A Two Stage Classification Approach to Tamil Handwriting Recognition

S. Hewavitharana
Department of Computer Science
University of Colombo
Colombo 03, Sri Lanka
sanjika@cmb.ac.lk

H. C. Fernando
Sri Lanka Institute of Information Technology
Colombo 03, Sri Lanka
chandrika@sliit.lk

Abstract

This paper describes a system to recognize handwritten Tamil characters using a two-stage classification approach, for a subset of the Tamil alphabet. In the first stage, an unknown character is pre-classified into one of the three groups: core, ascending and descending characters. Then, in the second stage, members of the pre-classified group are further analyzed using a statistical classifier for final recognition. A recognition rate of 80% was achieved for the 1st choice and 97% for the top 3 choices.

1. Introduction

Character and handwriting recognition has a great potential in data and word processing for instance, automated postal address and ZIP code reading, data acquisition in bank checks, processing of archived institutional records, etc. Combined with a speech synthesizer, it can be used as an aid for people who are visually handicapped. As a result of intensive research and development efforts, systems are available for English language [1], [2] Chinese/Japanese languages [3], [4],[5] and handwritten numerals [6], [7]. However, less attention had been given to Indian language recognition. Some efforts have been reported in the literature for Devanagari [8], Tamil [9],[10],[11] and Bangla [12] scripts.

The process of handwriting recognition involves extraction of some defined characteristics called features to classify an unknown character into one of the known classes. The strategy used for recognition can be broadly classified into three categories, namely: structural, statistical and hybrid. Structural techniques use some qualitative measurements as features. They employ different methods such as rule-based, graph-theoretic and heuristic methods for classification. Statistical techniques use some quantitative measurements as features and an appropriate statistical method for recognition. In hybrid approach, these two techniques are combined at appropriate stages for representation of characters and utilizing them for recognition. Depending on the problem, anyone of these techniques can be utilized while accommodating the variations in handwriting.

In this paper, we propose a recognition system for handwritten Tamil characters. It uses a two-stage classification approach, which is a hybrid of structural and statistical techniques. In the first stage, known as preliminary classification, the unknown character is classified into one of three groups of Tamil characters. Structural properties of the text line are used for this classification. In the second stage, a statistical classifier recognizes the unknown character as one of the members of the pre-classified group.

The organization of the paper is as follows. Section 2 gives an introduction into the Tamil language. In section 3, the proposed system is described in detail. Results of the experimentation are presented in section 4. Finally in section 5, we present our conclusions and future work.

2. Tamil Language

Tamil, which is a south Indian language, is one of the oldest languages in the world. Although it has been influenced by Sanskrit to a certain degree, Tamil along with other south Indian languages are genetically unrelated to the descendants of Sanskrit such as Hindi, Bengali and Gujarati. Most Tamil letters have circular shapes; partially due to the fact that they were originally carved with needles on palm leaves, a technology that favored rounded shapes. The Tamil script is used to write the Tamil language in Tamil Nadu state of India, Sri Lanka, Singapore and parts of Malaysia, as well as to write minority languages such as Badaga [13].

The Tamil alphabet consists of 12 vowels, 18 consonants and one special character (AK). Vowels and consonants are combined to form composite letters, making a total of 247 different characters. The complete Tamil alphabet and composite character formations are given in [9]. However, with the advantage of having a separate symbol for each vowel in composite character formations, there is a possibility to reduce the number of symbols used by the alphabet. In character recognition point of view, only 67 symbols have to be identified to recognize all 247 characters [11]. We have considered 26 characters of the Tamil alphabet for our study.

3. Experimental System

The experimental system consists of five major sections. The first section deals with the data collection followed by preprocessing. Segmentation, Preliminary classification, Feature extraction and Recognition are the other sections described in here.

3.1 Data Collection

Data samples were colleted from different writers on A4 sized documents. They were scanned using a flat-bed scanner at a resolution of 100 dpi and stored as 8-bit grey scale images. In preprocessing, the document images were binarized using a global thresholding technique [14] to eliminate noise and extract the handwriting.

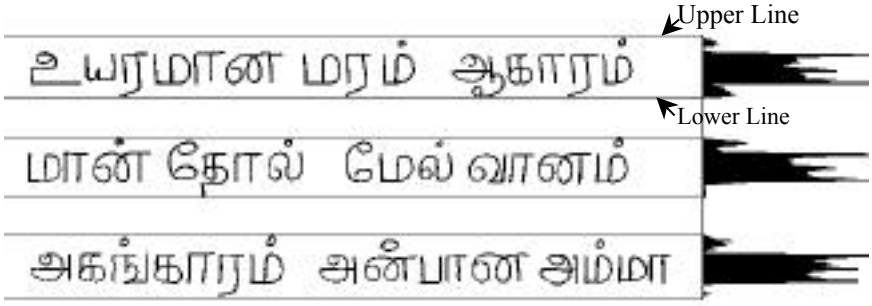

Fig. 1a: Line segmentation using horizontal projection profile

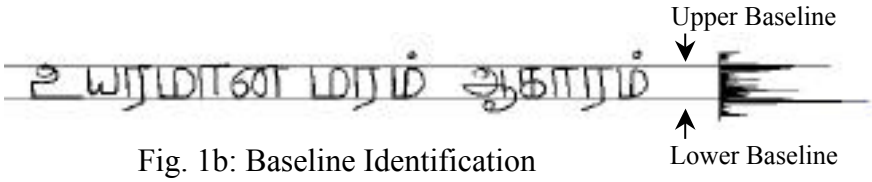

Fig. 1b: Baseline Identification

3.2 Segmentation

An input image consists of a uniform text area with distinct text lines. In the segmentation process, this is broken down into constituent text lines, words and finally into individual characters. The method is based on the horizontal projection profile of the image. Zero values or valleys in the projection profile correspond to the horizontal gaps between text lines. Each text line is identified using two reference lines; the *upper line* and the *lower line*. They correspond to the minimum and maximum zero value positions adjoining a text line, respectively. (See Fig. 1a)

We extract two more reference lines from each text line, namely, the *upper baseline* and the *lower baseline*. For this, we use a method similar to [1]. First derivative of the horizontal projection profile is calculated for each segmented text line. The local extrema of the first derivative in the two halves of the text line are taken to be the two baselines. The lines drawn across the two peaks in Fig. 1b indicate the two baselines.

We use a pre-formatted paper for the collection f handwriting so that we could guide the writer and simplify the process of reference line extraction. Each document has the four reference lines printed on it. However, these lines are completely eliminated during the binarization of the image and have no effect on the segmentation.

After the reference lines have been found, words and characters are extracted using the vertical projection profile of each text line. Word boundaries and character boundaries are distinguishable since the former are much wider than the latter. Ones all the characters have been segmented, the minimum bounding box of each character is identified eliminating the white space around it. Upper and lower boundary values of the minimum bounding box, along with the four reference lines, are sent to the next stage for preliminary classification.

3.3 Preliminary Classification

The aim of the preliminary classification is to reduce the number of possible candidates for an unknown character, to a subset of the total character set. For this purpose, the selected domain is categorized into three non-overlapping groups as in Fig. 2. The classification is based on the relative heights of each character in the three-strip frame determined by the four reference lines.

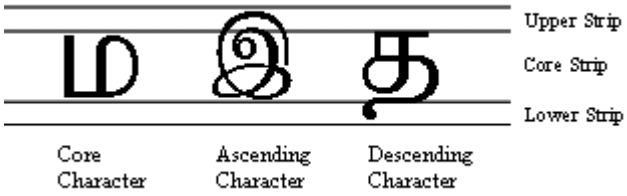

Fig. 2: Three pre-classification groups

Characters that lie within the two baselines are categorized into the *core characters* group. *Ascending characters* start in the core strip and extend towards the upper strip. *Descending characters* start in the core strip and extend towards the lower strip. The only character that does not fit into any of the groups is the 'virama' sign (_). However, it can be easily identified since it is the only character to appear within the upper strip.
Table 1 lists all the characters under consideration, classified into the above three pre-classification groups.

The subsequent recognition stage concentrates only on the pre-classified group and treat the members of the group as possible recognition candidates. Characters belonging to other groups are assumed to be invalid matches to the unknown and are not considered for the recognition.

3.4 Feature Extraction

All the segmented character images are then scaled into a common height and width (32x32 pixels) using a bilinear interpolation technique. The slant associated with the characters is negligible due to the use of a pre-formatted paper for data collection. Hence, no attempt was made to perform slant normalization. Each image is then divided into equal number of horizontal and vertical strips, producing a grid with square shaped zones. For each zone, the pixel density is calculated and therefore a vector created.

Different sizes of zones were used in the study ranging from 2x2 pixels to 16x16 pixels. When the zones size was small, it captured more detailed pixel variations. However, due to the varying nature of handwriting, there was high dissimilarity between the feature vectors of the same class. Large sized zones failed to capture the essential parts of characters, which make them distinct from others. The best results were produced by 4x4 pixel zones. Therefore, we decided to use 4x4 zones for feature extraction. This 64-dimension feature vector contained a value between 1 and 16 corresponding to the pixel density of each zone.

Table 1: Preliminary classification into 3 groups

Group	Characters of the group
Core Characters	க ங ச ட ண ப ம ய ல வ ள ன அ ஈ உ ஊ எ
Ascending Characters	இ ஃ
Descending Characters	ஞ த ந ர ழ ற ஆ ஏ ஐ ஒ ஓ ஔ

3.5 Recognition Process

A statistical classifier based on interval estimation is used for the recognition process. For each zone of size 4x4 pixels, we calculate an interval of values within which the mean pixel density of the population lies with 95% confidence. The upper and lower confidence limits of region *i* are calculated as follows:

Lower limit of conf. interval (LCL_i) = $(\bar{x}_i - 1.96 \times \frac{s_i}{\sqrt{n}})$

Upper limit of conf. interval (UCL_i) = $(\bar{x}_i + 1.96 \times \frac{s_i}{\sqrt{n}})$

where, $\bar{x}_i$ is the sample mean, S_i^2 is the sample variance and n is the sample size.

40 sample images from each character class are chosen as training data. For each class, LCL_i and UCL_i are calculated for each zone *i*, and is stored as the classifier.

Upon the receipt of an unknown image, the recognition process first extract the feature vector and then compare these values with the corresponding confidence intervals of the classifier. If a value is within the confidence interval, the response is 1, otherwise it is 0. The total number of matches is counted for each candidate class. The class with the highest matches is considered to be the one to which the character image belongs to. The best 3 matches are presented in the descending order.

4. Recognition Results

We trained the system with 1000 characters belonging to all the classes. The testing data contained a separate set of 800 characters. A portion of the training data was also used to test the system, to see how well the system represents the data it has been trained on.

A total of 50 text lines were subjected to segmentation and reference line identification. In all the cases, every character in each text line was correctly segmented. The reference line identification was almost 99% accurate resulting only 1% pre-classification error. Results of the recognition process is given in Table 2.

In the test set, a recognition rate of 79.9% was achieved for the 1st choice and 96.9% for the top 3 choices. Understandably, the training set produced much higher recognition rate than the test set.

Table 2: Recognition results

		Top 1	Top 2	Top 3	Mis-recognition	Total
Test Set	#	639	104	32	25	800
	%	79.9	92.9	96.9	3.1	100.0
Trained Set	#	179	15	3	3	200
	%	89.5	97.0	98.5	1.5	100.0

5. Conclusions

In this paper, we have presented a system to recognize Tamil handwriting using a two-stage classification approach. The number of possible candidates for a character is narrowed down to a smaller subset of the alphabet, in the first stage. Then, the actual recognition is performed using a statistical classifier. The results show 96.9% recognition rate for the top three choices. The main recognition errors were due to abnormal writing and ambiguity among similar shaped characters. Abnormal writing caused a character to be pre-classified into a wrong group, thereby resulting a misrecognition. Most of the confusion was between character pairs such as "¡"& "®", "_"& "¿" and "º"& "¸". This could be avoided by using a word dictionary to look-up for possible character compositions. The presence of contextual knowledge will help to eliminate the ambiguity.

Extension of the system to cater for the full Tamil alphabet would require splitting a composite character into basic recognizable symbols. The methods used in [15] and [16] are under investigation for this purpose. Future work could also include extracting more robust features for the classifier to achieve better discrimination power. We strongly feel that the method of pre-classification would have much higher recognition accuracy if applied to Optical Character Recognition, since printed characters preserve the correct positioning on three-zone frame.

References

[1] R. M. Bozinovic and S. N. Srihari, "Off-line cursive script word recognition", *IEEE Trans. on Pattern Anal. Mach. Intell.*, vol. 11, no. 1, pp. 68-83, Jan. 1989.

[2] Hu, M. K. Brown and W. Turin, "HMM based on-line handwriting recognition", *IEEE Trans. on Pattern Anal. Mach. Intell.*, vol. 18, no. 10, pp. 1039-1045, Oct. 1996.

[3] D. Deng, K. P. Chan, and Y. Yu, "Handwritten Chinese character recognition using spatial Gabor filters and self-organizing feature maps", *Proc. IEEE Inter. Confer. on Image Processing,* vol. 3, pp. 940-944, Austin TX, June 1994.

[4] C-H. Chang, "Simulated annealing clustering of Chinese words for contextual text recognition", *Pattern Recognition Letters*, vol. 17, no. 1, pp. 57-66, 1996.

[5] H. Yamada, K. Yamamoto, and T. Saito, "A non-linear normalization method for handprinted Kanji character recognition–line density equalization", *Pattern Recognition*, vol. 23, no. 9, pp. 1023-1029, 1990.

[6] S-W Lee, "Off-line recognition of totally unconstrained handwritten numerals using multiplayer cluster neural network", *IEEE Trans. on Pattern Anal. Mach. Intell.*, vol. 18, no. 6, pp. 648-652, June 1996.

[7] J. Cai and Z-Q Liu, "Integration of structural and statistical information for unconstrained handwritten numeral recognition," *IEEE Trans. on Pattern Anal. Mach. Intell*., vol. 21, no. 3, pp. 263-270, Mar. 1999.

[8] V. Bansal and R.M.K. Sinha, "On how to describe shapes of Devanagari characters and use them for recognition", *Proc. 5th Int. Conf. Document Analysis and Recognition*, Bangalore, India, pp. 410-413, Sept. 1999.

[9] P. Chinnuswamy, S.G. Khrishnamoorthy, "Recognition of handprinted Tamil characters", *Pattern Recognition,* vol. 12, pp. 141-152, 1980.

[10] N. Damayanthi, P. Thangavel, "Handwritten Tamil character recognition using Neural Network", *Proc. The Tamil Internet 2000 Conference*, Singapore, July 2000.

[11] R.M. Suresh, S. Arumugam and K.P. Aravanan, "Recognition of handwritten Tamil characters using fuzzy classificatory approach", *Proc. The Tamil Internet 2000 Conference*, Singapore, July 2000.

[12] B. B. Chaudhuri and U. Pal, "A complete printed Bangla OCR system", Pattern Recognition, vol. 31, no. 5, pp. 531-549, 1997.

[13] The Unicode Consortium, *The Unicode Standard 3.0*, Harlow: Addison Wesley publishers, 2000.

[14] R. C. Gonzalez, R. E. Woods, *Digital Image Processing*, Addison Wesley publishers, 1993.

[15] M. Lohakan, S. Airphaiboon and M. Sangworasil, "Single-character segmentation for handprinted Thai word", *Proc. 5th Int. Conf. Document Analysis and Recognition*, Bangalore, India, Sept. 1999.

[16] A. Bishnu and B. B. Chaudhuri, "Segmentation of Bangla handwritten text into characters by recursive contour following", *Proc. 5th Int. Conf. Document Analysis and Recognition*, Bangalore, India, Sept. 1999.