyabama University, Chennai, India
[2]St.Peter's University, Chennai, India
E-Mail: jaisakthi21@gmail.com

**ABSTRACT**

Tamil Text detection in natural scene picture is an important requirement for many content-based image analysis tasks. I propose an accurate and robust method for detecting Tamil texts in natural scene pictures. A fast and effective pruning algorithm is designed to extract Maximally Stable Extreme Regions (MSERs) as character candidates using the strategy of minimizing regularized variations. Character candidates are merged into text candidates by the single link clustering algorithm, where distance weights and clustering threshold are learned automatically by a novel self-training distance metric learning algorithm. The posterior probabilities of text candidates corresponding to non-text are estimated with a character classifier; text candidates with high non-text probabilities are eliminated and texts are identified with a text classifier. In this application the documents will be scanned as images and once the image is scanned the data from the image is extracted automatically and will be shown in the application as text. Then the text message is given to the translator tool which will convert the Tamil text into English Text message.

**Keywords:** pruning algorithm, maximally stable extreme regions, self-training.

## INTRODUCTION

Image processing is the process of enhancing the quality of the images by certain algorithm. Image processing is the form of signal processing. Here the input is an image and the output of the image will be a set of characters related to the image. In image processing technique, the image will be treated as two dimensional signals. Image processing is a method to convert an image into digital form and perform some work on it, in order to get an enhanced image or to extract some useful information from it.

Tamil Text detection is the mechanical or electronic conversion of images of typewritten or printed text into machine-encoded text. It is widely used as a form of data entry from printed paper data records, whether passport documents, invoices, bank statement, receipts, business card, mail, or other documents. It is a common method of digitizing printed texts so that it can be electronically edited, searched, stored more compactly, displayed on-line, and used in machine processes such as machine translation, text-to-speech, key data and text mining.

At the present time, keyboarding remains the most common way of inputting data into computers. This is probably the most time consuming and labor intensive operation. OCR is the machine replication of human reading and has been the subject of intensive research for more than three decades. OCR can be described as Mechanical or electronic conversion of scanned images where images can be handwritten, Typewritten or printed text. It is a method of digitizing printed texts so that they can be electronically searched and used in machine processes. It converts the images into machine-encoded text that can be used in machine translation, text-to-speech and text mining. This paper presents a simple, efficient, and less costly approach to construct OCR for reading any document that has fix font size and style. To achieve efficiency and less computational cost, OCR in this paper uses database to recognize Tamil characters which makes this OCR very simple to manage.

Optical character recognition, usually abbreviated to OCR, is the mechanical or electronic translation of scanned images of handwritten, typewritten, or printed text into machine-encoded text.

In this application the user will capture an image threw his smart phone as scanned image where the text will be extract, scanned images as involved into de-skewing. Image De-skew is the process of removing skew from images (especially bitmaps created using a scanner). Skew is an artifact that can occur in scanned images because of the camera being misaligned, imperfections in the scanning or surface, or simply because the paper was not placed completely flat when scanned.

The extraction steps will be followed by binarization of captured image, Character segmentation, and orientation. Crossing these steps the text will be extracted from the image.

Once the text is extracted from the image it will be uploaded in the web application for the purpose of Translation.

## LITERATURE REVIEW

Jaehwa Park *et al.* [3], this paper describes a character recognition methodology (henceforth referred to as Hierarchical OCR) that achieves high speed and accuracy by using a multiresolution and hierarchical feature space. Features at different resolutions, from course to fine-grained, are implemented by means of a recursive classification scheme. Typically, recognizers have to balance the use of features at many resolutions (which yields a high accuracy), with the burden on computational resources in terms of storage space and processing time. We present in this paper, a method that adaptively determines the degree of resolution necessary in order to classify an input pattern. This leads to optimal use of computational resources. The Hierarchical OCR

dynamically adapts to factors such as the quality of the input pattern, its intrinsic similarities and differences from patterns of other classes it is being compared against, and the processing time available. Furthermore, the finer resolution is accorded to only certain zones of the input pattern which are deemed important given the classes that are being discriminated. Experimental results support the methodology presented. When tested on standard NIST data sets, the Hierarchical OCR proves to be 300 times faster than a traditional K-nearest-neighbor classification method, and 10 times faster than a neural network method. The comparison uses the same feature set for all methods. Recognition rate of about 96 percent is achieved by the Hierarchical OCR. This is at par with the other two traditional methods.

Bhushan Sonawane *et al*. [10], This paper relies on Microsoft Office Document Imaging (MODI) and puts forward to give easy access to documenting image for blind people. We propose the system with simple architecture which provides better reliability than ordinary traditional reading methods for blinds. This system replaces existing system such as braille lipi which reduces the efforts for guessing the characters in the paper documents. The Proposed system adapts flexible functionality for recognizing the characters using Microsoft Office Document

Imaging. It is essential to tune the system with sharp, bright and proper format image. The System can adjust for various image structure & results in high recognition accuracy on poor quality print.

Youssef Es Saady *et al.* [5], in this paper a system of Amazigh handwriting recognition based on horizontal and vertical centerline of the character. After the image preprocessing, the text is segmented into lines and then into characters. The positions of the vertical and horizontal centerline of the character are used to obtain a set of independent and dependent features to those lines. These features are related to the densities of pixels and are extracted on binary images of characters using the sliding window technique. Finally, a multilayer perceptron is used for character classification. The system was tested on two bases of the Amazigh characters: on a printed database of Amazigh characters and on another one for handwritten characters created locally. The correct average recognition rate obtained using 10-crossvalidation was 99.28 % for the 19437 Amazigh printed characters and 96.32% for the 20150 Amazigh handwritten characters.

Vladimir Kluzner *et al.* [11], Optical character recognition (OCR) technology is widely used to convert scanned documents to text. However, historical books still remain a challenge for state-of-the-art OCR engines. This work proposes a new approach to the OCR of large bodies of text by creating an adaptive mechanism that adjusts itself to each text being processed. This approach provides significant improvements to the OCR results achieved. Our approach uses a modified hierarchical optical flow with a second-order regularization term to compare each new character with the set of super-symbols (character templates) by using its distance maps. The classification process is based on a hybrid approach combining measures of geometrical differences (spatial domain) and distortion gradients (feature domain).

Shalin A. Chopra *et al.* [4], at the present time, keyboarding remains the most common way of inputting data into computers. This is probably the most time consuming and labour intensive operation. OCR is the machine replication of human reading and has been the subject of intensive research for more than three decades. OCR can be described as Mechanical or electronic conversion of scanned images where images can be handwritten, typewritten or printed text. It is a method of digitizing printed texts so that they can be electronically searched and used in machine processes. It converts the images into machine-encoded text that can be used in machine translation, text-to-speech and text mining. This paper presents a simple, efficient, and less costly approach to construct OCR for reading any document that has fix font size and style or handwritten style. To achieve efficiency and less computational cost, OCR in this paper uses database to recognize Tamil characters which makes this OCR very simple to manage.

S. K. Thilagavathy *et al.* [7], Character recognition is an ever ending research application in the real world. Each character recognition should be accurate. So that it leads to understand the exact meaning and concept. Analyzing the distorted character is quite complicated work. In some unique languages like Telugu and Malayalam the distorted character may resemble like some other character. In this paper analysis of each character is done even though it is inconsistent in shape and irrespectively distorted.

Er. Kavneet Kaur *et al.* [14], Automatic Number Plate Recognition (ANPR) is a special form of Optical Character Recognition (OCR). ANPR is an image processing technology which identifies the vehicle from its number plate automatically by digital pictures. In this paper we have presented an algorithm for vehicle number identification based on Optical Character Recognition (OCR). OCR is used to recognize an optically processed printed character number plate which is based on template matching. This algorithm is tested on different ambient illumination vehicle images. OCR is the last stage in vehicle number plate recognition. In recognition stage the characters on the number plate are converted into texts. The characters are then recognized using the template matching algorithm.

Er. Neetu Bhatia [1], this paper presents detailed review in the field of Optical Character Recognition. Various techniques are determined that have been proposed to realize the center of character recognition in an optical character recognition system. Even though, sufficient studies and papers are describes the techniques for converting textual content from a paper document into machine readable form. Optical character recognition is a process where the computer understands automatically the image of handwritten script and transfer into classify character. This material use as a guide and update for readers working in the Character Recognition area. Selection of a relevant feature extraction method is probably the single most important factor in achieving

high character recognition with much better accuracy in character recognition systems without any variation.

Pranob K Charles *et al*. [9], Handwriting recognition has been one of the most interesting and challenging research areas in field of image processing and pattern recognition in the recent years. This paper describes the techniques for converting textual content from a paper document into machine readable form. The computer actually recognizes the characters in the document through a revolutionizing technique called

Optical Character Recognition. Several techniques like OCR using correlation method and OCR using neural networks are reviewed in this paper.

Alessandro Bissacco *et al.* [13], we describe Photo OCR, a system for text extraction from images. Our particular focus is reliable text extraction from smart phone imagery, with the goal of text recognition as a user input modality similar to speech recognition. Commercially available OCR performs poorly on this task. Recent progress in machine learning has substantially improved isolated character classification; we build on this progress by demonstrating a complete OCR system using these techniques. We also incorporate modern datacenter-scale distributed language modeling. Our approach is capable of recognizing text in a variety of challenging imaging conditions where traditional OCR systems fail, notably in the presence of substantial blur, low resolution, low contrast, high image noise and other distortions. It also operates with low latency; mean processing time is 600ms per image. We evaluate our system on public benchmark datasets for text extraction and outperform all previously reported results, more than halving the error rate on multiple benchmarks. The system is currently in use in many applications at Google, and is available as a user input modality in Google Translate for Android.

Ondrej Krejcar [15], the paper deal with a development of a mobile application for capturing digital photography and its subsequent processing by OCR (Optical Character Recognition) technologies. The developed solution adds to existing Smart Device a capability of a virtual keyboard to which it is possible to transfer recognized text for further work in SMS or text editor. For example, based on the limitation of mobile devices it is mainly targeting at short text sections (internet references, complex addresses, etc.). The accent is targeted on the simple, fast and intuitive working with a mobile device. Practical realization is verified at several Smart Devices with Windows Mobile OS.

## SMART CHARACTER RECOGNITION

In this system we are providing a smart phone and web based application which helps to retrieve Tamil text from image files. The user can scan the file to get the clear image file threw our application and if the image contains many Tamil text means that text data will be extracted from the image the user can edit and extract it from the web application.

In this application the Tamil text present in the image will be extracted using segmentation, denoising and edge detection algorithms. The extracted Tamil text can be converted to the English language and also to speech which helps visually challenged people.

## MODULES AND ALGORITHM DESCRIPTION

In our proposed model we perform five modules:

- Text Detection
- Language Conversion
- Offline Dictionary
- Text to Speech conversion
- Cloud Storage

### Text detection

In this module image will be captured, the Tamil text will be extracted from the captured image by following three step algorithm they are Segmentation, denoising and edge detection.

### Segmentation

It is the process of partitioning a digital image into multiple segments (sets of pixels, also known as super pixels). The goal of segmentation is to simplify and/or change the representation of an image into something that is more meaningful and easier to analyze. Image segmentation is typically used to locate objects and boundaries (lines, curves, etc.) in images. More precisely, image segmentation is the process of assigning a label to every pixel in an image such that pixels with the same label share certain characteristics.

### Denoising

It is an algorithm in image processing for image denoising. Unlike "local mean" filters, which take the mean value of a group of pixels surrounding a target pixel to smooth the image, non-local means filtering takes a mean of all pixels in the image, weighted by how similar these pixels are to the target pixel. This result in much greater post-filtering clarity and less loss of detail in the image compared with local mean algorithms.

### Edge detection

Edge detection is the name for a set of mathematical methods which aim at identifying points in a digital image at which the image brightness changes sharply or, more formally, has discontinuities. The points at which image brightness changes sharply are typically organized into a set of curved line segments termed *edges*. The same problem of finding discontinuities in 1D signal is known as step detection and the problem of finding signal discontinuities over time is known as change detection. Edge detection is a fundamental tool in image processing, machine vision and computer vision, particularly in the areas of feature detection and feature extraction.

### Language conversion

In this module the extracted text will be converted to desired language using google translator API, when the text is extracted to edit text, the user has to choose the language in the list, when the language is chosen the text will be converted to desired language. This

module depends on data connection; if data connection is not available the conversion will be failed.

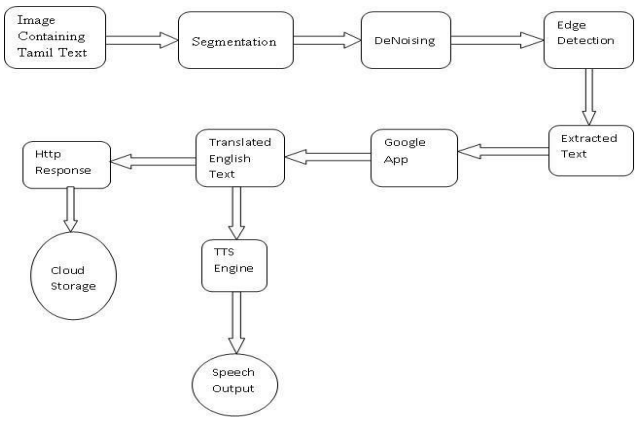


**Figure-1.** Architecture diagram.

**Offline dictionary**

In this module local dictionary is implemented using SQLite, to give a support to the words in extracted. The implemented words will be almost equal to oxford hand book dictionary.

**Text to speech conversion**

In this module the text will be converted to speech using TTS engine. The extracted Tamil text will be converted to English speech; this will be helpful for blind people by giving them a practice to use this app.

A Text To Speech instance can only be used to synthesize text once it has completed its initialization. Implement the *Text To Speech. On Init Listener* to be notified of the completion of the initialization.

**Cloud storage**

Cloud storage is introduced in this technique; user can store both image and extracted Tamil text and converted English Text in cloud using their username and password. If the user login in another mobile using their username and password, they can sync in cloud and download their image and text anywhere anytime. This cloud is developed by PHP and taken MySQL as backed data storage.

**Advantages**

- It is easy to extract Tamil text from image files.
- The user can export in the text in desired file type.
- The user can make the hard copy as soft copy within few minutes.

## RESULTS AND DESCRIPTION

Text detection is the mechanical or electronic conversion of images of typewritten or printed text into machine-encoded text. It is widely used as a form of data entry from printed paper data records, whether passport documents, invoices, bank statement, receipts, business card, mail, or other documents. It is a common method of digitizing printed texts so that it can be electronically edited, searched, stored more compactly , displayed on-line, and used in machine processes such as machine translation, text-to-speech, key data and text mining.

| Method | Recall | Precision | *f* |
|---|---|---|---|
| Our method | 85.22 | 95.32 | 90.35 |
| Text detection on | 84.21 | 93.49 | 88.61 |
| ICDAR 2011 | | | |
| TCC_textDetector | 88.64 | 81.46 | 84.90 |
| Anthimpulos | 81.88 | 87.35 | 84.53 |

## CONCLUSION AND FUTURE ENHANCEMENT

An accurate and robust method for detecting Tamil text in natural scene images. A fast and effective pruning algorithm is designed to extract Maximally Stable Extreme Regions (MSERs) as character candidates using the strategy of minimizing regularized variations.

The user can upload the extracted text to the web application. Threw the web application the user can modify the content changes if any. And the user can extract the text with his desired format for example doc, pdf, etc. Here user can edit the extracted content so he can easily add his own content or can modify the existing content changes if any. This edit and export options are given in web application so it is user friendly.

In future, this work can be extended by adding more constraints.

## REFERENCES

[1] Er. Neetu Bhatia. 2014. Optical Character Recognition Techniques: A Review. 4(5).

[2] H. Li, D. Doermann, and O. Kia. 2000. Automatic text detection and tracking in digital video. IEEE Trans. Image Process. 9(1): 147-156.

[3] Jaehwa Park, Venu Govindaraju, Senior Member, IEEE and Sargur N. Srihari, Fellow, IEEE. 2000. OCR in a Hierarchical Feature Space. IEEE Trans. Pattern Anal. Mach. Intell. 22(4): 400-407.

[4] Shalin A. Chopra, Amit A. Ghadge, Onkar A. Padwal, Karan S. Punjabi, Prof. Gandhali S. Gurjar. 2014. Optical Character Recognition. International Journal of Advanced Research in Computer and Communication Engineering. 3(1): 4956-4958.

[5] Youssef Es Saady, Ali Rachidi, Mostafa El Yassa, Driss Mammass. 2011. Amazigh Handwritten Character Recognition based on Horizontal and Vertical Centerline of Character. International Journal of Advanced Science and Technology. 33: 33-50.

[6] X. Chen and A. Yuille. 2004. Detecting and reading text in natural scenes. in Proc. IEEE Conf. CVPR. Vol. 2. Washington, DC, USA. pp. 366-373.

[7] S.K. Thilagavathy and Dr.R. Indra Gandhi. 2013. Recognition of distorted character using edge detection algorithm. International Journal of Innovative Research in Computer and Communication Engineering. 1(4): 1056-1061.

[8] K. Kim, K. Jung, and J. Kim. Texture-base approach for text detection in images using support vector machines and continuously adaptive mean shift algorithm. IEEE Trans. Pattern.

[9] Sukhpreet Singh. 2013. Optical Character Recognition Techniques: A Survey. Journal of Emerging Trends in Computing and Information Sciences. 4(6): 545-550.

[10] Pranob K Charles V. Harish M. Swathi CH. Deepthi. 2012. A Review on the Various Techniques used for Optical Character Recognition. International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com. 2(1): 659-662.

[11] Bhushan Sonawane, Kiran Patil, Nikhil Pathak, Ram Gamane. 2013. Third Eye: An Image Explorer. International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal. 3(4): 438-441.

[12] Vladimir Kluzner, Asaf Tzadok, Dan Chevion, Eugene Walach. Hybrid Approach to Adaptive OCR for Historical Books. 2011 International Conference on Document Analysis and Recognition. p. 900

[13] J. Zhang and R. Kasturi. 2008. Extraction of text objects in video documents: Recent progress, in Proc. IAPR Workshop DAS, Nara, Japan. pp. 1-13.

[14] Alessandro Bissacco, Mark Cummins, Yuval Netzer, Hartmut Neven. Photo OCR:Reading Tex t in Uncontrolled Conditions. p. 785.

[15] Er. Kavneet Kaur, Vijay Kumar Banga. Number plate recognition using OCR technique. IJRET: International Journal of Research in Engineering and Technology eISSN: 2319-1163 | pISSN: 2321-7308, pp. 286-290.

[16] Ondrej Krejcar. Smart Implementation of Text Recognition (OCR) for Smart Mobile Devices. INTELLI 2012: The First International Conference on Intelligent Systems and Applications. p. 19.

[17] J. Zhang and R. Kasturi. 2008. Extraction of text objects in video documents: Recent progress. in Proc. IAPR Workshop DAS, Nara, Japan. pp. 1-13.

[18] Yin and C.-L. Liu. 2009. Handwritten Chinese text line segmentation by clustering with distance metric learning. Pattern Reco gnit. 42(12): 3146-3157.

[19] X. Yin, X.-C. Yin, H.-W. Hao and K. Iqbal. 2012. Effective text localization in natural scene images with MSER, geometry-based grouping and ada boost. in Proc. Int. Conf. Pattern Recognit., Tsukuba, Japan. pp. 725-728.

[20] A. Jain, M. Murty and P. Flynn. 1999. Data clustering: A review. CM Comput. Surv. 31(3): 264-323.