IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 8, August 2016

Collection of Offline Tamil Handwriting Samples and Database Creation

D. Rajalakshmi [1], Dr. S.K. Jayanthi [2]

Part Time Doctoral Research Scholar, Department of Computer Science, Vellalar College for Women, Erode, India[1]

Head of the Department, Department of Computer Science, Vellalar College for Women, Erode, India [2]

Abstract: Handwriting analysis is an important research area in today's world. In many languages, recognition of handwritten letters, numerals and words were focused by the researchers. To the best of our knowledge, there is no public domain database of Tamil Scripts available for handwriting analysis. Hence the idea of creating own database for Tamil Handwriting raised. This paper describes the efforts to create a collection of handwritten samples in Tamil to support research activities in offline Tamil handwriting analysis and recognition. The database is created with a view to make it available publicly for research purpose, which contains handwritten- scripts of 300 document images of different writers. The collected documents were converted into document images and then preprocessed. The paper concludes with the status of the effort and future directions.

Keywords: Handwriting samples, Document Images, Tamil Script, Preprocessing

I. INTRODUCTION

Handwriting is part of our life from day to day commercial transactions and to personal communications. Handwriting analysis and recognition finds a lot of potential applications that encompass the reach of Information Technology to the common man. Handwriting analysis is generally classified into offline and online [1]. Online refers the inclusion of temporal information along with the handwriting data. Offline handwritten data is obtained by scanning the handwritten documents. Due to the lack of temporal information, off-line is considered more complex than online. Additionally, the off-line method is the one corresponds to the conventional reading and writing job performed by the human [2]. Most of the offline texts are recognized using Hidden Morkov Models and Neural Networks [3] [10].

The challenges in online handwriting recognition in Indic scripts presented by Bharath. A et al, and also pointed some resources available for Indic script recognition research [14]. A script dependent approach to segment online handwritten tamil words into symbols presented by Suresh Sundaram and A. G. Ramakrishnan [15].

The techniques for offline handwriting recognition inlcludes feature extraction and Radial basis function [16][17].Writer idenfication for offline Tamil handwriting presented by Jayanthi, S. K., and D. Rajalakshmi and also pointed our the unavailability of databases for Tamil handwritten documents [18]. In the field of handwriting research, databases are so important to analyse and evaluate a particular system or methodology. Databases for scripts like English, Roman and Chinese already exist, whereas no such databases exist for Indic Scripts. The ten official Indic scripts -Devanagari, Tamil, Gurmukhi ,Telugu, Kannada, Guajarati, Oriya, Bengali, Malayalam and Urdu - most of which have not seen much targeted research in human language technologies, regardless of the large number of users. One of the barrier for the language technology research in the Indian context has been the lack of significant linguistic resources, and this is especially true for handwriting. Tamil, the native language of Tamil Nadu state in India, is one of the oldest language has several million speakers across the world and is an official language in countries such as Sri Lanka, Malaysia and Singapore.

Tamil has a large alphabet size and hence text entry through QWERTY keyboard is cumbersome. The penetration of Information Technology becomes harder in a country such as India where the greater part read and write in their native language. Therefore, enabling interaction with computers in the native language and in a natural way such as handwriting is absolutely necessary.

One of the major stumbling blocks for language technology research in the Indian context has been the lack of significant publicly available linguistic resources, and this is especially true for handwriting research. It is imperative that tools and data formats be standardized and validated datasets be created and made available to change the condition.

It should be mentioned that such datasets would also benefit research in handwritten document analysis, writer identification, script identification, handwritten document indexing and retrieval, and so forth. Data collection is a challenge for Indic scripts due to the large number of characters.

The paper is organized as follows. In Section 2, the related work is portrayed. In Section 3, the data collection methodology is explained. Section 4 describe the results and Section 5 concludes the work.

DOI 10.17148/IJARCCE.2016.5836

II. RELATED WORK

The first corpus for modern written Tamil was built in the Central Institute of Indian Languages (CIIL), Mysore in 1987. The CIIL in collaboration with the Tokyo University of Foreign Studies, Japan built a corpus on Tamil textbooks. In 1991 under the scheme Technological Development for Indian Languages (TDIL), the Department of Electronics, and Govt. of India launched a project called Development of Corpora of Texts of Indian Languages.

The CIIL was entrusted to build Corpora for the four major Dravidian languages. Texts printed during the period 1981 to 1990 were selected to represent the modern Tamil. They are collected from 6 major categories, such as Aesthetics (Literature and Fine arts), Social Sciences, Natural, Physical and Professional Sciences, Commerce, Administration and Technology, and Translated Texts. They are further classified into 76 minor categories, to cover various domains of language use. The size of the Tamil corpus is 3.6 million words. Some of the practical challenges encountered in creating a handwriting corpus for an Indic script are described by Agrawal et al. [4] [12] [13].

Jayaram et al. Focuses about the corpus construction of EMILE database and discussed about the issues on south Indian Corpus collection [5]. Nethravathi, B., et al. describes how the database has been created for Kannada and Tamil handwriting recognition [6]. The process of creating a reduced symbol list, which includes all the basic symbols of the character set has been described. The focus is on the process of collecting data, the devices used, the criteria for selection of writers and why the reduction in number of symbols is required. A semi-automated tool for annotating handwritten data from the stroke to the word level is also described.

Thadchanamoorthy et al. [2013] developed a city name data base for the postal automation [7]. For online Handwritten Tamil Word Recognition, Bharath et al.[2007] proposed a data driven HMM model [8].
Bhattacharya et al. described an Image database of handwritten isolated numerals of three different Indian scripts (Devanagari, Bangla and Oriya) for research purpose [11].

III. DATA COLLECTION METHODOLOGY

For performing handwriting analysis experiments, we have created our own data base with a view to make it available for research purpose. The database can be used for the following purposes:

- Handwritten Document Segmentation
- Handwritten Character Recognition
- Handwriting Analysis
- Writer Identification and Verification
- Text to Speech Conversion

The writers were asked to write text lines in A4 size pages. No constraints were forced regarding the content and use of pen. Writers chosen were Undergraduate Students of a College. 300 writers contributed to the task of handwriting.

The scripts were collected and scanned using flatbed scanner at a resolution of 300 dpi and stored as jpg format images. Following factors were observed from handwritten documents

- Variation in script size
- Slant
- Difference in inter-space between lines and words of the page
- Variation in number of words/characters in a line

Our data collection methodology is shown in Figure 1 which depicts various phases in database creation.

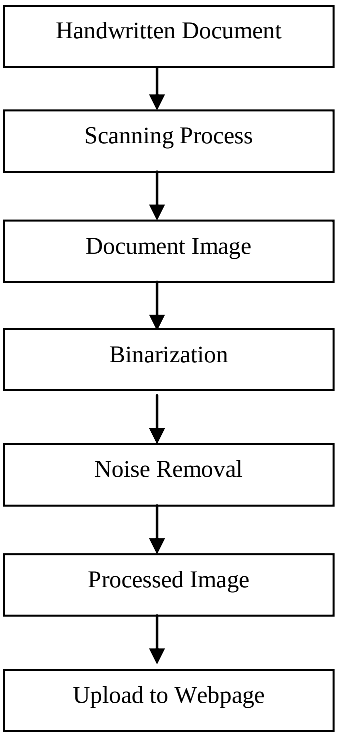


Figure 1 Steps in Creation of database

A. Preprocessing

Preprocessing techniques enhances the image features and support for better feature extraction. In our attempt, preprocessing techniques include the tasks such as binarization, noise removal and thinning.

Image pre-processing refers to the operations on images at the lowest level of abstraction. The aim of pre-processing is to improve the image data that suppresses undesired distortions and enhances some image features that are relevant for further processing and analysis. Median filter is applied to remove the noise and the resultant image is converted into gray scale image.

B. Image Binarization

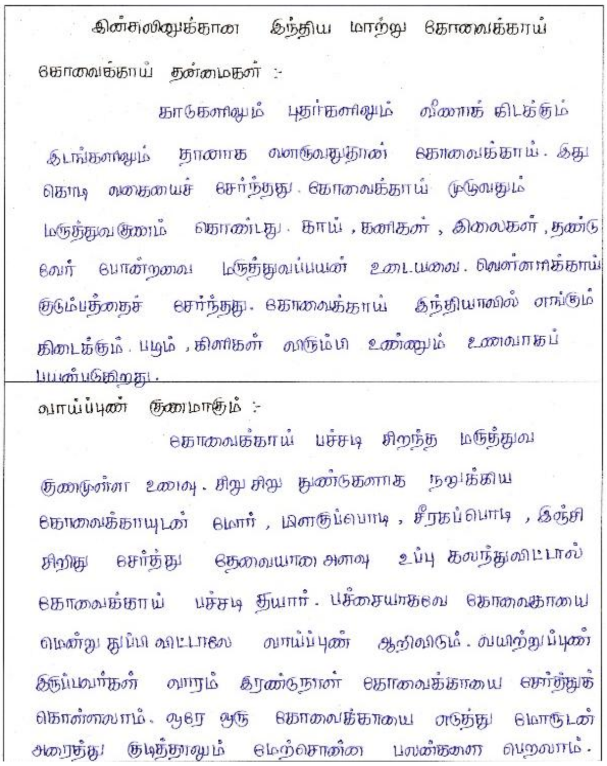
The binarization process converts the grayscale image into binary form which facilitates several tasks such as document analysis, script recognition and character recognition. In handwriting analysis, it is necessary to convert handwritten text image to binary because the text may be written using different pens. Thresholding plays a vital role in image binarization. Several method can be found in the literature and one of the popular method is Otsu's Global thresholding technique. This image is converted to binary form using Otsu's method [9]. Figure 2 shows the document image and its binary form.

C. Thinning

Thinning is a morphological operation that is used to remove selected foreground pixels from binary images. Thinning is applied to binary images and produces another binary image as output. It is necessary to apply thinning operation to the text image since the text written vary in thickness depending on the writers and type of pen used. Thinning process supports better feature extraction for the purpose of analysis.

D. Noise Removal

Noise can appear in a document image during the conversion process and also caused by dirt on the document. This noise can be composed of a group of pixels, but by definition, these are assumed to be much smaller than the size of the text objects. Noise can be removed by simple filters like median or morphological operators. In median filtering, the idea is to replace the current point in the image by the median of the brightness in its neighborhood.



இன்சுலினுக்கான இந்திய மாற்று கோவைக்காய்

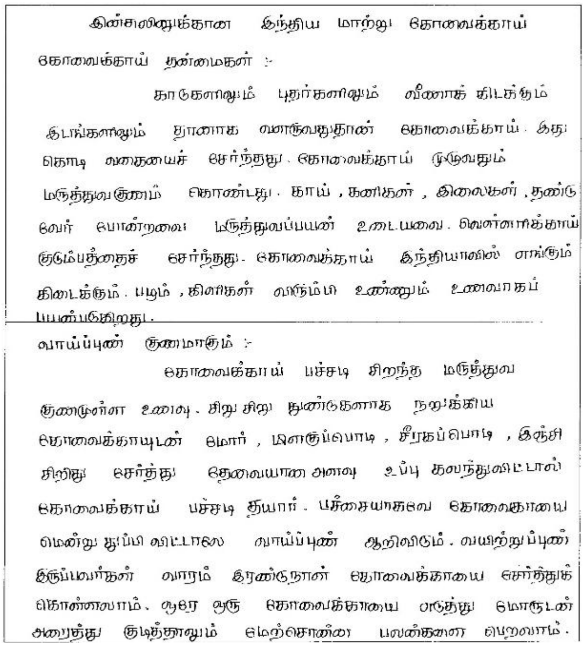
கோவைக்காய் தன்மைகள் :-

காடுகளிலும் புதர்களிலும் வீணாக் கிடக்கும் இடங்களிலும் தானாக வளருவதுதான் கோவைக்காய். இது கொடி வகையைச் சேர்ந்தது. கோவைக்காய் முழுவதும் மருத்துவகுணம் கொண்டது. காய், கனிகள், இலைகள், தண்டு வேர் போன்றவை மருத்துவப்பயன் உடையவை. வெள்ளரிக்காய் குடும்பத்தைச் சேர்ந்தது. கோவைக்காய் இந்தியாவில் எங்கும் கிடைக்கும். பழம், கிளிகள் விரும்பி உண்ணும் உணவாகப் பயன்படுகிறது.

வாய்ப்புண் குணமாகும் :-

கோவைக்காய் பச்சடி சிறந்த மருத்துவ குணமுள்ள உணவு. சிறு சிறு துண்டுகளாக நறுக்கிய கோவைக்காயுடன் மோர், மிளகுப்பொடி, சீரகப்பொடி, இஞ்சி சிறிது சேர்த்து தேவையான அளவு உப்பு கலந்துவிட்டால் கோவைக்காய் பச்சடி தயார். பச்சையாகவே கோவைக்காயை மென்று துப்பி விட்டாலே வாய்ப்புண் ஆறிவிடும். வயிற்றுப்புண் இருப்பவர்கள் வாரம் இரண்டுநாள் கோவைக்காயை சேர்த்துக் கொள்ளலாம். ஒரே ஒரு கோவைக்காயை எடுத்து மோருடன் அரைத்து குடித்தாலும் மேற்சொன்ன பலன்களை பெறலாம்.

Figure: 2 (a) Original document image



இன்சுலினுக்கான இந்திய மாற்று கோவைக்காய்

கோவைக்காய் தன்மைகள் :-

காடுகளிலும் புதர்களிலும் வீணாக் கிடக்கும் இடங்களிலும் தானாக வளருவதுதான் கோவைக்காய். இது கொடி வகையைச் சேர்ந்தது. கோவைக்காய் முழுவதும் மருத்துவகுணம் கொண்டது. காய், கனிகள், இலைகள், தண்டு வேர் போன்றவை மருத்துவப்பயன் உடையவை. வெள்ளரிக்காய் குடும்பத்தைச் சேர்ந்தது. கோவைக்காய் இந்தியாவில் எங்கும் கிடைக்கும். பழம், கிளிகள் விரும்பி உண்ணும் உணவாகப் பயன்படுகிறது.

வாய்ப்புண் குணமாகும் :-

கோவைக்காய் பச்சடி சிறந்த மருத்துவ குணமுள்ள உணவு. சிறு சிறு துண்டுகளாக நறுக்கிய கோவைக்காயுடன் மோர், மிளகுப்பொடி, சீரகப்பொடி, இஞ்சி சிறிது சேர்த்து தேவையான அளவு உப்பு கலந்துவிட்டால் கோவைக்காய் பச்சடி தயார். பச்சையாகவே கோவைக்காயை மென்று துப்பி விட்டாலே வாய்ப்புண் ஆறிவிடும். வயிற்றுப்புண் இருப்பவர்கள் வாரம் இரண்டுநாள் கோவைக்காயை சேர்த்துக் கொள்ளலாம். ஒரே ஒரு கோவைக்காயை எடுத்து மோருடன் அரைத்து குடித்தாலும் மேற்சொன்ன பலன்களை பெறலாம்.

(b) Result after Preprocessing

IV. RESULTS

Due to the non-availability of database for Tamil handwriting, samples were collected from 300 different writers. Writing samples were obtained by scanning the documents. After scanning, the gray scale images were converted into binary images by using Otsu's method [9]. The resolution of scanning was 300 dpi and the resulting images were stored as gray scale images.

V. CONCLUSION

A database of Tamil Handwritten documents collected from different writers is created. The aim of the paper is to tell about the handwritten document database available for unrestricted access. The handwritten samples are collected from 300 different writers who are undergraduate students. The documents are scanned using HP Scanner at the resolution of 300 dpi. The processed document images are uploaded to the website www.tamilhwsamples.in and available for free download.

REFERENCES

[1] Plamondon.R, and Sargur N. Srihari. "Online and off-line handwriting recognition: a comprehensive survey." Pattern Analysis and Machine Intelligence, IEEE Transactions on 22.1 (2000): 63-84.

[2] Kannan, R. Jagadeesh, R. Prabhakar, and R. M. Suresh. "Off-line Cursive Handwritten Tamil Character Recognition." 2008 International Conference on Security Technology. IEEE, 2008.

[3] Sumathi, C. P., and S. Karpagavalli. "Techniques and methodologies for recognition of Tamil typewritten and handwritten characters: A survey." International Journal of Computer Science and Engineering Survey 3.6 (2012): 23.

[4] Agrawal, Mudit, Ajay S. Bhaskarabhatla, and Sriganesh Madhvanath. "Data collection for handwriting corpus creation in Indic scripts." International Conference on Speech and Language Technology and Oriental COCOSDA (ICSLT-COCOSDA 2004), New Delhi, India (November 2004). 2004.[20].

DOI 10.17148/IJARCCE.2016.5836

[5] Baker, P., Hardie, A., McEnery, T., & Jayaram, B. D. (2003, April). Corpus data for South Asian language processing. In Proceedings of the 10th Annual Workshop for South Asian Language Processing, EACL.

[6] Nethravathi, B., et al. "Creation of a huge annotated database for Tamil and Kannada OHR." Frontiers in Handwriting Recognition (ICFHR), 2010 International Conference on. IEEE, 2010.

[7] Thadchanmoorthy,S., et al, “Tamil Handwritten City Name Database Development and Recognition for Postal Automation.” Document Analysis and Recognition (ICDAR), 2013 12th International Conference on. IEEE ,2013.

[8] Bharath, A. “Hidden Markov Models for online handwritten Tamil word Recognition.” Document Analysis and Recognition,2007. ICDAR 2007. Ninth International Conference on. Vol. 1. IEEE, 2007.

[9] N. Otsu, “A threshold selection method from grey level histogram” IEEE Trans. on SMC, vol. 9, pp. 62-66, 1979.

[10] Kannan, R. Jagadeesh, R. M. Suresh, and A. Selvakumar. "Applications of Hidden Markov Model to Recognize Handwritten Tamil Characters." Advances in Communication, Network, and Computing. Springer Berlin Heidelberg, 2012. 282-290.

[11] Bhattacharya, Ujjwal, and B. B. Chaudhuri. "Databases for research on recognition of handwritten characters of Indian scripts." Document Analysis and Recognition, 2005. Proceedings. Eighth International Conference on. IEEE, 2005.

[12] Ganesan, M. Tamil Corpus Generation and Text Analysis. In Conference papers Conference papers (p. 193).

[13] Bhaskarabhatla, Ajay S., and Sriganesh Madhvanath. "Experiences in Collection of Handwriting Data for Online Handwriting Recognition in Indic Scripts." LREC. 2004.

[14] Bharath. A., and Sriganesh Madhvanath. "Online handwriting recognition for Indic scripts." Guide to OCR for Indic Scripts. Springer London, 2010. 209-234.

[15] Suresh Sundaram and A. G. Ramakrishnan, “Attention-feedback based robust segmentation of online handwritten isolated Tamil words,” ACM Transactions on Asian Language Information Processing (TALIP), Vol. 12 (1), March 2013, Article No. 4.

[16] Safi, L. Anlo, and K. G. Srinivasagan. "Offline Tamil handwritten character recognition using zone based hybrid feature extraction technique." International Journal of Computer Applications 65.1 (2013).

[17] Ashok, J., And E.G Rajan. "Off-Line Hand Written Character Recognition Using Radial Basis Function." International Journal of Advanced Networking & Applications 2.4 (2011).

[18] Jayanthi, S. K., & Rajalakshmi, D.,”Writer identification for offline Tamil handwriting based on gray-level co-occurrence matrices.” In 2011 Third International Conference on Advanced Computing (pp. 187-192). IEEE.

DOI 10.17148/IJARCCE.2016.5836