Mezhinam-OCR

Barathy Kolappan A
BTech IT, VIT University
Vellore(TN), India

barathy.kolappan2017@vitstudent.ac.in

Optical Character Recognition (OCR) is recognizing and digitizing characters without human association, its one area where continuous enhancements in accuracy and newer proposals in detection emerge every day. Tamil is one of the many classical languages in India, it's also one amongst the oldest in the world that is still in common usage. Its letter form and type are considered to be one of the most archaic. The language comprises 12 vowels, 18 consonants and 217 consonantal vowels. It has evolved from origin Brahmi script to modern day usage Tamil Script. TDN OCR digitizes any printed material, which has Tamil text in it. In the application of OCR, three datasets have been synthesized, TDNMax (10386), TDN (3268), TDNs (106). Convolutional neural networks compare templates as a function of the displacement of one relative to the other and evaluate its similarity. It differentiates in the grounds of number of templates it contains, this way letters are fragmented and acted upon. Several optimization techniques are applied to increase the reliability of the machine. The accuracy of the proposed method were 94.6, 87.2, and 63.8 for the datasets of TDNMax, TDN and TDNs respectively. Also the proposed technique outperforms most existing benchmarks set in the language.

1 Introduction

With increasing growth of portable documents every day, project-based documents are ceasing to exist. Plenty of good text just go into the dump lumps to a remote village somewhere just to serve the needs of disposable food carriers or tissue projects. Decades of work, sleepless hours curating content, back and forth drafts and is this their destiny?

With the widely advanced technology we have today, what we are capable of materializing is little. This is not one of those don’t cares. Aforesaid Tamil is one of the oldest language in existence, obvious to the fact it contains millions of documents written. Adding to its authenticity is the letter types it comprises, 12 vowels, 18 consonants, 217 consonantal vowels, thousands of rules to form words, hundreds of rules to form sentences. Given its

complexity. For one to decipher it, takes expertise, nativity. What if we could digitize all of the documents we have in little more than a snap each?
Thanks to Optical Character Recognition techniques that has made it possible. Several techniques starting from deep learning networks to as simple as a built-in function, could effectively figure out and digitize the text in all actuality. But the catch is all of the existing techniques come with a downside. One can't recognize if the input text has a shadow, other can't read colored text, other can't read text in three dimensions. Precisely OCR errors are more likely to occur in languages that have large number of characters, one such is Tamil. Whatsoever, it is almost impossible to correct all of these OCR errors manually. Technology is not purported to share its work load and this incapacity proves wrong, it's very existence. This project intends to overcome the defects produced by conventional techniques, common OCR errors and improvise magnitude of reliability.

This project supposes the system to evaluate similarities between the fragmented letter from the processed image and the synthesized template collection through convolutional neural networks that effectively simulates cross correlation by performing template matching to the input image to be processed. When the adequacy of word detection is fulfilled, the word is run through a dictionary of over 300,000 most used words from the language and evaluated for similarity coefficient, the word that most closely matches the requiem is printed.

2 Related Work

A lot of reports on Optical Character Recognition using Convolutional Neural Networks/ Cross Correlation methods/Template matching have been reported. However most of them suppose the text in the document is English. Although some reports are purported for Tamil, they demand organic text without casted shadows and dimensional variation.

In another front various other techniques have been implemented to detect Tamil, few of which have been successful too. However, the aim of this project is to improve reliability on high frequency words overall.

3 Convolutional Neural Networks

Convolutional Neural Networks (CNN) is a type of feed forward neural networks that has applications in image recognition, natural language processing and recommender systems. It consists of an input layer, multiple hidden layers and an output layer. In actuality it performs more cross correlation than convolution. The project simulates the working of CNN with Template Matching.

3.1 Algorithm

1) Each image that is to be processed have to undergo,

    a) Conversion to Grayscale
    b) Conversion to Binary
    c) Dilation in Threshold
    d) Contour mapping and sorting

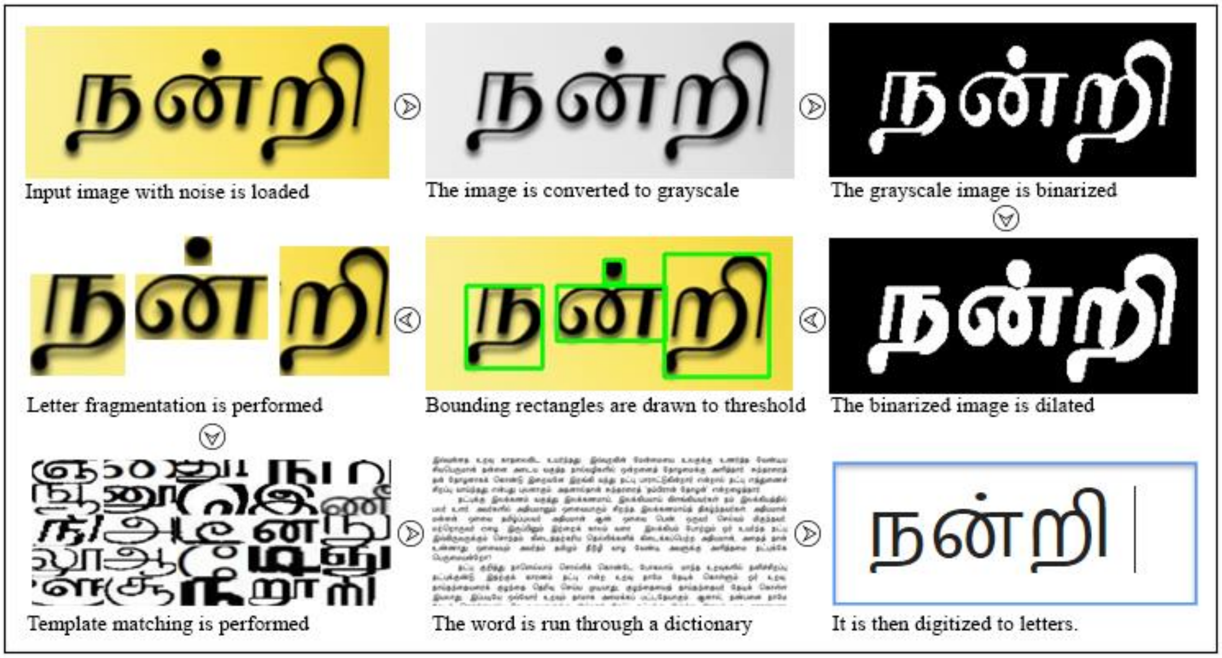
2) When the Contours are Sorted,

    a) Bounding Rectangles are drawn around threshold
    b) Region of Interest, ROI is extracted
        i) similarity_score=compare(template,ROI)
        ii) match_score = max(similarity_score)
        iii) match_letter = templates[match_score]
        iv) final_word+=match_letter

3) With the completion of first phase, final_word is,

    a) Compiled into UNICODE Tamil via Python dictionary
    b) Corrected for errors in spelling via Tamil dictionary

4) Final word detected and identified correct is printed.


Input image with noise is loaded

The image is converted to grayscale

The grayscale image is binarized

Letter fragmentation is performed

Bounding rectangles are drawn to threshold

The binarized image is dilated

Template matching is performed

The word is run through a dictionary

It is then digitized to letters.

Fig 1: Illustrating the process flow that the project intends to.

4 Proposed Method

The proposed method intends to extend the boundaries of computer vision and ConvNets to identifying and digitizing the one of the oldest Indic languages that's still in common usage. It also purports to serve to adopt models implemented for western scripts and examine its viability for Tamil script.   Also since the datasets have been manually synthesized to stand out, extreme measures have been classification of letters to detect and nomenclature has been taken to ensure maximal accuracy in shortest worst case speed possible in the system.

The project proposes alterations and additions to existing methods, leading to an entirely new model for examination, detection, identification and digitization of Tamil text on any surface, with any texture, any color and with manageable noise.

5 Experiments

The proposed has been to test with a plethora of printed and handwritten characters and proved to be adequately reliable on the terms of accuracy and speed. Below is the report from experiment with 500 characters on the synthesized three datasets of TDNMax, TDN and TDNs.

Dataset	**True Positives**		**Accuracy**		**Average Speed**
	Printed	Handwritten	Printed	Handwritten	
TDNMax (10386)	473	344	94.6	68.8	0:00:04.92/letter
TDN (3268)	436	291	87.2	58.2	0:00:02.43s/letter
TDNs (106)	319	117	63.8	23.4	0:00:01.54s/letter

5 Conclusion

The proposed method is a scalable system and still has scope of enhancements. However it has already outperformed most conventional techniques, what remains is to break its own feat. The project out of subject also seeks technological advancements for regional sects, its first step to achieve that being an advanced OCR system for Tamil. It will constantly be under improvements and drafts, only to make it better.

References

1.Tamil character recognition from ancient epigraphical inscription using OCR and NLP
T Manigandan ; V. Vidhya ; V Dhanalakshmi ; B Nirmala
2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)

2.Identification of Tamil ancient characters and information retrieval from temple epigraphy using image zoning
R. Giridharan ; E. K. Vellingiriraj ; P. Balasubramanie
2016 International Conference on Recent Trends in Information Technology (ICRTIT)

3.Feature selection for an automated ancient Tamil script classification system using machine learning techniques
T S Suganya ; S Murugavalli
2017 International Conference on Algorithms, Methodology, Models and Applications in Emerging Technologies (ICAMMAET)

4.Tamil Character Recognition Using Canny Edge Detection Algorithm
P. Selvakumar ; S. Hari Ganesh
2017 World Congress on Computing and Communication Technologies (WCCCT)

5.Optical character recognition using template matching and back propagation algorithm
Ashima Singh ; Swapnil Desai
2016 International Conference on Inventive Computation Technologies (ICICT)