Indian Journal of Science and Technology, Vol 8(35), DOI: 10.17485/ijst/2015/v8i35/86811, December 2015
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Offline Character Recognition of Printed Tamil Text using Template Matching Method of Bamini Tamil Font

D. Pugazhenthi* and S. Arul Vallarasi

Department of Computer Science, Quaid-E-Millath Government College for Women (Autonomous), Anna Salai, Chennai – 600002, Tamil Nadu, India; pugazh006@gmail.com, arul.s.vallarasi@gmail.com

Abstract

Background/Objectives: Character recognition of the English alphabet using template matching method is a simple method to implement. This paper proposes Tamil character recognition of Bamini Tamil font using Template Matching method. **Materials/Methods:** The document image without skew is binarized at the preprocessing step. The preprocessed image is then segmented. Every line of text is segmented using horizontal projection analysis. Every character in a line is segmented using connected component processing. Then each character segmented is correlated with the preloaded templates of the system. The maximum correlation judges the character. In the same way, every segmented input is checked with the preloaded templates. These templates are mapped onto Tamil Unicode for recognition. The text is reconstructed using Unicode fonts and finally produces the Machine editable Unicode text in a text file. **Conclusion/Findings:** The system gives results considerably greater than 20 pixel base height of a character in the document image. **Applications/ Improvements:** The possibilities of using other fonts character recognition is applied in future. Other methods are also considered for further implementation.

Keywords: Connected Component Labeling, OCR, Offline Character Recognition, Segmentation, Template Matching

1. Introduction

Optical Character Recognition (OCR) is a field of research in pattern recognition, artificial intelligence and computer vision. OCR is the mechanical or electronic conversion of typewritten or printed text images into editable text. It is a technology that converts different kinds of inputs, such as scanned documents, images captured by a digital camera into editable and searchable data. In other words, producing documents like a text file from a scanned image of a printed or handwritten page. OCR is a common method of digitizing printed texts so that it can be electronically edited, searched and used in machine processes such as machine translation (translate text or speech from one language to another), text-to-speech and text mining (process of deriving high-quality information from text).

The Optical Character Recognition can be widely categorized into two; online and offline[1–3]. In offline character recognition scanned images of text is converted into machine editable text. Whereas in online character recognition the input of the OCR is received by means of digital pen, digital stylus like devices and the input is converted into machine editable text. There are many methods available for printed text recognition. In templates matching method, individual character templates are added into the system initially. Once the templates creation is over, the test image is fed into the system. The system segments the image into individual characters and the segmented image is correlated with the preloaded templates[4]. The maximum correlated character is declared as the character present in the image. The same process is repeated for every segmented character in the image.

* Author for correspondence

The character recognition of Roman script is a universal one. And complete OCR system is available for Roman script. But for the Indian scripts (Tamil script included) there is no complete OCR system is available. Tamil character set is large when compared to English character. Not much work has been done for Tamil language. English language has only 26 upper case letters and 26 lower case letters totaling 52 characters. Tamil language has one special character (Aayuthaezhutthu), 12 vowels, 18 consonants and 12x18 compound letters, totaling 247 characters and some Sanskrit characters like [,],\,`. The Tamil character set is vast when compared to English character set. So the character recognition of Tamil language is a big job. The compound letters are formed by adding each consonant with each vowel. For example we furnish the first consonant with all vowels.

க் + அ = க க் + ஆ = கா

The pitch of the letters in Tamil varies character by character. The pitch of 'கா' is greater than that of 'க'. Some characters are split into two or three parts. For example the compound letter 'கா' consists of two parts 'க' and 'h'. In Tamil, characters consonants have a dot symbol on top of the characters 'க்'. The Tamil character set has multiple pitch and 2 or 3 parts. The proposed Optical Character Recognition system has two stages 'template creation' and 'character recognition stage'.

The paper is organized as follows. Section 1, the basics of Optical Character Recognition is discussed. In materials and methods, the process of the proposed system is explained. In results and discussion, the input and output of the system is discussed. In conclusion, the positive and negative of the system is talked about.

2. Materials and Methods

2.1 Template Creation Stage

The proposed system has two stages; template creation stage and character recognition stage.

In template creation stage, images of every letters in Bamini Tamil font with normal font (neither bold nor italic) style are gathered. The images are further split into connected component images. That is the image of split character 'ngh' is divided into three say 'n', 'g', 'h'. Then all the divided images are converted into a common dimension say 42x24. Then the resultant images are stored in system drive with one label 1,2,3 and so on. After naming every final image, a template database is created with the labeled images.

2.2 Character Recognition Stage

The steps involved in the character recognition stage are given below.

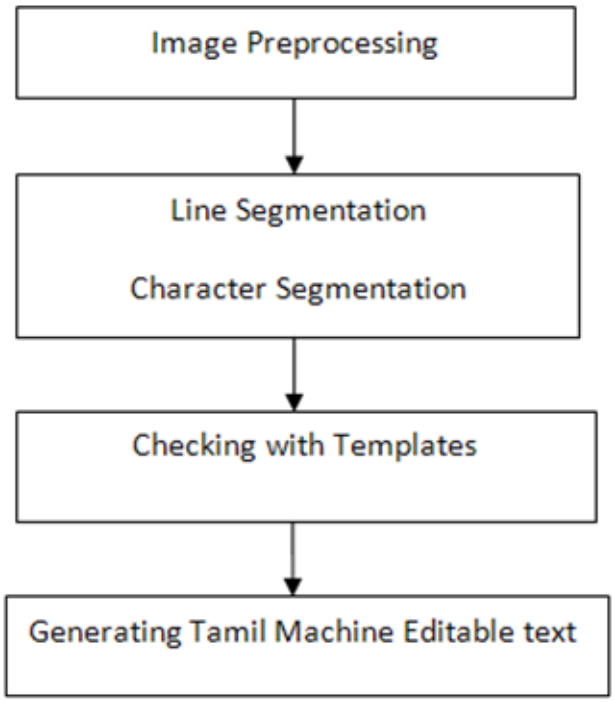

Figure 1. Steps in character recognition.

In the character recognition stage the test document image is sent through the system. Initially the image is preprocessed[5,6]. The color image is converted into grayscale image. Image binarization converts an image of 256 gray levels into a black and white image. Using global thresholding technique the Image is binarized[7,8]. The preprocessed image is in the form of matrix with binary values (0,1). The lines in the preprocessed imageare separated according to the white space between lines of text. The horizontal projection analysis method is used to split each line[9–11]. The image matrix is traversed from top to bottom, wherever the white space count in a line is zero, the line is separated.

After the line segmentation, every character in the line is separated. Each character is segmented using the connected component labeling technique. It uses 8-connectivity to detect the connected image of a character[12]. Every connected component gets a separate label to distinguish one from the other. While using connected component to separate characters, some Tamil compound letters get 2 or 3 labels because of the split parts. That is, the image of compound character 'ngh' is

divided into three say 'n', 'g', 'h'. Likewise the compound letter 'gh' is divided into two say, 'g', 'h'. This is because the character has split parts in other words it has the white space between the three parts. And gets three individual labels.

Template matching is a technique in digital image processing for finding small parts of an image which match a template image[13]. The labeled image from the previous stage is split according to the label number. Every labeled image is correlated with the preloaded templates. The correlation coefficient is calculated using the below formula

$$\gamma = \frac{\sum_m \sum_n (A_{mn} - \overline{A})(B_{mn} - \overline{A})}{\sqrt{\left(\sum_m \sum_n (A_{mn} - \overline{A})^2\right)\left(\sum_m \sum_n (B_{mn} - \overline{B})^2\right)}} \quad (1)$$

Where m is row number and n is column number. Here A_{mn} is the pixel intensity or the gray scale value at a point (m,n) in the template image. B_{mn} is the gray scale value at a point (m,n) in the test image. = mean of (A) and = mean of (B)

The maximum correlated character is declared as the character present in the labeled image. The same process is repeated for every segmented character (labeled image) in the image. Then the text is reconstructed using Unicode font 'Arial Unicode MS'. Tamil characters range from '0B80' to '0BFF' in the Unicode consortium[14]. It finally produces the machine editable unicode text in a text file. The system is worked with Bamini Tamil True type font.

3. Results and Discussion

The system is implemented in Matlab R2013b. The Sample Input Image of 40 pixel height (single character base height) is given below. The corresponding output for the input image is also given below. The sample output image in a text file. The document image is neither bold nor italic. The test image is passed through the system, the test document image is preprocessed, each character is separated and correlated with the preloaded templates, the maximum correlated template judges the character, and the output is given in a text file.

Table 1. Experimental Results

Number	Base Height of a character in pixel	Percentage of Correctly judged characters	Percentage of Wrongly judged characters
1	10	48.14815	51.85185
2	11	43.51852	56.48148
3	13	41.66667	58.33333
4	15	26.85185	73.14815
5	17	56.48148	43.51852
6	19	82.40741	17.59259
7	21	88.88889	11.11111
8	23	93.51852	6.481481
9	25	90.74074	9.259259
10	27	90.74074	9.259259
11	29	92.59259	7.407407
12	31	91.66667	8.333333
13	33	92.59259	7.407407
14	35	92.59259	7.407407
15	37	91.66667	8.333333
16	39	90.74074	9.259259
17	41	93.51852	6.481481
18	43	96.2963	3.703704

Offline cursive handwritten character recognition and skew corrections[15–17] are not added in the system. The

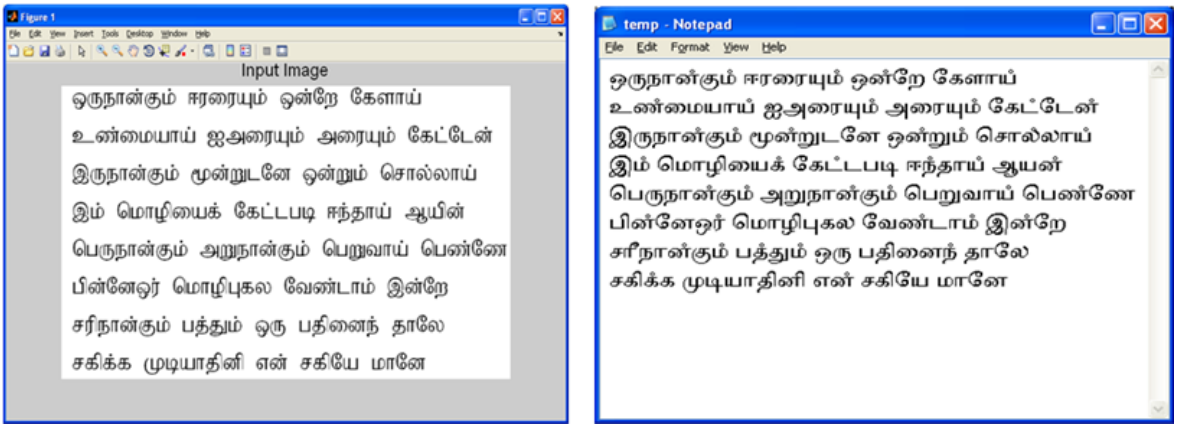

Figure 2. tInput document image and its corresponding output in notepad.

Tamil character document image of 108 characters long is recognized in the system and the result is listed in below Table 1. For every character base height in pixel, correctly judged and wrongly judged character count is listed. For small font images the system gives more wrong judgment. For big font images the system gives better results. It gives reliable results greater than 20 pixel height (single character base height). If the pixel height of the character in the document image is greater than 20 pixels, then the output text is readable (with minimum error).

4. Conclusion

This paper presents Tamil Optical Character Recognition system which uses image binarization, Line segmentation, Character segmentation, Template matching and Tamil character reconstruction. The advantage of this method, Template matching makes Tamil character recognition simple. The system gives considerable results greater than 20 pixel base height of a character in the document image. The discovered result gives good accuracy for big fonts. The system loses its accuracy when the font size in the document image is small. But in normal practice while scanning the document, the image with high character pixel height is not possible. So the template matching method is less desirable for the Tamil character recognition system.But mostly the documents have small fonts. The system has not considered the skewed image. The future direction of the work is to apply skew correction in the scanned document image, and the Tamil character recognition system for all size of fonts.

5. Reference

1. Chopra SA, Ghadge AA, Padwal OA, Punjabi KS, Gurjar GS. Optical character recognition. International Journal of Advanced Research in Computer and Communication Engineering. 2014; 3(1):4956–8.
2. Dedgaonkar SG, handavale AA, Sapkal AM. Survey of methods for character recognition. International Journal of Engineering and Innovative Technology. 2012; 1(5):180–9.
3. Kannan RJ, Prabhakar R. A comparative study of optical character recognition for Tamil script. Euro J Scientific Res. 2009; 35(4):570–82.
4. Charles PK, Harish V, Swathi M, Deepthi CH. A review on the various techniques used for optical character recognition. International Journal of Engineering Research and Applications. 2012; 2(1):659–62.
5. Shrivastava V, Sharma N. Artificial neural network based optical character recognition. arXiv preprint arXiv:1211.4385. 2012; 3(5):73–80.
6. Alon J, Athitsos V, Sclaroff S. Online and offline character recognition using alignment to prototypes. IEEE Proceedings 8th International Conference on Document Analysis and Recognition; 2005. p. 839–43.
7. Yasser A. Preprocessing techniques in character recognition. INTECH Open Access Publisher; 2010. p. 1–20.
8. Su B, Lu S, Tan SL. Combination of document image binarization techniques. Document Analysis and Recognition (ICDAR). International Conference on IEEE; 2011. p. 22–6.
9. Richard GC, Lecolinet E. A survey of methods and strategies in character segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1996; 18(7):690–706.
10. Karmakar P, Nayak B, Bhoi N. Line and word segmentation of a printed text document. International Journal of Computer Science and Information Technologies. 2014; 5(1):157–60.
11. Aparna KG, Ramakrishnan AG. A complete Tamil optical character recognition system. Document Analysis Systems V. Springer Berlin Heidelberg; 2002. p. 53–7.
12. Kanimozhi VM, Muthumani I. Optical character recognition for English and Tamil script. IJCSIT. 2014; 5(2):1008–10.
13. Due TO, Jain AK, Taxt T. Feature extraction methods for character recognition-A survey. Pattern Recognition. 1996; 29(4):641–62.
14. Seethalakshmi R, Sreeranjani TR, Balachandar T. Optical character recognition for printed Tamil text using Unicode. Journal of Zhejiang University Science A. 2005; 6(11):1297–305.
15. Kannan RJ, Prabhakar R, Suresh RM. Off-line cursive handwritten Tamil character recognition. IEEE International Conference on Security Technology, SECTECH'08; 2008. p. 159–64.
16. Dan SB, Kopec GE, Dasari L. Measuring document image skew and orientation. IS&T/SPIE's Symposium on Electronic Imaging: Science and Technology. International Society for Optics and Photonics; 1995. p. 302–16.
17. Adams M. Algorithm for text document de-skewing. EECS 490-Digital Image Processing Session-1; 2004.