Available online at www.elixirpublishers.com (Elixir International Journal)

Advanced Engineering Informatics

Elixir Adv. Engg. Info. 36 (2011) 3584-3586

ISSN: 2229-712X

Digitalization of handwritten classical Tamil language using first order and second order statistical features

R.S. Sabeenian

Department of ECE, Sona College of Technology, Salem- 636005, TamilNadu.

ARTICLE INFO

Article history:

Received: 21 May 2011;

Received in revised form:

8 July 2011;

Accepted: 18 July 2011;

Keywords

Optical Character Recognition (OCR),
Scanned images,
Neural network,
Tamil characters.

ABSTRACT

Rapid growth of technology and prevalent use of computers in the business and other areas, more and more organizations are converting their handwritten paper documents into electronic documents that can be processed by computers. Handwriting recognition has attracted voluminous research in recent times. Optical Character Recognition (OCR) deals with machine recognition of characters present in an input image obtained using scanning operation. It refers to the process by which scanned images are electronically processed and converted to an editable text. Almost all the existing handwritten character recognition techniques use neural network approach, which requires lot of pre-processing and hence accomplishing these problems using neural network is a tedious task. In this paper we propose a novel solution for performing character recognition in TAMIL, the official language of the south Indian province of TamilNadu. Pursued by the pre-processing techniques, Segmentation and Feature Extraction are done for recognizing handwritten Tamil characters, which improves the efficiency. The tolerance of the system is evident as it can overwhelm the complexities arise out of font variations and proves to be flexible and robust. These initial results are promising and warrant further research in this direction.

Introduction

Machine simulation of human functions has been a very challenging research field since the advent of digital computers. Character and handwriting recognition has a great potential in data and word processing[5]. Combined with a speech synthesizer, it can be used as an aid for people who are visually handicapped. However, less attention had been given to Indian language recognition. Some efforts have been reported in the literature for Tamil script. Therefore, enabling interaction with computers in our native language TAMIL and in a natural way such as handwriting is absolutely necessary.

Tamil is a Dravidian language spoken predominantly by Tamil people all over the world. It is the first Indian language to be declared as a classical language by the government of India in 2004.Tamil literature has existed for over two thousand years [2]. Tamil has a large alphabet size, which has 12 vowels, 23 consonants and one special character (AK). Vowels and consonants are combined to form composite letters; making a total of 247 different characters. The rise of the Internet has triggered a dramatic growth in the number of Tamil blogs and specialist portals catering to political and social issues. Hence there is a need of digitizing Tamil documents from the ancient and old era to the latest, helps in sharing the data through the Internet.

System Analysis

Problem Statement

When a document is scanned, it gets saved as an image file, in which editing the content is not possible. Considering this as the problem statement, we have proposed this project.

Existing System

There are few algorithms proposed for digitizing the handwritten Tamil language. They are as follows [2].

- Using Octal graph
- Using Discrete Wavelet Transform
- Using Neural Networks

Drawbacks

The results for the segmentation and recognition of touching handwritten characters has not been very good and still there is a need for improvement so that they can be used in real world applications.

These algorithms require lot of pre-processing steps and hence accomplishing these problems using neural network is a tedious task [2-4].

Proposed System

A novel system for recognition of handwritten Tamil characters is presented. In this effort, handwritten recognition system for Tamil based on System Training & Interpretation is proposed. The scanned document image is pre-processed to ensure that the characters are in a suitable form. Then the line, word and characters are segmented and features are extracted from the segmented characters.

Higher degree of accuracy in results can be obtained with the implementation of this approach on a comprehensive database.

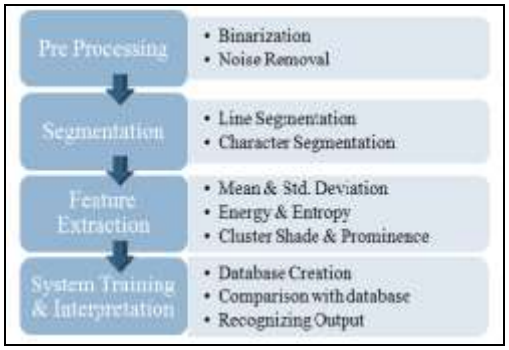

Figure 1: Proposed System

Tele:

E-mail addresses: sabeenian@gmail.com, sabeenian@sonatech.ac.in

The Proposed System Architecture

Handwritten character recognition system includes three stages,

- Image pre-processing
- Segmentation
- Feature extraction
- System Training and Interpretation

The process of handwriting recognition involves extraction of some defined characteristics called features to classify an unknown character into one of the known classes. Pre-processing is primarily used to reduce variations of handwritten characters. A feature extractor is essential for efficient data representation and extracting meaningful features for later processing. A classifier assigns the characters to one of the several classes.

Methodology

Pre-Processing

These preceding tasks make certain the scanned document is in a suitable form so as to ensure the input for the subsequent recognition operation is intact. The process of refining the scanned input image includes several steps.

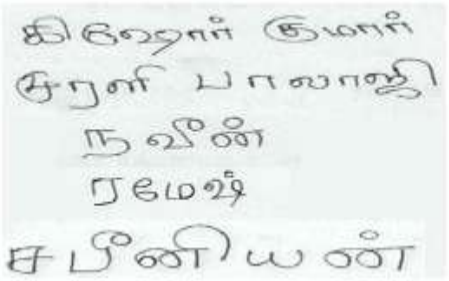

Figure 2: Scanned input Image

The pre-processing stage comprises the following steps:

- Binarization
- Noise Removal

Binarization

Extraction of foreground (ink) from the background (paper) is called as Thresholding. The global Thresholding [5] gray-scale [6] value of the document image is taken and the image is converted into binary image.

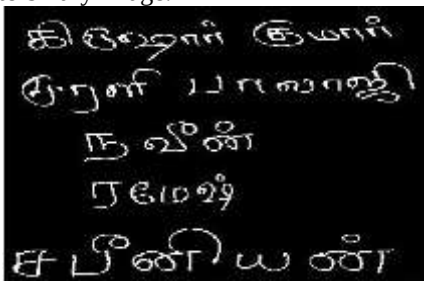

Figure 3: Binary Image

Noise Removal

The presence of noise can cost the efficiency of the character recognition system. We have used median filtering [1] for the removal of the noise from the image.

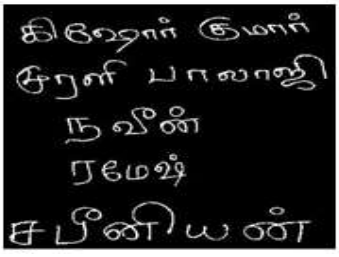

Figure 4: Noise Removed Image

Segmentation

Segmentation is a process of distinguishing lines and even characters of a hand written or machine printed document, a crucial step as it extracts the meaningful regions for analysis.

The Segmentation stage comprises two steps,

- Line Segmentation
- Character Segmentation

Line Segmentation

Each word of the line resides on the imaginary line that people use to assume while writing and a method has been formulated based on this notion.

The local minima [9] points are calibrated from each component to approximate this imaginary baseline. To guesstimate and categorize the minima of all components and to recognize different handwritten lines clustering techniques can be deployed [3].

Figure 5: Segmented Lines

Character Segmentation

Most of the character segmentation issues usually concentrate on discerning the gaps between the characters to distinguish them from one another.

This process of discriminating words emerged from the notion that the spaces between words are usually larger than the spaces between the characters[2].

Segmentation of words [3] in to its constituent characters is touted by most recognition methods.

Features like ligatures and concavity are used for determining the segmentation points. The algorithm exploits the caps between character segments and heights of character segments too.

Figure 6: Segmented Characters

Feature Extraction

In this stage, each pre-processed sample is transformed into a sequence of feature vectors. This is done by using following features,

- Mean
- Standard Deviation
- Entropy
- Energy
- Cluster Shade
- Cluster Prominence

Mean

Mean calculates the mean values of the elements along different dimensions of an array [5-6].

Standard Deviation

It computes the standard deviation of the values in Array.

Entropy

Entropy [10] is a statistical measure of randomness that can be used to characterize the texture of the input image.

$$s = \left(\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right)^{\frac{1}{2}}$$

where

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Energy

Energy is a measure of sum of squared value of each pixel value in the Co occurrence matrix.

$$\text{Energy} = \sum_{i,j} C(i,j)^2$$

Cluster Shade & Cluster Prominence

They are the measures of skewness of the matrix. Lack of symmetry can be found through this. If the values are high the image is not symmetry.

$$\text{Cluster Shade} = \sum_{i,j} ((i - \mu_i) + (j - \mu_j))^3 C(i,j)$$

$$\text{Cluster Prominence} = \sum_{i,j} ((i - \mu_i) + (j - \mu_j))^4 C(i,j)$$

Where,

$$\mu_i = \sum_i i \sum_j C(i,j)$$

$$\mu_j = \sum_j j \sum_i C(i,j)$$

C (i, j) – Gray level co-occurrence [4] matrix

System Training and Interpretation

Data Base

The features of the standard literals are extracted using a function in MATLAB. The MAT (matrix) file is created using SAVE command.

Comparison

The standard MAT file [8] created is retrieved using LOAD command. Now the features for segmented literal are weighed against the standard database. Finally the corresponding digital literal is printed.

Interfacing with Text pad

Text pad is interfaced with the MATLAB coding text pad using the FILE concept [7]. Handwritten document is printed in Text pad which can be edited by the user.

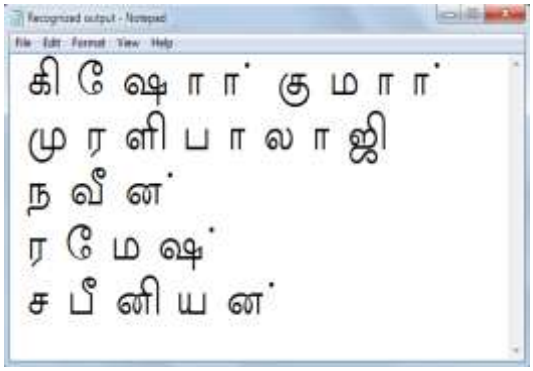

Figure 7: Digitalized Output

Conclusion

An approach to recognize cursive Tamil script based on system training and interpretation. We proposed an approach described the Pre-processing, Segmentation and Feature extraction process in detail. Considerable increase in accuracy levels has been found on comparison of our method with the others for character recognition. Furthermore, this recognition model poses to be more compatible for other Indian scripts too. With the addition of sufficient pre processing the approach offers a simple and fast structure for fostering a full OCR system.

References

1. Andrew W. Senior and Anthony J. Robinson, "An Off-Line Cursive Handwriting Recognition System", IEEE Transactions on pattern analysis and machine intelligence, vol. 20, pp: 309-320, October (1996).
2. R. Jagadeesh Kannan and R. Prabhakar "An Improved Handwritten Tamil Character Recognition System using Octal Graph" Journal of Computer Science 4 (7): 509-516, 2008, ISSN 1549-3636 2008 Science Publications
3. Mohamed Cheriet, Nawwaf Kharma, Cheng-Lin Liu, Ching Suen, Wiley- Interscience, "Character Recognition Systems: a guide for students and practioners" (2007) .
4. Dr. R.S. Sabeenian, "Handwritten text to digital text conversion using Radon Transform And Back Propagation Network" (RTBPN), published in Springer international conference on advances in information and communication technologies ICT 2010, held on September 7 to 9 2010 at Cochin, India. Volume 101, issue 3, pp 498-500.
5. Senior, A.W., "Off-line handwriting recognition: A review and experiments'\", Technical Report 105, Cambridge University Engineering Department, England, December 1992.
6. Sabeenian, R.S., Palanisamy, V.: Rotation Invariant Texture Characterization and Classification using Radon and Wavelet Transform. Published in the International Journal of Computational Intelligence and Health Care Informatics 1(2), 95–100, (2008)