Novel Approaches to Machine Recognition of Handwritten Tamil Documents

In spite of the wide spread use of computers in document processing, hand printed documents play a major role in our life. The machine analysis of hand printed documents has great significance in this era of paperless offices. The machine analysis of documents prepared with paper and pen is referred to as offline handwritten document recognition systems. Compared to this, the online handwritten character recognition system focuses on the recognition of characters written on sensor device like touch pad which can be directly converted into digital form.

Recognition of individual character recognition is the first step in the machine analysis of handwritten documents. This study focuses on the recognition of isolated characters obtained from handwritten Tamil documents.

The efficiency of offline handwritten character recognition (HCR) system depends on the script used. In foreign languages like English and Chinese, offline HCR system technology is matured. In the case of Indian languages, several HCR system frame works with high accuracy are reported, especially in languages/scripts like Devanagari, Bangla, Kannada, Telugu and Malayalam. In Tamil, though many research papers are published in this area, the results reported are inadequate for the design of efficient HCR systems. This is the motivation behind this thesis.

Unlike online HCR systems, which utilize factors like order of strokes, speed, pen up and pen down information, offline systems use scanned images of handwritten documents. Hence offline HCR system design is more challenging. An offline HCR system has to address issues such as variation in writing style of individuals at different times and also among different individuals (size, shape, thickness of characters etc). These issues limit the accuracy of offline HCR systems. Due to these challenges design of offline HCR system has become an interesting research area in image processing pattern recognition.

The following are the objectives of this work:

- Study the features of Tamil characters.
- Creation of handwritten isolated Tamil character database.

- A detailed study on different image preprocessing and feature extraction methods.
- A detailed study on different classifiers
- Performance evaluation of different classifier -feature set pairs in recognizing handwritten isolated Tamil characters.
- To propose new or modified feature extracting method suitable for Tamil character.
- To propose new or modified preprocessing algorithms suitable for Tamil character.

Tamil is a Dravidian language spoken predominantly by Tamil people of South India and North-east Sri Lanka. It is one of the 22 scheduled languages of India and was the first Indian language to be declared as a classical language by the Government of India in 2004. Tamil is one of the longest surviving classical languages in the world. Tamil is a diglossic language. The Tamil script consists of 12 Vowels or Uir alphabets, 18 Consonants or Mai alphabets, one special character the āytam, written as ∴, 216 Syllabic alphabets or Uir Mai alphabets and 6 Grantha script alphabets. Some features of Tamil character are:

- Direction of writing: left to right in horizontal lines.
- There is no cursive writing.
- There is no upper and lower case difference.
- There exits similarities between some characters.
- The basic geometric features of Tamil character are the number of loops, number of T joints, total number of end points, vertical and horizontal bars and total no. of arcs etc.

As a first step, a database consisting of 72 characters × 300 samples/character is created. The samples include documents written by people in different age groups and different educational backgrounds. An HCR system typically consists of the following steps:-

1. Scanning of documents at an appropriate resolution, typically 300-1000 dots per inch (as gray images).
2. Preprocessing : a) Binarization (two-level thresholding)
   b) Segmentation to isolate individual character
3. Feature Extraction: Extracting meaningful features
4. Classification: Recognition using one or more classifier
5. Contextual verification on post processing

The different preprocessing steps employed in this research work are detailed below:

To bring about size uniformity, the character image samples are normalized to 72X72 pixel sizes using nearest neighbor interpolation method. Normalized image is then converted in to binary image using Otsu's threshold selection technique. This is followed by thinning or contour tracing approaches are explained with the help of experiments conducted on handwritten Tamil characters. The thinning algorithms viz Improved Parallel Thinning (IPT), Rotation Invariant Thinning (RIT), Nagandraprasad Wang Gupta (NWGT) Thinning, Zhang Suen Thinning (ZST), Parker Thinning (PT) and Stentiford Thinning (ST) are implemented and applied on Tamil handwritten characters. The drawbacks of these thinning algorithms are identified. A modification for Stentiford Thinning algorithm is proposed. The proposed thinning algorithm is found to give better result compared to the other six thinning algorithms.

A fast, efficient and accurate contour extraction method using eight sequential Euclidean distance (8SED) map and connectivity criteria based on maximal disk is proposed. The connectivity criterion is based on a set of point pairs along the image boundary pixels, which are the nearest point under consideration and its neighbors. The performance of the proposed algorithm in terms of execution time and contour shape is compared with two well known contour tracing algorithms- the Moore method and the Canny edge detection method. The results established the efficacy of the proposed algorithm as it has less computing time and never loses connectivity.

An important issue in character recognition is the selection of best discriminative features. This research work discusses various statistical and structural features and different combinations of them. The different feature extraction techniques employed in this work are chain code, gradient, zero crossing, normalized vector distance, division point and view based feature. As part of the work, a novel view based feature is proposed. The effectiveness of these features on handwritten Tamil character recognition is experimented using MultiLayer Perceptron (MLP), Support Vector Machine (SVM) and Extreme Learning Machine (ELM) classifiers. The proposed modified view based feature is found to be more suitable than the other features. A detailed analysis of the results of the experiments carried out with different features/combination of features and classifiers is presented.

The classifier employed in this work includes SVM, MLP and ELM. MLP and SVM available as part of popular WEKA tool is used for the work and for ELM which is a relatively new paradigm in ANN, Mat Lab code is used.

Major contributions:

- A database for research in the area offline Tamil HCR systems is created.
- A comparative study of different thinning algorithm is carried out. Based on that a modified thinning algorithm is proposed.
- A new contour tracing algorithm is proposed and is compared with two existing algorithms to establish its merits.
- A detailed investigation of performance of a set of selected features for the recognition of isolated Tamil characters is carried out.
- A modified view based feature extraction method is proposed and its effectiveness is established through experiments.

Publications based on the work specified

1) **Sobhana Mari S and G.Raju** "Different Meshing Techniques for Handwritten Tamil Character Recognition" Proceedings of the International Conference on Mathematics and Computer Science (ICMCS 2010), pp. 626-631, (**ISBN: 978- 81-908234-2-5),** GK Publishing Pvt.Ltd, Chennai.
2) **Sobhana Mari S and G.Raju** "Handwritten Tamil Character Recognition using chain code method" Proceedings of the International Conference on Mathematics and Computer Science (ICMCS 2011),pp.138-142 (**ISBN: 978- 81-920490-0-7),** GK Publishing Pvt.Ltd, Chennai.
3) **Sobhana Mari S and G.Raju** " Centroid based Feature Extraction for Handwritten Tamil Character Recognition" Proceedings of the National Conference on Image Processing (NCIMP 2010), pp. 203-209, **(ISBN: 978-81-8424-574-5)** Allied Publishing Pvt.Ltd, New Delhi
4) **Sobhana Mari S and G.Raju** "Chain code Based Approach with Meshing Techniques for Handwritten Tamil Character" Proceedings of the International Conference on Mathematical Computing and Management (ICMCM 2010), MACFAST, Kerala
5) **Sobhana Mari S and G.Raju** "Zero crossing Method for Handwritten Tamil Character Recognition Using MLP classifier" Proceedings of the National

Conference on Indian Language Computing (NCILC 2011), Cochin University of Science and Technology, Kochi, Kerala

6) **Sobhana Mari S and G.Raju** "Performance Comparison of Different Thinning Algorithms on Offline Handwritten Tamil Character Recognition" Book chapter on Computing and Communication (NCCC 2011), pp.124-130, **(ISBN: 978-81-8487-178-4),** Narosa publishing house Pvt. Ltd, 2011
7) **Sobhana Mari S and G.Raju** "Performance Comparison of SVM and MLP Classifier for Recognizing Handwritten Tamil Characters" Conference on Computing Paradigms and Bio-Informatics (CPBI-2012) M.G.University, Kerala
8) **Sobhana Mari S and G.Raju** "Tamil Handwritten Character Recognition Using Wavelet Transformations " Proceedings of the 2nd National Conference on Indian Language Computing (NCILC 2012), pp.44-50 Cochin University of Science and Technology, Kochi, Kerala